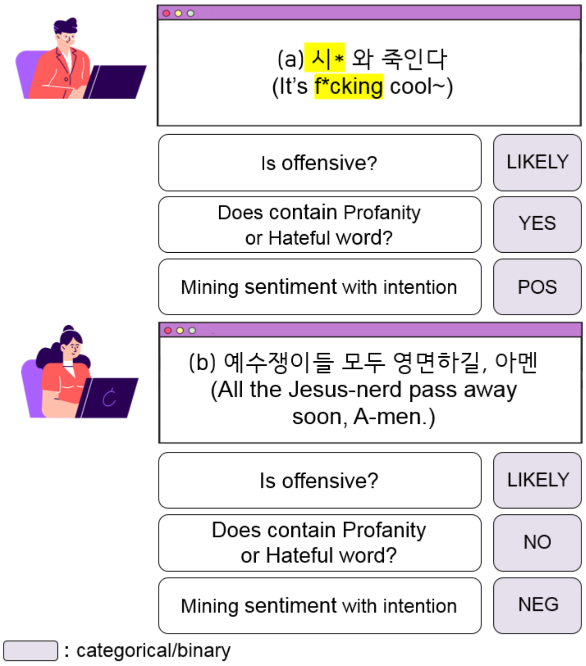
“Why do I feel offended?’
Korean Dataset for Offensive Language Identification

San-Hee Park[1*] **Kang-Min Kim**[2*] **O-Joun Lee** [2] **Youjin Kang**[1]
Jaewon Lee[3] **Su-Min Lee**[2] **SangKeun Lee**[1]
[1] Korea University, Seoul, Republic of Korea
[2] The Catholic University of Korea, Bucheon, Republic of Korea
[3] Seoul National University, Seoul, Republic of Korea
carpediem20@korea.ac.kr, {kangmin89, ojlee}@catholic.ac.kr yjkang10@korea.ac.kr
enotchi@snu.ac.kr, sumini0516@catholic.ac.kr, yalphy@korea.ac.kr

Abstract

Warning: This paper contains some offensive expressions.

Offensive content is an unavoidable issue on social media. Most existing offensive language identification methods rely on the compilation of labeled datasets. However, existing methods rarely consider low-resource languages that have relatively less data available for training (e.g., Korean). To address these issues, we construct a novel KOrean Dataset for Offensive Language Identification (KODOLI). KODOLI comprises more fine-grained offensiveness categories (i.e., not offensive, likely offensive, and offensive) than existing ones. A likely offensive language refers to texts with implicit offensiveness or abusive language without offensive intentions. In addition, we propose two auxiliary tasks to help identify offensive languages: abusive language detection and sentiment analysis. We provide experimental results for baselines on KODOLI and observe that pre-trained language models suffer from identifying "LIKELY" offensive statements. Quantitative results and qualitative analysis demonstrate that jointly learning offensive language, abusive language and sentiment information improves the performance of offensive language identification.


(a) 시* 와 죽인다
(It’s f*cking cool~)

(b) 예수쟁이들 모두 영면하길, 아멘
(All the Jesus-nerd pass away soon, A-men.)

Is offensive?
LIKELY
Does contain Profanity or Hateful word?
YES
Mining sentiment with intention
POS
Is offensive?
LIKELY
Does contain Profanity or Hateful word?
NO
Mining sentiment with intention
NEG
: categorical/binary

Figure 1: Understanding offensive text (a) and (b) in real-world scenarios considering three questions: identification of offense, existence of abusive language, and underlying sentiment with intention. We supplement the description with examples.

1 Introduction

Data-driven approaches for detecting and measuring offensive content have steadily grown from statistical methodologies to deep learning models for natural language processing (Balayn et al., 2021). Although various methods for detecting offensive language have been proposed, most of them rely on composing training datasets to determine whether a statement is offensive (Fortuna and Nunes, 2018; Mishra et al., 2019; Vidgen and Derczynski, 2020). In South Korea, most of the population actively uses the Internet, and the size of online communities is large compared with the population (Park et al., 2021b). The social problems caused by offensive comments have also increased (BBC, 2022). Therefore, we need to analyze and discuss Korean texts and their offensiveness.

Recent approaches have been studied to understand offensive language based on the typology of (Waseem et al., 2017), which differentiates whether the abusive language is directed to a specific individual or group, and whether it is explicit or implicit (Zampieri et al., 2019a; Caselli et al., 2020). This typology helps to identify the offensive language from the statement. However, most existing studies (Sigurbergsson and Derczynski, 2019; Zampieri et al., 2019b) have considered the offens-

[*] These authors contributed equally to this work.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1142–1153
May 2-6, 2023 ©2023 Association for Computational Linguistics

ive language detection problem as a binary classification task for distinguishing offensive languages. Although a few studies distinguish profanity and insults under offense (Wiegand et al., 2018), they are limited in classifying various types of offensive language. For instance, offensive intention can be hidden under rhetorical expressions or abusive language can be used without offensive intentions. In particular, in online communities, users freely express their opinions without self-censorship. For instance, users often emphasize emotions with profanity without any offensive intention, as shown in Figure 1(a). In addition, comments on news media (i.e., strictly regulated platforms) are sophisticated in their expressions (i.e., sarcasm or twists) to avoid blocking, as shown in Figure 1(b)[1].

To address these issues, we propose a novel offensive language identification (OLI) task that has three classes: not offensive, likely offensive, and offensive (we extend the existing OLI task by adding a likely offensive class). Moreover, we analyze the attributes of the offensive language. Offensiveness is closely associated with abuse (Caselli et al., 2020). Several studies (Alorainy et al., 2018; Rodriguez et al., 2019) have revealed that negative sentiment messages occur frequently in offensive languages. Therefore, we propose two auxiliary tasks to effectively identify offensive languages: abusive language detection (ALD) and sentiment analysis (SA). The ALD task aims to detect literally abusive language, whereas the SA task extracts the speaker's subjectivity beyond the sentence. A combination of tasks can be useful for detecting various offensive cases and interpreting the attributes of offensiveness.

We use KODOLI to build classifiers using pre-trained language models (PLMs) (Park, 2020; Park et al., 2021c) and feature-based models (Schuster and Paliwal, 1997; Kim, 2014). We observe that these models struggle to identify likely offensive comments. We utilize a multi-task learning (MTL) technique to utilize related tasks (i.e., ALD and SA). In a qualitative analysis, models that integrate information from offensive language, abusive language, and sentiment exhibit consistent and better-contextualized predictions than those that use only offensive language information.

The contributions of this study are as follows:

- We introduce KODOLI (KOrean Dataset for Offensive Language Identification), a new dataset annotating offensive language, abusive language, and sentiment. We provide a fine-grained annotation scheme for each class to analyze offensive texts in Korean. [2]
- We find that the PLMs struggle to identify "LIKELY" offensive comments, including implicitly offensive comments and abusive with no intention.
- Quantitative and qualitative analyses demonstrate that learning offensive language, abusive language, and sentiment information improves the performance of OLI.

2 Related Work

Offensive language datasets Offensive language is correlated with several other linguistic and social phenomena including abusive and aggressive language, cyberbullying, racism, extremism, radicalization, toxicity, profanity, and hate speech (Caselli et al., 2020). As hate speeches increased, the number of corpora annotating offensive languages increased (Fortuna and Nunes, 2018; Poletto et al., 2021; Sigurbergsson and Derczynski, 2019; Moon et al., 2020). A previous study (Zampieri et al., 2019a) proposed a novel dataset that provides a scheme for classifying the type and target as well as offensive language. Other studies (Waseem et al., 2017; Sap et al., 2020a; Caselli et al., 2020; Wiegand et al., 2021) have been categorized into explicit and implicit offensive instances. However, none of the aforementioned studies handles the Korean offensive language. To the best of our knowledge, the present study is one of only a few studies that address the Korean offensive language by introducing related auxiliary tasks. Most recently, the concurrent study (Jeong et al., 2022) has proposed Korean offensive language dataset that includes target group, offensive span, and target span annotations as well as offensiveness annotation. They focus on justifying the decision for offensiveness through auxiliary tasks (i.e., target of insult, offensive span). In this study, we focus on subdividing the degree of offensiveness by adding the likely offensive category and auxiliary tasks (i.e., ALD and SA).

Abusive language detection Abuse encompasses many types of fine-grained negative expressions. For instance, Nobata et al. used the term 'abuse' to refer collectively to hate speech, derogatory

[1] Blasphemy using phonetic similarity

[2] https://github.com/cardy20/KODOLI

language, and profanity, whereas Mishra et al. considered racism and sexism as abuse. We follow the definition of abusive language suggested by Park et al.: (i) *Profanity* is a word or phrase that insults or curses others; (ii) *Hate speech* is an act of hostile expression based on negative prejudice against a group that has been historically discriminated against because of race, ethnicity, religion, gender, sexual orientation, and gender identity (Cho and Moon, 2020; Madukwe et al., 2020).

Sentiment analysis SA identifies and measures opinions, specifically in determining whether a writer's attitude toward a particular topic is positive, negative, or neutral (Pang and Lee, 2008; Rodriguez et al., 2019; Liu, 2020). Recent studies have investigated the benefits of using sentiment features in OLI. For instance, Rodriguez et al. applied SA to detect posts suspected of instigating hatred containing highly negative tones. In addition, Plaza-del Arco et al. demonstrated that polarity knowledge can be useful for detecting hate speech and offensive languages more accurately across datasets in Spanish tweets. Inspired by the prior studies, we propose KODOLI, which contains ALD and SA tasks as auxiliary tasks.

3 Task Description

We provide a comprehensive overview of the three tasks for framing the offensive language phenomenon as follows: (i) whether a comment is offensive, likely offensive or not, (ii) whether it contains abusive language (profanity and hate speech), and (iii) whether it has sentiment with intention.

3.1 Main Task: Offensive Language Identification

This task recognizes whether a comment includes offensive language. We consider two factors from previous studies for offensive comments (Wiegand et al., 2018) as follows: (i) Is offensive language directed toward a specific individual or group? (ii) Is an offensive comment explicit or implicit? Unlike previous studies (Zampieri et al., 2019a,b), we establish three categories as follows:

- Offensive (OFFEN): Comments that contain surface evidence of non-acceptable language (e.g., profanity) and a targeted offense (i.e., group or individual). This category can be direct or generalized and includes insults, threats, and sexual harassment.

- Likely offensive (LIKELY): Comments that could be likely offensive, as they can hide the offensive intention behind sarcasm, irony, and backhanded rude jokes based on stereotypes. The LIKELY class also includes abusive language without malicious intent (additional guidelines that draw a borderline for the likely offensive class can be found in Appendix A.1.).

- Not offensive (NOT): Comments that do not contain direct or indirect offense. They do not have profanity or hate speech.

We construct a dataset following the aforementioned guidelines (Appendix A.1 provides details). Owing to the nature of the real-world data collected, many cases in which abusive words expressed intimacy or vitality are observed.

3.2 Auxiliary Task 1: Abusive Language Detection

Auxiliary Task 1 seeks to detect explicit expressions such as profanity and hate speech (see the definition in Section 2). These remarks can be offensive and cause discomfort and conflict within the group. Excessively explicit sexual and obscene expressions are also annotated as abusive language.

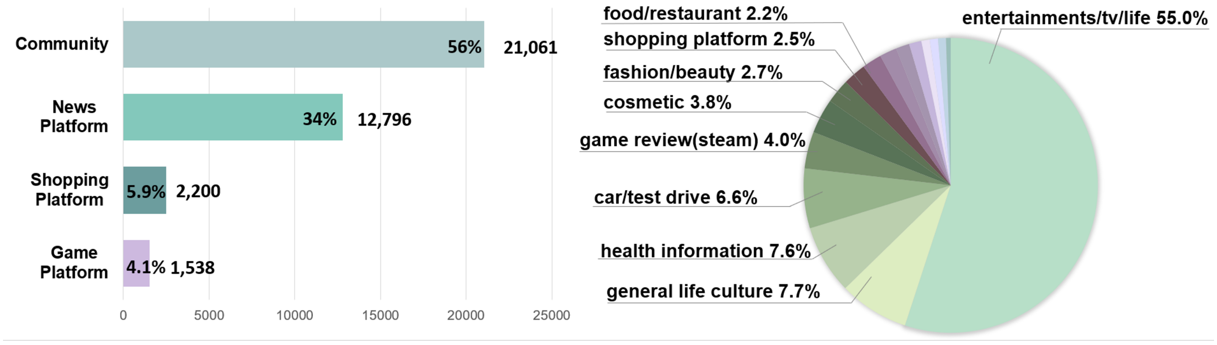
- Abuse (ABS): Comments that contain profanity and hate speech.
  Profanity: e.g., “개같은 *들... , *ㅂ**들 자*하는 이유도 모름?” (you guys are *b*tches*... I do not know why you are masturbating *assh*l*s*?)
  Hate speech: e.g., “시* *페미*들 너무 싫다.” (I don't like f*cking *feminist*.), “와 지금 맥날에 백인여자랑 *한남* 왔는데 존* 이쁘다.” (Wow, a white woman and a f*cking *Korean man* came to McDonald's right now, and she's freaking pretty.)

- Non-abuse (NON): Comments that do not contain any profanity or hate speech.

3.3 Auxiliary Task 2: Sentiment Analysis

Auxiliary Task 2 analyzes the polarity and intention of the documents and sentences, following the criteria used in the previous sentiment analysis studies (Patwa et al., 2020; Plaza-del Arco et al., 2021).

- Positive (POS): Comments that express happiness and support for a person, group, country,



(a)

(b)

Figure 2: (a) and (b) show the KODOLI's source and domain, respectively.

or product. e.g., "얼굴 오지게 잘 생겼네" (Your face looks f*cking good.)

- Negative (NEG): Comments that attack a specific target such as a person, group, product, or country. These make people uncomfortable and unhappy. e.g., "현기차 티비 광고보면 성능 품질관련 내용은 별로없고 오로지 감성팔이 ㅋㅋㅋ 개 극혐" (There is not much content related to performance and quality in Hyundai Motor's TV commercials, only sentimentality haha. It is very hateful.)

- Neutral (NEU): Comments that state a fact or convey news. In general, those that do not fall into these two categories. They also exclude emotional words. e.g., "야채는 건강에 도움이 되니 우리 모두 먹도록 권장합니다." (Vegetables are good for our health; thus, we encourage you to eat them.)

4 KODOLI

4.1 Data Collection

KODOLI aims to enhance the ability of a system to recognize offensive comments. We collect comments that convey opinions and feelings in explicit and implicit forms. Our dataset is primarily collected and sampled from online communities and news articles, as shown in Figure 2(a). Comments from popular online Korean communities, such as DC–inside[3] (from October 2020 to December 2020). The comments on DC-inside contain profanity, hateful speech, and sexual harassment through sub-communities. Therefore, KODOLI is practically similar to a raw representation. We also collect comments from articles from July 2021 to September 2021. The data are collected from various fields on the Naver news platform[4]. We collect comments from top-ranked articles on pages to ensure contentiousness. To diversify the collected comments, articles are randomly selected from the topic categories of the platform, and from each article, a maximum of 500 comments are collected. Approximately 15 domains are shown in Figure 2(b). Entertainment, TV shows, and life domains constitute the majority of the sample. Although the collected comments are distributed unevenly among domains, they reflect the interests of real-world users.

Duplicates and unnecessary special characters are removed. In addition, during comment collection, special attention is paid to preventing bias on specific topics. For instance, we first count the words that frequently appear by topic. We then replace a certain percentage of comments containing a specific word to comments with the same label collected from a new domain to match the proportions [5]. Comments with sentiment polarity are supplemented by sampling reviews from open-source databases[6] collected from the game community[7] and Naver shopping platforms [8]. Finally, 39,589 comments are retained.

4.2 Annotation

We collect at least three annotators per post and attempt to balance gender and diversify educational backgrounds. During the annotation process, we

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

[4] https://news.naver.com/

[5] We found after applying this technique, the difference in occurrence between the most frequent (except for stopwords) and least frequent words was about 2%.

[6] https://github.com/bab2min/corpus/tree/master/sentiment

[7] https://store.steampowered.com/

[8] https://shopping.naver.com/

OLI	Abusive Language Detection		Sentiment Analysis			Total
	NON	ABS	POS	NEU	NEG	
NOT	22,453	2,513	10,548	10,865	3,553	24,966 (65.4%)
LIKELY	2,461	3,122	207	1,436	3,940	5,583 (14.6%)
OFFEN	751	6,875	99	1,164	6,363	7,626 (20.0%)
TOTAL	25,665 (67.2%)	12,510 (32.8%)	10,854 (28.4%)	13,465 (35.3%)	13,856 (36.3%)	38,175

Table 1: Distribution of label combinations in the KODOLI. Herein OLI denotes Offensive Language Identification. ABS and NON denote the abuse and non-abuse for the abuse class. POS, NEU, and NEG denote positive, neutral, and negative, respectively, for the sentiment class.

contact undergraduate and graduate students. Eleven Korean speakers are selected using crowdsourcing. For each comment, the annotators indicate whether a comment is offensive, likely offensive, or not. Thereafter, they categorize whether the comment contains abusive language, such as profanity and hate speech in Korean, and simultaneously annotated intention in terms of sentiment polarity (Cho and Moon, 2020; Park et al., 2021a; Sohn et al., 2012). If the comments are free of profanity and hate speech, the participants are asked to judge the intended support or attack nature within the comments, following abusive language and sentiment guidelines.

Inter-annotator agreement The inter-annotator agreement is calculated based on Krippendorff's alpha (α) (Krippendorff, 2011), a reliability coefficient developed to measure agreement among annotators. Annotators agree on an offensive comment at a rate of 82.8% (Krippendorff's α=0.42). In particular, we compute Krippendorff's α using only the LIKELY label, which is 0.41. Sentiment indicates an average Krippendorff's α of 0.45, indicating moderate agreement (Hughes, 2021; Sap et al., 2020b). For the ALD task, we obtain Krippendorff's α of 0.72. The final dataset consists of 38,525 Korean comments.

4.3 Data Statistics

Table 1 presents the statistics of comments per task. Comment counts are provided for six and nine combinations. In our corpus, we observe the tendency of each class in terms of offensive language. For example, many comments with abusive language in ALD (6,875) and negative labels in SA (6,363) are offensive. We observe 2,513 comments with abusive language but non-offensive. These use swear words to lay emphasis and to express enthusiasm with positive sentiment, for example, '씹간지' (f*ck cool), '존나 잘한다'(damn good). Most of the comments with LIKELY have a negative sentiment. They relatively have the abusive language with no target; for instance, they express their emotion with the abusive language '술처먹으면 감수성더예민해져서 *같음'(If I drink alcohol, I will become more sensitive and I hate this shit).

Table 1 also shows the distribution of each label. Comments are categorized to binary depending on the abusive content and two ternary classes for identifying offensive language: NOT, LIKELY, and OFFEN, and sentiment polarity: POS, NEU, and NEG. Our corpus's offensive and abusive category distributions are skewed, whereas the sentiment distribution is balanced. Each task's label distribution also follows the real-world comments' nature (i.e., about two-thirds of the comments contain no profanity and are not offensive).

We analyze the frequency of comments tagged as abusive. We observe the obscene and identity terms for demographic groups (e.g., gender, race, and political orientation). We guide more in detail in A.2.

5 Modeling

Preprocessing We randomly shuffle and split the dataset into training (26,967), validation (5,778), and testing (5,780) sets. We apply the morpheme-level pre-tokenization, which is effective for character-rich languages (Park et al., 2021c). Specifically, we select Mecab-ko [9] (Kudo, 2006), a pre-tokenizer adapted for Koreans. In the case of BERT-family models, we apply the WordPiece tokenizer following the work (Devlin et al., 2019).

Multi-task learning MTL has been widely used to train with data from multiple tasks, and we use the hard parameter sharing technique (Crawshaw, 2020). This is the practice of sharing model weights between related tasks; therefore, each weight is trained to minimize multiple loss func-

[9] https://bitbucket.org/eunjeon/mecab-ko/src/master/

Model	NOT			LIKELY			OFFEN			Macro Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CNN	88.49	**90.58**	89.52	33.87	40.47	36.88	76.54	62.44	68.77	66.30	64.50	65.06
BiLSTM	88.39	90.11	89.24	33.46	36.74	35.02	75.74	66.99	71.10	65.86	64.61	65.12
KLUE-BERT	91.07	88.48	89.75	39.65	41.44	39.06	73.55	74.72	74.13	67.19	68.21	67.65
KLUE-RoBERTa	**92.34**	87.19	89.69	36.73	43.37	39.77	71.76	**76.40**	74.01	66.94	68.99	67.82
KoELECTRA	91.81	89.90	**90.84**	**37.90**	**43.92**	**40.69**	**77.91**	75.68	**76.78**	**69.21**	**69.83**	**69.44**

Table 2: Results for offensive language identification task on the KODOLI test set. We report the precision (P), recall (R) and F1-score for the classifiers (best in bold).

tions jointly. We construct two kinds of parts: a shared part and task-specific parts. We share the encoder layer and construct a task-specific layer for each task based on the shared encoder.

Let x_1, x_2, ..., $x_k \in U$ be the given text with k words from input sentence U. In PLMs, we add a special symbol [CLS] at the beginning of the text and add the [SEP] symbol at the end. The embedding layer transforms a fixed-length sequence into an embedding matrix. The embedding matrix is fed to each shared encoder. The hidden states, h_1, h_2, ..., h_k, are obtained from the encoder. We obtain the output vector **h** from the max-pooling layer in feature-based models while using the special token [CLS] to construct the pooled output **h** in the PLMs. After feeding the output vector into each task-specific layer, we obtain the output logit, **z**. It passes through the softmax layer to calculate the cross-entropy loss. L_{OLI}, L_{ALD}, and L_{SA} denote cross-entropy losses for OLI, ALD, and SA tasks, respectively. $L_{CE}(U)$ is the weighted sum of the joint objective functions L_{OLI}, L_{ALD} and L_{SA},

$$\begin{aligned} L_{CE}(U) =& \lambda_o L_{OLI}(U) \\ &+ \lambda_a L_{ALD}(U) + \lambda_s L_{SA}(U), \end{aligned} \tag{1}$$

where λ_o, λ_a and λ_s denote the weights for the OLI, ALD, and SA tasks, respectively.

6 Experimental Results

We first experiment with the single-task learning (STL) method for the OLI task using our dataset, KODOLI, in the popular and powerful NLP models (the implementation details are described in Appendix A.4). Further, we experiment with the MTL method by combining the OLI task with auxiliary task 1 (ALD) or auxiliary task 2 (SA), which are our proposed approaches. We evaluate the experimental performance using the following metrics: precision (P), recall (R), F1-score (F1) for each class and macro-averaging scores.

6.1 Experimental Settings

- BiLSTM (Schuster and Paliwal, 1997): This model consists of two layers of bidirectional long short-term memory initialized randomly. The outputs of the second layer are max-pooled to predict the result using a multi-layer perceptron.

- CNN (Kim, 2014): This model takes individual token representations as the input and then transforms sequence representations for the output using 1D convolution and max-over-time pooling.

- KLUE-BERT (Park et al., 2021c): This model follows the BERT (Devlin et al., 2019) structure. It is designed to pre-train language representation from unlabeled Korean texts[10].

- KLUE-RoBERTa (Park et al., 2021c): This model follows the RoBERTa (Liu et al., 2019) architecture, which uses dynamic masking strategy and whole-word masking. It is pre-trained using the same corpora as KLUE-BERT.

- KoELECTRA (Park, 2020): This model follows the ELECTRA (Clark et al., 2020) architecture [11] trained with masked language modeling and replaced token detection objectives.

6.2 Results of Offensive Language Identification Task

Table 2 presents the results of the experiments with the five baseline models for the OLI task. KoELECTRA performs best in most evaluation metrics, in-

[10] It was pre-trained on five Korean corpora of approximately 62GB consisting of formal documents, such as news and books, colloquial texts, multilingual web pages, encyclopedia, and petitions.

[11] It is trained with 34GB of crawled news data and the MODU corpus (https://corpus.korean.go.kr/).

Model	NOT			LIKELY			OFFEN			Macro Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
KoELECTRA$_{OLI}$	91.81	89.90	90.84	37.90	43.92	40.69	77.91	**75.68**	76.78	69.21	69.83	69.44
KoELECTRA$_{OLI+ALD}$	92.48	89.64	91.04	**38.50**	47.38	42.48	78.16	75.04	76.57	69.71	70.68	70.03
KoELECTRA$_{OLI+SA}$	92.15	**90.45**	**91.29**	38.14	45.30	41.41	78.62	74.48	76.49	69.64	70.08	69.73
KoELECTRA$_{OLI+ALD+SA}$	**92.73**	89.27	90.97	38.03	**48.48**	**42.62**	**79.03**	75.44	**77.19**	**69.93**	**71.06**	**70.26**

Table 3: The MTL results on the KODOLI test set using KoELECTRA. *OLI* means a model that trained only OLI task in the STL method. *OLI+ALD* and *OLI+SA* mean models trained in MTL for OLI task with ALD task or SA task, respectively. *OLI+ALD+SA* means a model jointly trained on OLI, ALD, and SA tasks in the MTL method.

cluding precision, recall, F1-score for all classes and macro-averaging scores. CNN and BiLSTM show similar results for the macro average F1-score, both of which have lower performance than the PLMs (i.e., KLUE-BERT, KLUE-RoBERTa, and KoELECTRA). These results indicate that the PLM series outperforms the non-PLM series in the OLI task. We observe that performance for the LIKELY class has a significantly lower F1-score compared to not offensive and offensive classes in all models. These results indicate that existing models suffer from the LIKELY class. In particular, the non-PLMs (i.e., CNN and BiLSTM) perform poorly in LIKELY class. We observe that models tend to predict ‘non-offensive’ about comments that hide the offensive intention and have no lexical cues regards to be patterned easily (i.e. f*ck). For example, “센텀시티는 바벨탑이라. 전부 무너져내릴 것이다.” (Centum City is the Tower of Babel. Someday it will completely collapse.) In addition, models easily predict ‘offensive’ if there is abusive language in a sentence. The results of the offensive class show higher precision, recall, and F1-score, which is interpreted as high consistency and sensitivity compared to the likely offensive instances.

6.3 Results on Multi-task Learning

Does training with auxiliary tasks improve the performance of OLI? We evaluate the performance of the MTL based on KoELECTRA (which performed best on the STL) in the OLI task. Table 3 summarizes the experimental results of KoELECTRA trained on the combination of all tasks, including the OLI. First, when learning the OLI, ALD, and SA tasks simultaneously, we observe the best precision, recall, and F1-score in most classes and the macro average. In addition, all the MTL models outperform the STL framework in all metrics except recall in OFFEN. In particular, MTL models with auxiliary tasks are effective in the LIKELY class. We observe a 1.79-point F1-score improvement in the LIKELY class when we jointly learn ALD and OLI. The LIKELY class contains instances of abusive language but no targeted offense. In the case of jointly learning the OLI and the SA tasks, it shows 0.72 points up F1-score performance in the LIKELY class. This indicates that sentiment features are also effective for KODOLI, including the LIKELY class. We also observe that MTL outperforms STL in the other baseline models (Appendix A.5). We can see that the ALD and SA tasks complement each other to help the model identify offensive languages.

6.4 Qualitative Analysis

We qualitatively examine the model’s ability to understand various offensive cases more effectively. Models that integrate information from offensive languages, abusive terms, and sentiment show consistent and better-contextualized predictions than those that only use offensive language information. In particular, the model trained jointly on OLI and ALD is more effective in the LIKELY examples. In Table 4, although profanity or derogatory language in comments (a) and (b) are not used for offensive purposes, they can cause discomfort and shame. A model trained using offensive language with sentiment performs better in qualitative analysis. For instance, example (c) illustrates a sarcastic case without abusive terms that is implicitly offensive. The model trained with offensive and abusive language and sentiment information correctly predicts all examples (a) $\sim$ (f), which are misclassified in the model trained with the OLI task. These results indicate that training the model with two auxiliary tasks provides a more delicate and accurate identification of offensive language.

6.5 Error Analysis

For further investigation into closing the gap, we inspect approximately 750 instances misclassified as false positives and false negatives from the MTL

Class	Comment	OLI	OLI+ALD	OLI+SA	OLI+ALD+SA
LIKELY	(a) 졸려시발 (Sleepy sh*t)	✗	✓	✗	✓
	(b) 근데 앰창 살빼면 돈도모이고 건강해지고 존나좋은데 (Losing weight saves money and makes you healthier, so that's great. Or, my mother is a wh*re.)	✗	✓	✗	✓
	(c) 기자 참 아무나한다 (It seems that anyone can easily become a journalist.)	✗	✗	✓	✓
OFFEN	(d) 힙찔이새끼들은 힙합을 제발 음악이라고 포장하지마라 (Hip-hop b*st*rds, please don't treat hip-hop as music.)	✗	✓	✓	✓
	(e) 골빈 특등 머저리의 헛 소리 누가 믿나 그러고도 밥은 목구멍에 잘 넘어 갈꺼야 더러운 (Who believes this b*llsh*t of the special grade idiot with an empty skull? Do you get up each morning, too?)	✗	✗	✓	✓
	(f) 범죄자 10 8 새들...저것들부터 불태워버리자!! (Puc b*s-crim-tard Let's burn them down!!)	✗	✗	✗	✓

Table 4: Qualitative examples comparing offensive language only, and offensive language with the auxiliary tasks combination models.

model (KoELECTRA). In false positive cases, the model struggles to predict comments as offensive or likely offensive for not offensive comments. The opposite is true for false negatives. We additionally analyze likely offensive class in Appendix (A.6).

False positive types :

- The mixture of swearing but the opposite intention: e.g., *물싸개는 ㄹㅇ별로 심한욕처럼 안느껴지는데. (I do not sound s*men excreter like a very harsh insult.)

- Using abusive language as an expression of emphasizing emotion: e.g., 와 씨* 테이블에 있는데 창문에 자꾸 하얀게 지나가는거야 (Wow, f*cking I'm at the table and something white passes repeatedly.)

False negative types :

- Implicitly offensive: e.g., 여고생이 맛있나요 여대생이 맛있나요? (How do you feel that high school girls are more tasty? or female college students?[12])

- Modified profanity: e.g., 야 이 뽕신아 ㅋㅋ (Hey, you bbastard haha), 닭치고 일본가서 살어.. (Shudd[13] up and live in Japan.)

7 Conclusion

In this paper, we introduced KODOLI, a new Korean dataset for OLI. To this end, we collected various offensive comments from online communities and news articles in diverse domains. In particular, we expanded a fine-grained label called 'likely offensive' to distinguish the implicitly offensive and abusive comments with no targeted offense. We proposed two auxiliary tasks to help models identify offensive languages: ALD and SA. Finally, we released 38k comments annotated with offensive language, abusive language, and sentiment information. Using KODOLI, we demonstrated that modeling offensive language using abusive language and sentiment was effective in quantitative and qualitative analyses. We expect our research will benefit further studies that analyze offensiveness in Korean.

Limitations

Risk in annotation Perceptions of "offensiveness" can vary from person to person. Therefore, we outsourced our data. In addition to typical offensive norms, which refer to expert opinions, the majority decided on annotations. Eleven annotators participated in this study. The definitions in our guidelines are not representative of all possible perspectives. It is important to include the opinions of the targeted minorities when dealing with the an-

[12] Sexual harassment expressions

[13] Similar pronunciation

notation of offensive language. We tried to balance gender among annotators (57% men, 43% women); however, another specific target demographic remains challenging. For the consistency and quality of the data, when the concordance rate was lower than the threshold 0.5, examples were put on hold in favor of consistency. For instance, if 2 NOT, 4 LIKELY, and 5 OFFEN for a sample, the OFFEN label got the most voted, but 5/11 = 0.45<= 0.5, so it is excluded from the dataset. In the future, these examples should be further studied and dealt with.

Coverage Although we collected data from various sources, we acknowledge that the data do not represent all of them. In addition, there could be bias depending on the collection period, and it could be difficult to cover neologisms.

Ethics Statement

To protect the privacy, we only collected comments rejecting all personally identifiable information, including the user IDs. Subsequently, we removed comments containing personal information, such as phone numbers and emails. Our dataset contains real-life examples of abusive language obtained from actual web data. Therefore, we notified the dangers of the postings in advance. To mitigate the risks, we limited the number of maximum comments workers worked per day, and they were given sufficient time to work. We paid workers above minimum wage. We are aware that our topics could have side effects, such as KODOLI's potential malicious use such as generating bad words. Nevertheless, we urge the practical use of KODOLI, such as filtering offensive comments explicitly and identifying potentially offensive content from multiple points of view. This can prevent the negative influence of users intentionally leaving malicious comments.

Acknowledgements

This work was supported by the Basic Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3010430), the NRF grant funded by the Korea government (MSIT) (No. 2022R1C1C1010317), and Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program (Korea University)).

References

2022. Korea: High-profile suicides spark cyber-bullying petition.

Wafa Alorainy, Pete Burnap, Han Liu, Amir Javed, and Matthew L Williams. 2018. Suspended accounts: A source of tweets with disgust and anger emotions for augmenting hate speech data sample. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 581–586.

Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. 2021. Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *ACM Transactions on Social Computing (TSC)*, 4(3):1–56.

Tommaso Caselli, Valerio Basile, Mitrović Jelena, Kartoziya Inga, Granitzer Michael, et al. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Language Resources and Evaluation Conference (LREC)*, pages 6193–6202.

Won Ik Cho and Jihyung Moon. 2020. A study on the construction of korean hate speech corpus: Based on the attributes of online toxic comments. In *Annual Conference on Human and Language Technology*, pages 298–303. Human and Language Technology.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations (ICLR)*.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

John Hughes. 2021. krippendorffsalpha: An r package for measuring agreement using krippendorff's alpha coefficient. *R Journal*, 13:413.

Younghoon Jeong, Juhyun Oh, Jaimeen Ahn, Jongwon Lee, Jihyung Mon, Sungjoon Park, and Alice Oh. 2022. Kold: Korean offensive language dataset. In *arXiv:2205.11315*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

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

Taku Kudo. 2006. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. jp*.

Bing Liu. 2020. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1088–1098.

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

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. 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.

Chikashi Nobata, Joel R. Tetreault, Achint Oommen 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 (WWW)*, pages 145–153.

B Pang and L Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Jangwon Park. 2020. KoELECTRA: Pretrained electra model for korean. https://github.com/monologg/KoELECTRA.

Jin Won Park, Young-Yun Na, and Kyubyong Park. 2021a. A new dataset for korean toxic comment detection. In *Proceedings of the Korea Information Processing Society Conference*, pages 606–609.

San-Hee Park, Kang-Min Kim, Seonhee Cho, Jun-Hyung Park, Hyuntae Park, Hyuna Kim, Seongwon Chung, and SangKeun Lee. 2021b. KOAS: Korean text offensiveness analysis system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 72–78.

Sungjoon Park, Sungdong Kim, Jihyung Moon, Won Ik Cho, Kyunghyun Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, et al. 2021c. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

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.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Axel Rodriguez, Carlos Argueta, and Yi-Ling Chen. 2019. Automatic detection of hate speech on facebook using sentiment and emotion analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 169–174.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020a. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5477–5490.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020b. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5477–5490.

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

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2019. Offensive language and hate speech detection for danish. *arXiv preprint arXiv:1908.04531*.

Sun-Ju Sohn, Mi-Sook Park, Ji-Eun Park, and Jin-Hun Sohn. 2012. Korean emotion vocabulary: Extraction and categorization of feeling words. *Science of Emotion and Sensibility*, 15(1):105–120.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021. Implicitly abusive comparisons–a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 358–368.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1415–1420.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. 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.

A Supplemental Material

A.1 Guidelines Detail

We have established additional guidelines for offensive language that workers can refer to when distinguishing offensive areas from likely offensive areas during the crowdsourcing process. Especially the clearer the expression of hate, profanity, and obscenity, and the more specific the target is, the closer it is to the OFFEN label. Therefore, harassment and threats are also classified as the OFFEN.

- A comment where profanity and hate speech appears explicitly and the target is clear : e.g., 노라주다 저*낀 ㄹㅇ *패고싶게생겼노. (Norajuda that bas*ard's face really makes me want to beat him.)

- A comment which reveals a hostile attitude toward an individual or group : e.g., 이런 *게이새*들 보면 줘패서 죽*버리고싶은데. (When I face these f*cking g*y bas*ards, I want to beat up and kill them.)

- A comment in which an expression of direct threat is present : same as second example.

- A comment which contains expressions that sexually harass or humiliate : e.g., 이 짤보고 고로상 빙의됐다 애* 존맛겠다. (After seeing this meme, I was possessed with Goro and wanted to f*ck your m*ther.)

We define the terms in the guidelines to the participants well before the annotation work, and the annotators were trained follow them precisely. Through the pilot study, we analyze the types and set the following guidelines with examples to draw a borderline for the LIKELY class.

- A comment judged to have offensive intent due to intrinsic bias: e.g., 장애인들 몸비틀고 잘 노내. (Disabled people twist their bodies and play well.)

- A comment with disrespectful expressions (reproaching, admonishing, etc.): e.g., 신이 있어? 그딴건 없음. 정신차리고 사세요. (Did you say that there is a god? Nothing like that. Calm down.)

- A comment may cause shame: e.g., 섹*할때 필수용품.. (A must-have item for s*x.)

- A comment with a cynical tone: e.g., 조작일보 다운 기사네.. (It is an article from a daily forgery..)

- A comment considered to be implicitly disparaging: e.g., 근데 화*는 왜 빠는거야? (Why did you suck Hw*s*[14]?)

- A comment with abusive language but judged to be acceptable: e.g., 와 미친 개잘한다. (Wow, it's crazy, you are doing f*cking great!)

A.2 Abusive Language in KODOLI

We analyze comments with abusive labels, extract the profane term and hate term based on the frequency, and organize them into a bag of words. [15]

[14] Celebrity

[15] https://github.com/cardy20/KODOLI/tree/main/bow

A.3 Experimental Results on Each Auxiliary Task in the STL Settings

We evaluate both the auxiliary tasks, ALD and SA. Table 5 and Table 6 summarize the baseline results of the STL setup.

Model	Abusive Language Detection		
	P	R	F1
BiLSTM	89.03	87.27	88.05
CNN	90.53	88.22	89.22
KLUE-BERT	88.60	88.22	88.41
KLUE-RoBERTa	88.96	88.61	88.78
KoELECTRA	90.96	90.02	90.47

Table 5: Precision, recall, F1-score of abusive language detection

Model	Sentiment Analysis		
	P	R	F1
BiLSTM	73.31	72.16	72.61
CNN	74.32	73.46	73.81
KLUE-BERT	77.02	76.78	76.85
KLUE-RoBERTa	76.88	76.51	76.68
KoELECTRA	77.70	77.69	77.64

Table 6: Precision, recall, F1-score of sentiment analysis

A.4 Implementation Details

a. Hyperparameters: We used a batch size of 32 examples for each model and a fixed sentence length of 128. We used the AdamW optimizer (Loshchilov and Hutter, 2017). We set 48 seed and explored the learning rate to obtain the best results for each model. For CNN and BiLSTM, the learning rate was searched for between 1e-04, 2e-04, 3e-04, 4e-04, 5e-04, 6e-04, 7e-04. We searched for the following learning rates: 7e-06, 9e-06, 1e-05, 2e-05, 3e-05, 4e-05, for KLUE-BERT, KLUE-RoBERTa, and KoELECTRA. In the case of MTL, we initially set all lambda weights to 1.0. We searched for an appropriate lambda weight by using a grid search.

b. Training conditions: We implemented the model using PyTorch (Paszke et al., 2019) and used an NVIDIA GeForce RTX 3090 with 24 GB of VRAM to train all baseline models. We used the HuggingFace library for our BERT-family models[16].

Model	Task	Macro Average		
		P	R	F1
CNN	OLI	66.30	64.50	65.06
	OLI + ALD + SA	**67.42**	**67.03**	**66.33**
BiLSTM	OLI	65.86	64.61	65.12
	OLI + ALD + SA	**66.98**	**65.06**	**65.91**
KLUE-BERT	OLI	67.19	68.21	67.65
	OLI + ALD + SA	**68.12**	**69.17**	**68.53**
KLUE-RoBERTa	OLI	66.94	68.99	67.82
	OLI + ALD + SA	**68.10**	**70.22**	**68.67**
KoELECTRA	OLI	69.21	69.83	69.44
	OLI + ALD + SA	**69.93**	**71.06**	**70.26**

Table 7: STL(OLI) vs MTL(OLI+ALD+SA)

A.5 Experimental Results on the Baseline Models in the MTL Settings

Table 7 presents the experimental results obtained using KODOLI on the STL method for the OLI task and the MTL method combining the OLI task with auxiliary task 1 (ALD) and auxiliary task 2 (SA) in the five baseline models. This result indicates that the performance is improved when two auxiliary tasks are jointly learned in all baseline models.

A.6 Error Analysis Details

We conduct an in-depth analysis of the LIKELY class, which shows relatively low performance on classifiers, with auxiliary labels. Of the 718 examples of the LIKELY class in the validation set, 208 examples misclassified LIKELY as NOT and 197 LIKELY examples as OFFEN. Among the cases misclassified as NOT, 136 cases are labeled as non-abusive language, which means that they have no explicit expression (i.e., hate words, profane). We find that a large portion of the cases is sarcastically or twisted as considering the context of the sentence. Especially, if a comment is likely offensive under the social and cultural background (e.g., first and fourth examples in A.1), the distribution of prediction scores tends to appear evenly. In addition, most misclassified cases as OFFEN (72%) contain an explicit and emphasized expression. We conjecture that classifiers predict OFFEN by looking at the specific word itself. However, humans take it differently in feeling offended.

[16] https://huggingface.co/klue/bert-base