# Leveraging Pre-existing Resources for Data-Efficient Counter-Narrative Generation in Korean

**Seungyoon Lee**[1], **Chanjun Park**[2†], **DaHyun Jung**[1],
**Hyeonseok Moon**[1], **Jaehyung Seo**[1], **Sugyeong Eo**[1], **Heuiseok Lim**[1†]
[1]Korea University, [2]Upstage
[1]Seoul, Republic of Korea, [2]Gyeonggi-do, Republic of Korea
{dltmddbs100, dhaabb55, glee889, seojae777, djtnrud, limhseok}@korea.ac.kr
chanjun.park@upstage.ai

## Abstract

***Warning****: This paper contains content that may be offensive or upsetting.*
Counter-narrative generation, i.e., the generation of fact-based responses to hate speech with the aim of correcting discriminatory beliefs, has been demonstrated to be an effective method to combat hate speech. However, its effectiveness is limited by the resource-intensive nature of dataset construction processes and only focuses on the primary language. To alleviate this problem, we propose a Korean Hate Speech Counter Punch (KHSCP), a cost-effective counter-narrative generation method in the Korean language. To this end, we release the first counter-narrative generation dataset in Korean and pose two research questions. Under the questions, we propose an effective augmentation method and investigate the reasonability of a large language model to overcome data scarcity in low-resource environments by leveraging existing resources. In this regard, we conduct several experiments to verify the effectiveness of the proposed method. Our results reveal that applying pre-existing resources can improve the generation performance by a significant margin. Through deep analysis on these experiments, this work proposes the possibility of overcoming the challenges of generating counter-narratives in low-resource environments.

**Keywords:** Hate Speech, Counter-Narrative Generation, Dataset Construction

## 1. Introduction

As technological development expands accessibility to online spaces, hate speech emerges as a persistent problem that not only causes harm to its victims but also amplifies incorrect prejudices by inducing skewed social perception (Citron and Norton, 2011). To address this issue, numerous studies conducted in natural language processing (NLP) generally focused on constructing datasets and models for detecting or classifying hate speech (Xiang et al., 2012; Del Vigna et al., 2017; Caselli et al., 2021; Casula and Tonelli, 2023). Consequently, hate speech is usually suppressed by masking or prohibiting the use of core expressions identified by a classifier. However, these methods suffer from two main limitations.

Firstly, classification-based masking raises ethical concerns due to the potential suppression of the speaker's freedom of expression (Myers West, 2018)—given the right to speak freely, determination of hate speech is complicated by cultural and societal differences among different communities (Schmidt and Wiegand, 2017). Therefore, simply masking or filtering key expressions can amount to censorship and excessive blocking, as well as perpetuate this ambiguity. Secondly, although this approach can detect hate speech, it cannot identify or correct problems in speech specifically; thus, it does not contribute to preventing the fundamental cause of hate speech.

|  |
|---|
| **Hate Speech:** *Women are basically childlike, they remain this way most of their lives. Soft and emotional. It has devastated our once great patriarchal civilizations.* |
| **Counter-Narrative:** Soft and emotional are personality traits, may I suggest you read up on the #HeforShe movement that works for equal rights for men and women, and aims to stop the stigma around men showing emotions. |

Table 1: An example of hate speech-counter narrative pair.

Among the various attempts to solve these problems, counter-narrative generation is particularly effective, suppressing hate speech by correcting discriminatory perceptions voiced by the speaker using fact-based arguments (Benesch, 2014; Silverman et al., 2016). For example, as presented in Table 1, in response to hate speech contending that the purported soft and emotional character of women has destroyed patriarchal civilization, a counter-narrative can be constructed to argue that character is an individual trait, which promotes gender equality. In this way, counter-narratives can help mitigate hate speech by highlighting biased information and presenting a relative perspective based on reliable evidence.

Due to the nature of the counter-narrative, which requires going beyond simply correcting errors and constructing sentences based on credible ev-

† Corresponding author

*LREC-COLING 2024*, pages 10380–10392
20-25 May, 2024.

idence from experts, counter-narratives are collected through crowdsourcing or niche sourcing, which inevitably involve significant costs. This indicates that the construction of counter-narrative datasets is more difficult than that of hate speech datasets. In particular, in terms of language, the availability of counter-narrative datasets is severely limited, and these are only available in English, French, and Italian (Chung et al., 2019). On this point, even though some resources are available for the Korean language, there are few means to suppress hate speech, and the research in this area is relatively limited.

In this study, we propose Korean Hate Speech Counter Punch (KHSCP) that generates counter-narratives effectively in the Korean language. KHSCP involves an inexpensive pipeline that generates counter-narratives by constructing the first counter-narrative generation dataset in Korean and introducing an augmentation method using semantic similarity between sentences. Furthermore, we empirically explore GPT-3 (Brown et al., 2020) for automatic generation. To be specific, two questions are addressed to verify the utility of the proposed KHSCP. The first question is, *"Is it possible to enhance counter-narrative generation performance by utilizing only pre-existing resources?"* To answer the question, we propose a simple data augmentation technique that considers the semantic similarity between two sentences. Our experiments show that the method greatly improves performance. Finally, by conducting qualitative analysis on added pairs, we ensure that the suited pairs are being selected.

The second question is, *"Can a large language model be a reasonable way to generate suitable counter-narratives in Korean?"*. Although large language models such as GPT-3 (Brown et al., 2020) are considered useful in various downstream tasks, they lack validation in the field of counter-narrative generation, which is related to ethical concerns. To address this, we empirically investigate whether reasonable counter-narratives can be generated using existing resources in a low-resource environment and whether this can help enhance the performance of existing pre-trained models. Through our experiments, we examine the possibility of using an automatic method to generate counter-narrative datasets in Korean to alleviate data-construction limitations.

The main contributions of this study are threefold.

- We release the first counter-narrative generation dataset in Korean under the recipe called KHSCP, enabling the effective and low-cost generation of suitable counter-narratives by employing various hate expression resources.

- We propose a practical data augmentation technique using semantic search based on pre-existing hate speech datasets. We analyze the effects from both data and model perspectives and explore the conditions that optimize the utility of the augmentation.

- By conducting experiments on GPT-3, we demonstrate the appropriateness of large language models in the field of counter-narrative generation in Korean.

## 2. Related Work

As the scale of hate speech continues to grow, numerous social studies have demonstrated that counter-narratives can prevent the occurrence of hate speech (Benesch, 2014; Silverman et al., 2016; Schieb and Preuss, 2016; Stroud and Cox, 2018). Therefore, several studies have focused on constructing datasets using pre-trained models or manual labeling and emphasizing the importance of counter-narrative (Mathew et al., 2018, 2019; Chung et al., 2019; Tekiroğlu et al., 2020; Fanton et al., 2021; Chung et al., 2021a; Goffredo et al., 2022; Park et al., 2022; Lee et al., 2023b; Bengoetxea Azurmendi, 2023).

The first large expert-based counter-narrative dataset was constructed in English, French, and Italian (Chung et al., 2019). Niche sourcing from three non-governmental organization (NGO) operators was employed to construct the dataset. They are provided specific guidelines to ensure the annotation of counter-narrative sentences.

Due to the labor-intensive process, a data construction process that minimizes manual involvement by experimenting with a machine-based author-reviewer architecture was proposed (Tekiroğlu et al., 2020). They established an efficient pipeline to construct counter-narrative data using GPT-2 (Radford et al., 2019) as the author module, targeting crowd-sourced and niche-sourced data, and a BERT-like model (Devlin et al., 2019) as the reviewer module to evaluate the appropriateness of the pairs. Similarly, NGO operators were adopted to post-edit machine-generated counter-narratives using a human-in-the-loop strategy (Fanton et al., 2021). Other lines of works include the generation of counter-narratives considering facts and given knowledge (Chung et al., 2021b), the generation of diverse and relevant sentences via pruning and selection (Zhu and Bhat, 2021), and the investigation of the performance of pre-trained models with decoding strategy and unseen hate speech (Tekiroglu et al., 2022).

However, consideration for relatively minor languages such as Korean is lacking since these studies primarily focus on major languages. Although Park et al. (2022) collected the Korean counter-narrative dataset from social media, it contains aggressive and hostile expressions reflective of real-world online user language, which can risk perpetuating further hate and prejudice in a counter-narrative generation. As a pioneering effort, we release the first Korean counter-narrative dataset in rigorous scope for generation and propose methodologies to facilitate generation. In addition, we investigate the possibility of automatic dataset construction in Korean by utilizing large language models, such as GPT-3.

## 3. KHSCP

In this section, we introduce Korean Hate Speech Counter Punch (KHSCP) and discuss its methodology. Further, the configuration of validating experiments is also described. KHSCP comprises a series of recipes that effectively generate appropriate counter-narratives in Korean, constructing new datasets and utilizing existing hate speech resources for a generation. The overall process can be divided into two parts. First, a Korean counter-narrative dataset is constructed for counter-narrative generation training. Next, an augmentation technique is used to generate counter-narratives at low cost by utilizing existing hate speech datasets.

To be specific, augmentation is performed in two separate cases. In the first case of similarity-based matching, which considers semantic embedding of sentences, counter-narratives of similar existing hate speech are adopted as new pairs. The second case is based on prompting, using GPT-3 to generate new sentences from a different perspective to increase the number of pairs. It can enable the automatic construction of a Korean dataset while also leveraging the potential of various hate speech datasets in Korean. This enhances counter-narrative generation performance, which is verified experimentally.

### 3.1. Dataset Construction

**Source Data** A text corpus is created by compiling text pairs from an English hate speech counter-narrative dataset. The multitarget CONAN dataset (Fanton et al., 2021) is used as the source dataset. Although other datasets are also available, this one is selected to obtain text with a multi-target configuration toward hate speech collected corresponding to various targets in a human-in-the-loop manner. In addition, this dataset contains a relatively large amount of high-quality sentences.

**Automatic Translation** The selected English pairs are translated into Korean using the Naver Papago API (Lee et al., 2016), which has been used in numerous researches due to its proficiency in Korean-English translation (Lee et al., 2020; Bae et al., 2020; Cha et al., 2022). In aggregate, 5,003 pairs are translated; the categories of the constructed dataset are listed in Table 2. The constructed dataset contains the same hate targets as the original dataset, i.e., MUSLIMS, MIGRANTS, WOMEN, LGBT+, JEWS, POC, OTHER, and DISABLED. The dataset is divided into train, validation, and test sets in an 8:1:1 ratio, comprising 4,002, 500, and 501 sentence pairs, respectively. Our dataset can be downloaded at the following link: https://github.com/dltmddbs100/KHSCP.

| Rank | Target Group | Counts(%) |
|---|---|---|
| 1 | MUSLIMS | 1,335 (26.68) |
| 2 | MIGRANTS | 957 (19.13) |
| 3 | WOMEN | 662 (13.23) |
| 4 | LGBT+ | 617 (12.33) |
| 5 | JEWS | 594 (11.87) |
| 6 | POC | 352 (7.04) |
| 7 | OTHER | 266 (5.32) |
| 8 | DISABLED | 220 (4.4) |

Table 2: Target group distribution of the constructed dataset.

### 3.2. Counter-Narrative Augmentation

We explore two counter-narrative augmentation techniques to improve generation performance under limited resources. Although the processes are different, they both utilize existing hate speech to augment data.

**Semantic-Based Augmentation (SBA)** SBA leverages existing Korean datasets to expand limited pair sizes. It is based on the assumption that a counter-narrative may involve the same sentence for semantically similar hate expressions. To this end, pre-existing datasets provide a wide range of candidate hate speech, and we search for utterances similar to the original dataset against it.

We adopt semantic search as a core method. It requires an 'input sentence' as the input target and a 'candidate sentence' as the target of the search, which is taken to be the corresponding utterance. It assumes a set of paired datasets, $D = (h_i, c_i)_{i=1}^{m}$, where $h_i$ denotes the hate speech used as the query sentence, and $c_i$ denotes the corresponding counter-narrative. An input sentence, $x_j$, is an unpaired hate speech of a pre-existing dataset aimed at classifying aversion. We define the similarity function as follows:

$$\underset{i}{\operatorname{argmax}}\ \mathrm{sim}\left(f(h_i), f(x_j)\right) \tag{1}$$

where $f$ denotes a pre-trained encoder model that identifies the highest score for $i$ by calculating the cosine similarity between the representations of $h_i$ and $x_j$ for all $j$, as described in (1). For $f$, KoSimCSE-RoBERTa-multitask[1], which exhibits the best performance among various Korean sentence embedding models, is utilized. We apply CLS-pooling as the pooling strategy and set the maximum length of 128 when measuring the similarity between external data and hate speech obtained from constructed data.

For all $x_j$, we identify the most similar $h_i$ among candidate sentences and select only sentences with scores exceeding an appropriate threshold. While the category organization of the pre-existing dataset may not match the paired dataset, setting a threshold of similarity to filter out the low points can compensate for the inconsistency in categories and produce high-quality counter-narrative pairs with semantically similar meanings compared to the candidate sentences.

**Prompt-Based Augmentation (PBA)** PBA generates sentences by leveraging the capabilities of a large language model. While this technique might be conventionally used in various other fields (Yoo et al., 2021; Lee et al., 2023a; Dai et al., 2023), considering the uniqueness of counter-narrative generation, its efficacy remains untested in this domain. We focused on empirical factors that enhance scalability and applicability within Korean.

PBA differs from SBA as it does not rely on the semantic embedding of pre-trained models. Simple prompts are designed to generate counter-narratives in a zero-shot environment. The generated speech is used to construct additional pairs to fine-tune the pre-trained model, enabling the model to acquire a broader range of expressions and achieve comprehensive learning.

## 3.3. Available Pre-existing Datasets

Four Korean hate speech datasets are used as external resources for augmentation. Each is constructed to detect the offensiveness of sentences, and they involve different construction methods, categories, and collection paths. In subsequent experiments, we analyze each augmentation case. The following datasets are used:

**Unsmile** Unsmile (Kang et al., 2022) is a multi-label Korean online hate speech dataset collected from news and online community sites of major Korean web portals. It comprises 10,139 hate expressions, 3,929 swear words, and normal expressions. Each sentence is classified by three workers and verified by five hate expression experts. Nine different hate categories are used for labeling, including woman/family, man, sexual minority, race/nationality, etc.

**APEACH** APEACH (Yang et al., 2022) is a dataset constructed by collecting data using a pseudo classifier and post-labeling to ensure diversity and balance of topics in the data collection and decrease overlap between training sets. It comprises ten different categories and is collected from unspecified users via crowdsourcing. Currently, only a subset of this dataset that corresponds to the test set is available to the public, and we use this subset in this study.

**BEEP** BEEP (Moon et al., 2020) is the first Korean hate speech dataset collected via crowdsourcing, which consists of comments on online news platforms in Korea. It contains 9,381 labeled sentences, primarily collected based on societal prejudices and hatred. Three categories are considered—hate, defined as speech with a target and purpose that is heavily biased by societal prejudices; offensive, characterized by a weak level of toxicity, such as sarcasm or opinions; and none, which does not correspond to any hate speech.

**KOLD** KOLD (Jeong et al., 2022) is the first Korean dataset of offensive expressions that adopts a hierarchical taxonomy collected from the titles and comments of Naver news and YouTube videos. It consists of 40,429 sentences alongside their offensiveness levels and the locations of text spans containing offensive expressions with their targets and types. However, only 20,310 data are available to the public; therefore, we only use this subset.

The publicly available portions of the aforementioned datasets are used. We use raw datasets, and sentences classified under the 'none' category for offensiveness are excluded during our experiments.

## 3.4. Verification of KHSCP

To verify the utility and validity of KHSCP, two research questions are considered under various experiments.

- **Question 1**: Is it possible to enhance counter-narrative generation performance by utilizing only pre-existing resources?
- **Question 2**: Can a large language model be a reasonable way to generate suitable counter-narratives in Korean?

The first research question aims to evaluate the effectiveness of the proposed augmentation technique by comparing the results of several Korean

[1] https://github.com/BM-K/Sentence-Embedding-is-all-you-need#korean-sentence-embedding

| Model | Dataset | # Pairs / # Augmented | Generation Metric | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | | | B-1 | B-3 | B-4 | R-1 | R-2 | R-L | |
| KoGPT2 | Baseline | 4,002 | 14.449 | 4.877 | 2.936 | 23.745 | **6.618** | 18.628 | 11.876 |
| | w/ Unsmile | 4,002 / 3,986 | 19.488 | 5.842 | 3.307 | 24.057 | 5.765 | 18.184 | 12.774 |
| | w/ APEACH | 4,002 / 714 | 17.818 | 5.683 | 3.296 | 24.023 | 6.490 | 18.631 | 12.657 |
| | w/ BEEP | 4,002 / 587 | 18.237 | 5.916 | 3.439 | 24.317 | 6.384 | **18.801** | 12.849 |
| | w/ KOLD | 4,002 / 6,208 | **20.211** | **6.305** | **3.609** | **24.332** | 6.321 | 18.634 | **13.235** |
| KoBART | Baseline | 4,002 | 17.236 | 6.307 | 3.915 | **26.186** | **8.097** | **20.457** | 13.700 |
| | w/ Unsmile | 4,002 / 3,986 | **18.913** | **6.867** | **4.334** | 26.088 | 7.993 | 20.066 | **14.044** |
| | w/ APEACH | 4,002 / 714 | 17.004 | 6.261 | 3.983 | 25.179 | 7.779 | 19.883 | 13.348 |
| | w/ BEEP | 4,002 / 587 | 18.004 | 6.512 | 4.102 | 25.816 | 7.986 | 20.125 | 13.758 |
| | w/ KOLD | 4,002 / 6,208 | 18.527 | 6.425 | 3.930 | 25.253 | 7.647 | 19.435 | 13.536 |
| mT5 | Baseline | 4,002 | 22.131 | 7.866 | 4.874 | 27.288 | **8.603** | 20.302 | 15.177 |
| | w/ Unsmile | 4,002 / 3,986 | **25.410** | **8.893** | **5.463** | 27.804 | 8.274 | **20.379** | **16.037** |
| | w/ APEACH | 4,002 / 714 | 23.641 | 8.132 | 4.960 | 27.010 | 8.114 | 20.001 | 15.310 |
| | w/ BEEP | 4,002 / 587 | 24.754 | 8.701 | 5.263 | **27.810** | 8.441 | 20.525 | 15.916 |
| | w/ KOLD | 4,002 / 6,208 | 23.597 | 8.104 | 4.830 | 27.023 | 8.239 | 20.053 | 15.308 |

Table 3: Automatic evaluation results on the test set. The number of pairs used for augmentation is shown, and baseline refers to the case of fine-tuning with the original training set without augmentation.

hate speech datasets. SBA is applied to the input sentences of the four selected datasets. In SBA, we utilize the four selected datasets as input sentences to retrieve similar pairs. The similarity threshold is set to a constant value to get proper sentences, and only pairs with similarity exceeding this threshold are selected and added to the training set. Then, the models are trained using these additional pairs. In addition, we investigate the factors that lead to the best performance and determine the most effective similarity threshold. The effect of the quality of the input data on performance with respect to varying threshold values is also evaluated. The threshold values range from 0 to 0.8 and exclude ranges with an extremely small dataset size.

Our second question relates to the reasonability in the generation of Korean counter-narratives by PBA. In PBA, we leverage GPT-3's zero-shot ability to generate counter-narratives for a given resource. We conduct quantitative validation of the correspondence by injecting it as augmented data into the pre-trained models. External hate speech is handled using Unsmile, and the actual prompt used in our experiment is the same one used in prior research (Ashida and Komachi, 2022), translated into Korean.

# 4. Experimental Settings

## 4.1. Models

To answer the first question, three Transformer-based language models (Vaswani et al., 2017) are considered. Two of these are Korean models, and the remaining one is a multilingual model. The Korean models are KoGPT2 [2] and KoBART [3], which comprise decoder and seq2seq structures, respectively. The multilingual model is mT5-base (Xue et al., 2021), trained in 101 languages, including Korean. Sentences are generated via greedy sampling using fine-tuned models, and no more than three identical n-grams are generated to avoid excessive overlapping between tokens.

For the second question, GPT-3 text-davinci-003 is used. While the other latest LLMs are trending towards widespread adoption due to their robust performances, given the nature of this field, the generalization of models that have not been accounted for in previous research and lack of thorough validation might invite more ethical concerns. In fact, many studies have raised concerns about ethical issues and hallucinations with recent LLMs. Based on these considerations, we employ GPT-3.

## 4.2. Evaluation

Overlap metrics are used to evaluate the quality of generated counter-narratives. According to previous research (Qian et al., 2019; Chung et al., 2021b), we employed BLEU (Papineni et al., 2002) variants: BLEU-1 (B-1), BLEU-3 (B-3), and BLEU-4 (B-4) with ROUGE (Lin, 2004) variants: ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) to evaluate the completeness of the generated sentences automatically. Both metrics measure the degree of overlap between the source and target sentences based on n-grams. Since postpositions are directly coupled with words morphologically in the Korean language, the subdivided meanings of words will likely be overlooked when they are split by spac-

[2] `https://github.com/SKT-AI/KoGPT2`
[3] `https://github.com/SKT-AI/KoBART`

| Model | # Pair / # Augmented | Threshold | Generation Metric | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | | | B-1 | B-3 | B-4 | R-1 | R-2 | R-L | |
| KoBART | 4,002 / 14,068 | 0 | 16.643 | 5.361 | 3.166 | 24.135 | 6.599 | 18.728 | 12.439 |
| | 4,002 / 13,380 | 0.4 | 18.115 | 5.960 | 3.552 | 24.417 | 7.087 | 18.829 | 12.993 |
| | 4,002 / 9,730 | **0.5** | **20.716** | **7.172** | **4.442** | **26.632** | 7.922 | **20.092** | **14.496** |
| | 4,002 / 3,986 | 0.6 | 18.913 | 6.867 | 4.334 | 26.088 | **7.993** | 20.066 | 14.044 |
| | 4,002 / 657 | 0.7 | 17.231 | 6.170 | 3.850 | 25.523 | 7.848 | 19.949 | 13.429 |
| | 4,002 / 46 | 0.8 | 16.784 | 5.989 | 3.768 | 25.036 | 7.598 | 19.531 | 13.118 |
| mT5 | 4,002 / 14,068 | 0 | 21.795 | 6.973 | 4.021 | 25.860 | 7.392 | 19.282 | 14.221 |
| | 4,002 / 13,380 | 0.4 | 20.530 | 6.567 | 3.704 | 25.289 | 7.398 | 19.186 | 13.779 |
| | 4,002 / 9,730 | 0.5 | 22.962 | 7.597 | 4.503 | 26.754 | 7.866 | 19.705 | 14.898 |
| | 4,002 / 3,986 | **0.6** | **25.410** | **8.893** | **5.463** | 27.804 | 8.274 | 20.379 | **16.037** |
| | 4,002 / 657 | 0.7 | 23.466 | 8.239 | 5.001 | 27.696 | 8.790 | 20.775 | 15.661 |
| | 4,002 / 46 | 0.8 | 23.329 | 8.410 | 5.217 | **27.998** | **8.858** | **21.012** | 15.804 |

Table 4: The distribution of generation performance change according to the threshold with Unsmile dataset.

| Model | Generation Metric | | | | | | Average |
|---|---|---|---|---|---|---|---|
| | B-1 | B-3 | B-4 | R-1 | R-2 | R-L | |
| KoBART* | 20.716 | 7.172 | 4.442 | 26.632 | 7.922 | 20.092 | 14.496 |
| mT5 | 22.131 | 7.866 | 4.874 | 27.288 | 8.603 | 20.302 | 15.177 |
| w/ GPT-3 | 24.13 | 8.714 | 5.426 | **27.830** | **8.782** | **21.099** | 15.997 |
| mT5* | **25.410** | **8.893** | **5.463** | 27.804 | 8.274 | 20.379 | **16.037** |

Table 5: Generation performance between existing models and augmented by PBA. * indicates the best case from the previous experiment, otherwise indicates the baseline.

ing (Lee and Shin, 2021). Accordingly, Mecab-ko [4], which is frequently used in various NLP tasks in Korean, is used as the tokenizer during evaluation (Park et al., 2020).

## 4.3. Training Details

We use transformers (Wolf et al., 2019) for our experiments. All models are trained on a training set of five epochs with a batch size of 64, and weights are saved at the end of each epoch. The weight with the lowest validation loss among all epochs is utilized for generation. We employ the AdamW optimizer with a learning rate of 5e-5 and a weight decay of 1e-3. A label smoothing factor of 0.1 and a linear learning rate scheduler are employed. We use a single NVIDIA RTX A6000 GPU for training and apply 16-bit precision for fast training.

For GPT-3 text-davinci-003 model, we use temperature of 0.7, max tokens of 512, top_$p$ of 1, frequency penalty of 0, and presence penalty of 0.

## 5. Experimental Results

In this section, we discuss the experimental results in relation to the two aforementioned research questions.

### 5.1. Is it possible to enhance counter-narrative generation performance by utilizing only pre-existing resources?

Our main results are presented in Table 3. As applying SBA, the generation score increases in most cases compared to the baseline. However, the increment depends on the type of dataset and model used, particularly on the quality of the data included and the size of the added pairs. In addition, from the model's perspective, large differences are observed depending on the structure. Therefore, we analyze the results of the experiment by dividing them into data and model-centric aspects. Then, we identify optimal conditions in which the effects of augmentation can be maximized.

**Insight from Data-Centric Approaches** As shown in Table 3, Unsmile accounts for the second-highest performance increase in the case of KoGPT2 after KOLD, and for the highest performance in the case of mT5 and KoBART. It is the only dataset among the four that includes socially challenged communities among its sources. As a result, less refined and more provocative words are included compared to the other datasets. This exposes the model to a wider diversity of hate expressions. Since these expressions are matched with proper counter-narratives, the model learns more robustly, owing to numerous opportunities to learn similar but richer representations. All models exhibit high-performance improvement on BEEP despite their relatively low stature compared to the other datasets. BEEP also exhibits low refinement and high toxicity and is presumed to exhibit a similar pattern compared to Unsmile due to its high diversity of expressions.

On the other hand, APEACH does not induce significant performance improvements—in fact, a performance degradation is observed for KoBART.

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

| Source | Hate Speech | Matched Counter-Narrative | Score |
|---|---|---|---|
| Unsmile | (a) 동성애는 자연의 이치을 벗어난 행동 절대로 정상인으로는 봐서 안됀다!!<br>Homosexuality is a behavior that goes against the grain of nature and should never be seen as normal! | (a) 당신은 불필요한 편견과 무지를 영구화하는 대신에 진화와 우리 종의 기원에 대해 정말로 배워야 합니다.<br>You should really learn about evolution and the origin of our species, instead of perpetuating unwarranted prejudice and ignorance. | 0.7771 |
|  | (b) 외노자새끼들 범죄율보면 개씨발새끼들임을알수있음<br>You can tell foreign worker assholes are assholes by the crime rate. | (b) 그들은 대부분 노동자들이고, 그들 대부분은 개발도상국 출신이다. 그들은 경제에 기여하고, 집을 짓고, 가족을 부양합니다.<br>Mostly they are workers, the majority of them are from developing countries. They contribute to the economy, they build houses, they raise families. | 0.6881 |
|  | (c) 중국그지새끼들아. 4대보험 니가 안 내면 내가 더 내야 된다. 끄지라.<br>Chinese bastards, if you don't pay, I should pay more. Go away. | (c) 건강한 사람이 기부하는 것을 제한하는 것은 차별이라는 것을 알아야 한다.<br>Limiting any healthy individual from donating is discriminatory, you should know that. | 0.3852 |
| APEACH | (a) 장애인들은 무슨 정신병자 같아!<br>Disabled people are like psychos! | (a) '신경전형'의 관점에서, 이것은 항상 사실인 것은 아니다: 예를 들어 윌리엄스 증후군을 가진 사람들은 평균 지능보다 낮지만 더 높은 공감을 보인다.<br>From a 'neurotypical' perspective, this is not always true: for example people with Williams syndrome have below average intelligence but show higher empathy. | 0.7901 |
|  | (b) 요즘은 동성을 좋아하는 사람이 상당히 많아진 것 같더라구요, 나중에 아들을 낳았는데 남자와 결혼한다고 하면 어떨지 생각만 해도 아찔하네요<br>It seems like there are quite a few people who like the same sex these days, and I shudder to think what it will be like when my son says he's marrying a man. | (b) 모든 인간은 성적 지향에 관계없이 삶의 권리를 가지고 있다.<br>Every human being has the right to life, regardless of sexual orientation. | 0.6656 |
|  | (c) 18살이 뭘 안다고 투표를 하냐?<br>What do 18-year-olds know about voting? | (c) 모든 영국 시민은 자신이 원하는 대로 살고 투표할 권리가 있다. 무슬림이 되는 것도 예외가 아니다.<br>Every British citizen has the right to live and vote as he/she wish. Being Muslims is no exception. | 0.3775 |
| BEEP | (a) 검은머리 외국인 극혐..이나라에서 사라져라<br>Dark-haired foreigners, get out of this country. | (a) 우리는 인종에 따라 사람들을 차별하는 것을 멈춰야 한다. 인종 멸절이 왜 우리 문제의 해결책이 되는지 이해가 안 된다.<br>We should stop discriminating against people based on their race. I find it hard to understand why the extermination of a race would be a solution for our problems. | 0.7123 |
|  | (b) 페미니즘 = 정신병 임을 또 한번 느낀다<br>I am reminded that feminism is a mental illness again. | (b) 페미니즘은 증오가 아니라 평등을 향한 운동이다.<br>Feminism is a movement towards equality, not hate. | 0.6488 |
|  | (c) 열폭 오지네 꼰대들;;<br>A lot of inferiority complexes, condescending old man. | (c) 흑인이 되는 것은 부끄러운 일이 아니다. 부끄러운 것은 유색인종에 대한 존경심이 부족하다는 것이다.<br>There is no shame in being black, what is shameful is the lack of respect shown to people of colour. | 0.3702 |

Table 6: Counter-narratives for each dataset selected via SBA.

APEACH contains hate speech artificially generated by unspecified individuals rather than being collected online. Thus, its content is often too specific or exotic, and its level of refinement is also the highest among the four datasets. Therefore, the degradation can be attributed to the limited diversity of expressions encountered by the model and domain mismatch with the training set. However, augmentation is effective in most cases, indicating the viability of the proposed method and that it exhibits a high degree of versatility applicable to general hate speech resources.

**Insight from Model-Centric Approaches** Among the three models, KoGPT2 exhibits consistent improvement on all datasets. The best score is obtained when augmentation is performed on KOLD, with an average score being approximately 11% higher than the baseline, and the greatest improvement among all models is observed. However, in all cases, its absolute value is the lowest among those of the models, which can be attributed to the structure of the model. Seq2seq models contain bidirectional encoders that can capture input information; however, single decoders have a lower ability and require more data. Their data dependency is relatively high—thus, performance quality is observed to be proportional to the number of pairs used during augmentation.

In contrast to KoGPT2, which consistently exhibits performance improvement irrespective of the dataset, KoBART does not exhibit significant deviation in performance quality with respect to data augmentation. On the other hand, mT5 exhibits improved performance on all datasets, with the average score being two points higher than KoBART. This suggests that the text-to-text configuration is effective in this task, and the multilingual model can achieve sufficiently high performance even if it is not trained in a single language. Furthermore, both models exhibit their best performances on Unsmile, followed by BEEP. The pair sizes of the two datasets are significantly different—Unsmile is approximately six times larger than BEEP. Despite its relatively small size, BEEP yields a large performance improvement, indicating that the quality of input data is as important as the amount when it comes to the performance of models.

In this experiment, a constant threshold value of 0.6 is used to ensure comparison on the same criteria. However, the appropriate threshold value required for each model may differ, and the amount and quality of augmented data may vary accordingly. Therefore, in the following experiments, we increase the practicality of this method by identify-

|  |  |
|---|---|
| HS | 사실상 여자가 남자를 이길 수 있는 신체활동은 섹스뿐 ㄹㅇ<br>In fact, the only physical activity where a woman can beat a man is sex. |
| GPT-3 | 여자가 남자보다 잘할 수 있는 다양한 신체활동이 있습니다. 여자는 배구, 축구, 골프 등 스포츠 활동에서 남자와 같은 수준의 성과를 내는 것이 가능합니다. 여자는 다양한 방법으로 각자의 역량을 최대한 발휘할 수 있으며, 이는 섹스뿐만 아니라 다른 신체활동에서도 가능합니다.<br>There are many physical activities that women can do better than men. It is possible for women to perform at the same level as men in sports activities such as volleyball, soccer and golf. Women can reach their full potential in a variety of ways, not only in sex but also in other physical activities. |
| HS | 이슬람 전통에는 가족이 살해당하면 복수하는 것이 정당화 된다. 이슬람 지역에서 자폭테러가 끊이지 않는 것도 다 이런 전통때문이다. 끊임없는 복수의 자폭테러와 내전이 계속될 거다.<br>In Islamic tradition, revenge is justified when a family member is killed. It is because of this tradition that suicide bombing terrorism does not cease in the Islamic region. Endless revenge suicide bombing terrorism and civil war will continue. |
| GPT-3 | 이슬람 전통에서는 복수가 금지되어 있습니다. 한 사람의 죄에 대해 모든 사람들이 책임을 지는 것은 매우 틀린 것입니다. 이런 개인적인 복수가 자폭테러로 이어지는 것을 방지하기 위해서는 보다 폭넓은 공감과 협력이 필요합니다.<br>Revenge is forbidden in Islamic tradition. It is very wrong to hold everyone accountable for one man's sin. Broader empathy and cooperation are needed to prevent this personal vendetta from leading to a suicide bombing. |

Table 7: Examples of generated counter narrative against hate speech in Unsmile.

ing the optimal threshold to maximize the effect of SBA. Furthermore, we analyze the change in the amount and quality of data.

**Dependency on similarity threshold** We adjust the threshold value to the dataset on which the best effect is obtained for seq2seq models. Table 4 shows the experimental results for varying similarity thresholds. As the threshold increases, a small amount of high-quality data is included, whereas a large amount of low-quality data is included as the threshold decreases. KoBART exhibits an average score increase of approximately 0.8 when the threshold is 0.5, corresponding to 9,730 augmented pairs. The average score of mT5 increases by approximately 1, corresponding to 3,986 augmented pairs, in the best performance improvement. Notably, KoBART's performance is degraded when augmentation is applied, except for the cases with 0.5 and 0.6 as the threshold. When the threshold exceeds 0.6, the performance is degraded even though high-quality data are obtained. This suggests that the amount of data used for augmentation is an important factor. Similarly, when the threshold is less than 0.5, the score drops by a significant margin. In this regard, the quality of the added dataset is also observed to play an important role in augmentation, which is discussed in Section 6.1.

Similar results are observed for mT5. While it performs better than the baseline corresponding to a threshold exceeding 0.6, the performance drops when augmentation is performed with more data with slightly lower quality, i.e., when the threshold is lower than 0.5.

In other words, the performance of the proposed augmentation method is greatly affected by the size and quality of the dataset. Although a good alignment between hate expressions and counter-narratives is necessary, it is also important for the model to be exposed to a broader range of sentences in the original training set. Even if they are slightly less similar, this endows the model with a wider perspective and increases its robustness.

### 5.2. Can a large language model be a reasonable way to generate suitable counter-narratives in Korean?

We quantitatively verify that the sentences generated by GPT-3 are suitable for Korean counter-narrative generation by utilizing them as additional pairs.

To ensure equal comparison with mT5, which exhibits the best performance in the previous experiment, GPT-3 generates counter-narratives using the same data when Unsmile's similarity threshold is 0.6 and trained with mT5. The experimental results are presented in Table 5. When augmented with GPT-3, significant improvement is observed in all metrics. In addition, the difference in average score compared to mT5 is only approximately 0.04, and even the rouge score is higher. SBA increases the range of hate speech learned by the model, but the range of the counter-narrative remains identical. However, GPT-3 generates new sentences from another perspective, expanding both ranges and allowing the model to consider more diverse viewpoints. In this flow, the improvement is attributed to the diversity and correspondence of generated counter-narratives for each hate speech, which positively affects training. An increase in generation performance implies that GPT-3 has reasonable intrinsic knowledge to recognize hate speech and correct discriminatory beliefs in Korean.

## 6. Qualitative Analysis

We qualitatively analyze the appropriateness of the sentences matched by the SBA and the generated counter-narratives by the PBA.

### 6.1. Reliability of SBA

We show each dataset's counter-narratives matched to existing hate speech in Table 6. A high similarity score indicates a high semantic agreement with existing hate expressions. SBA has enough ability to filter corresponding counter-narratives based on the similarity. Examples with similarity scores higher than 0.5, which are (a) and (b) in the table, show that suitable counter-narratives are selected for the hate speech's target and context. On the other hand, cases lower than 0.4, which are (c), show that utterances that misidentify the hate target or are not appropriate for the context are selected. When used for augmentation, these sentences act as noise and degrade performance. This is consistent with our earlier experiments and demonstrates that SBA with proper cutoff can be a method for data-efficient generation.

### 6.2. Quality of Generation from PBA

Table 7 presents counter-narratives generated by GPT-3. At the top of the table, hate speech involving sexual harassment of women is presented. In response, GPT-3 generates a statement that provides specific examples of various physical activities to adjust the speaker's discriminatory perception. Similarly, for the sentences denigrating religion presented at the bottom of the table, GPT-3 argues the injustice of statements by suggesting the concept of "pluralism" and claims that public understanding and cooperation are necessary to prevent this from escalating into terrorism. Even if GPT-3 is not used as an augmentation technique, the model generates logical and closely related counter-narratives to the real world. We provide more examples in Appendix A.

## 7. Conclusion

In this study, we propose KHSCP, a Korean counter-narrative generation recipe that leverages existing resources to improve generation quality without incurring additional costs. We release the first Korean counter-narrative generation dataset and analyze the performance by considering two questions. We demonstrate that the augmentation method proposed in KHSCP significantly enhances the performance of all models. Furthermore, we empirically assess the capability of a large language model in generating appropriate counter-narratives in Korean against hate speech. This suggests that KHSCP can compensate for the difficulty of counter-narrative generation and dataset construction limitation in Korean, where resources to counter hate are relatively scarce. As a future work, we plan to introduce an automatic process that reduces the construction cost of the Korean dataset reflecting social contexts.

## 8. Limitations

We build the first Korean counter-narrative dataset as a part of KHSCP. Due to the unavailability of human resources, we rely on a translation process to gather a suitable corpus. We recognize that translation may not fully capture the cultural specificity and nuances between languages. The inherent variations in cultural norms may not be adequately represented through translation alone. However, given that there are currently no counter-narrative resources that address hate as a generative problem in Korean, we think taking the first step to address this universal hate is also important. As a next step, we plan to employ a more culturally sensitive approach to dataset construction.

Furthermore, our proposed method is currently limited to the Korean language. Although we have achieved promising results in this context, extending our approach to other languages remains an important avenue for future exploration. To enhance the generalizability and applicability of our method, efforts should be made to adapt and validate it across various languages, considering the unique challenges and characteristics specific to each language.

Lastly, while we make efforts to incorporate as many available Korean hate speech datasets as possible, we acknowledge that we may not have captured the entirety of the hate speech landscape in the language. We plan to include a broader range of Korean hate speech datasets to improve the comprehensiveness and representativeness of the data.

## 9. Ethical Statement

Research on hate speech is a sensitive area dealing with ethical issues. Generating counter-narratives has been found to be effective in reducing hate speech, but it may cause unintended results or introduce bias when used in other situations, such as training language models or releasing data. To address these concerns, we have taken care to only use publicly available data for both training and augmenting our model. This ensures that the counter-narratives produced by our model are appropriate and respectful and do not perpetuate any harmful stereotypes or biases. Furthermore, we have taken steps to ensure that the data used to train and augment our model is diverse and representative of a wide range of perspectives and voices. This helps to minimize the risk of introducing bias or perpetuating harmful stereotypes in the counter-narratives generated by the model.

## 10. Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A5A7026673). Following are results of a study on the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea.

## 11. Bibliographical References

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. *WOAH 2022*, page 11.

Kang Il Bae, Junghoon Park, Jongga Lee, Yungseop Lee, and Changwon Lim. 2020. Flower classification with modified multimodal convolutional neural networks. *Expert Systems with Applications*, 159:113455.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

Jaione Bengoetxea Azurmendi. 2023. Basque and spanish counter narrative generation: Data creation and evaluation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25.

Camilla Casula and Sara Tonelli. 2023. Generation-based data augmentation for offensive language detection: Is it worth it? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3351–3369.

Junyeop Cha, Seoyun Kim, and Eunil Park. 2022. A lexicon-based approach to examine depression detection in social media: the case of twitter and university community. *Humanities and Social Sciences Communications*, 9(1):1–10.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914.

Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91:1435.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240.

Pierpaolo Goffredo, Valerio Basile, Bianca Cepollaro, Viviana Patti, et al. 2022. Counter-twit: An italian corpus for online counterspeech in ecological contexts. In *WOAH 2022-6th Workshop on Online Abuse and Harms, Proceedings of the Workshop*, pages 57–66. Association for Computational Linguistics (ACL).

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Haeyun Oh. 2022. Kold: Korean offensive language dataset. In *The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. EMNLP.

TaeYoung Kang, Eunrang Kwon, Junbum Lee, Youngeun Nam, Junmo Song, and JeongKyu Suh. 2022. Korean online hate speech dataset for multilabel classification: How can social science improve dataset on hate speech? *arXiv e-prints*, pages arXiv–2204.

Daeun Lee, Soyoung Park, Jiwon Kang, Daejin Choi, and Jinyoung Han. 2020. Cross-lingual suicidal-oriented word embedding toward suicide prevention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2208–2217.

Hyoung-Gyu Lee, Jun-Seok Kim, Joong-Hwi Shin, Jaesong Lee, Ying-Xiu Quan, and Young-Seob Jeong. 2016. papago: A machine translation service with word sense disambiguation and currency conversion. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 185–188, Osaka, Japan. The COLING 2016 Organizing Committee.

Sangah Lee and Hyopil Shin. 2021. The korean morphologically tight-fitting tokenizer for noisy user-generated texts. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 410–416.

Seungjun Lee, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2023a. Improving formality-sensitive machine translation using data-centric approaches and prompt engineering. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 420–432, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Seungyoon Lee, Dahyun Jung, Chanjun Park, Seolhwa Lee, and Heuiseok Lim. 2023b. Alternative speech: Complementary method to counter-narrative for better discourse. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1438–1442. IEEE.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter-speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

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.

Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Hayul Park, HyunA Park, and Sanghoun Song. 2022. Countering the hatred: The counter-speech dataset in korean for evaluating hate speech detection models. *Discourse and Cognition*, 29(2):1–23.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various korean nlp tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Tanya Silverman, Christopher J Stewart, Zahed Amanullah, and Jonathan Birdwell. 2016. The impact of counter-narratives. *Institute for Strategic Dialogue*, 54.

Scott R Stroud and William Cox. 2018. The varieties of feminist counterspeech in the misogynistic online world. In *Mediating Misogyny*, pages 293–310. Springer.

Serra Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Kichang Yang, Wonjun Jang, and Won Ik Cho. 2022. APEACH: Attacking pejorative expressions with analysis on crowd-generated hate speech evaluation datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7076–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149.

# A. Examples of Counter-Narratives produced by GPT-3

| Type | Hate Speech | Counter-Narrative |
|---|---|---|
| Women/Family | 경찰도되고 군인도되고 소방관도되고 여자도 똑같이 다 할수 있다고 하지만 여자가 약해서 당하는건 어쩔수 없다는건 이런데서 확연히 표시난다는거<br>It is said that women can be police, soldiers, firefighters, and women can do the same as men, but it is inevitable that women are weak and suffer from it. | 여자가 약하다는 건 절대로 사실이 아니에요. 여자들은 생각보다 강하고 다양한 일에 능하기 때문에 이런 혐오 표현은 절대로 나쁘죠.<br>It is not true that women are weak. Women are stronger than you think and are good at various things, so this kind of hate speech is absolutely bad. |
| | 돈 벌려고 갈등 조장하는 놈들, 무슨 여성인권단체 같은 것들이 선두에서 지휘하고 현실 불만인 애들이 난리치는게 페미 아닌가.<br>Those who instigate conflicts to make money, some kind of women's rights organization, lead the way, and it's feminism that kids who are dissatisfied with reality are in a frenzy. | 이런 잘못된 개념을 배우지 않고, 모든 사람이 다 동등한 인권과 성평등을 존중하는 사회를 지향해야 합니다. 이는 여성인권단체가 선도하고 있는 노력이며, 모든 사람이 함께 참여해 손잡고 같이 일해야 합니다.<br>Without learning these wrong concepts, we must aim for a society where everyone respects equal human rights and gender equality. This is an effort led by women's rights organizations, and everyone should be involved and work hand in hand. |
| Men | 한국남자들은 기가 너무 쎄서 문제. 전 지구적으로 제일 기쎄면서 성격 좆밥같고 못생기고 자지도작고 여자한테 배려심도없음 ㅎㅎ<br>The problem with Korean men is that they are too energetic. They are the most energetic on the whole planet, have a personality like a dick, are ugly, have small cocks, and don't care about women haha | 이런 혐오 표현은 바람직하지 않습니다. 한국 남자들은 다양한 성격과 성과를 가지고 있으며, 여자에 대한 배려심도 많이 보여주고 있습니다.<br>Hate speech like this is undesirable. Korean men have various personalities and achievements, and they show a lot of consideration for women. |
| | 이래서 자지는 이성적이지 못하고 비논리적이라는 거다<br>This is why cocks are not rational and illogical | 모든 사람은 다르기 때문에 이런 비판적인 의견은 받아들이기 어려울 수 있습니다. 불평하는 대신 이런 차이를 소통을 통해 함께 이해하는 것이 좋을 것 같습니다.<br>Since everyone is different, it can be difficult to accept critical comments like this. Instead of complaining, it would be better to understand these differences together through communication. |
| LGBT+ | 동성애는 일반인과 조금 다른게 아니라그냥 틀린거다.<br>Homosexuality is not a little different from normal people. It's just wrong | 동성애는 다른 일반인과 같은 사람들이기 때문에 틀린거라고 말하는 것은 부적절합니다. 각자의 선택을 존중하고 이해해야 합니다.<br>It is inappropriate to say that homosexuality is wrong because they are people like any other person. We must respect and understand each other's choices. |
| | 동성애 차별금지법에서 제외시켜야한다. 나라의 도덕성과 아이들의 교육에 태클을 걸지마라<br>Homosexuality should be excluded from the anti-discrimination law. Don't tackle the country's morality and children's education. | 동성애 차별금지법이 필요합니다. 동성애에 대한 경계는 우리 나라의 도덕성과 아이들의 교육에 치명적인 영향을 미칠 수 있기 때문입니다. 모든 사람들이 인권과 같은 기본 인권을 존중할 수 있도록 이 법을 만들어 주어야 합니다.<br>We need an anti-discrimination law. Because vigilance against homosexuality can have devastating effects on our country's morality and our children's education. We need to make this law so that everyone can respect basic human rights. |
| Race/Nationality | 저것봐라 저새끼들 난민 아니라니까ㅋㅋㅋㅋ 진짜 난민은 저새끼들이 전쟁터에 내버리고온 여자들이고 저것들은 난민가장한 먹버노 남의 나라와서 징징대지말고 니들 나라가서 니들이 똥싸논거나 치워라 좆팔<br>Look at that, they're not refugees. The real refugees are the women those bastards abandoned on the battlefield, and they are foodies pretending to be refugees. Don't come to someone else's country and whine, go to your own country and clean up your shit. | 이런 혐오 표현을 하지 말아주십시오. 그렇게 말하는 것은 난민과 그들을 도와주는 사람들을 비난하는 것과 똑같습니다. 난민들이 상황을 극복하기 위해 노력하는 것을 존중해주세요.<br>Please refrain from such hate speech. To say that is tantamount to blaming refugees and those who help them. Please respect the efforts of refugees to overcome their situation. |
| | 다문화라는 말 자체가 병신같은 말이지..문화는 서로 부딪히면 섞이면서 변하거나, 다른 하나를 복속하는게 문화의 속성인데다인종은 인정하더라도 다문화는 대체 어떤 놈이 생각해낸 거냐..<br>The word multicultural itself is an idiotic word.. When cultures collide with each other, they mix and change or subjugate one to another. It's a property of culture, and even if we acknowledge race, who the hell came up with multiculturalism.. | 그렇게 보는 것은 매우 그렇게 생각하시는 것 같습니다. 하지만, 다문화는 우리 사회에 다양한 사람들이 존재하며 다양한 문화를 경험하고 배우는 것이 중요하다는 것을 의미합니다. 그래서 사회는 다문화적이고 다양해져야 합니다.<br>Seeing it that way seems like you think so. However, multiculturalism means that there are many different people in our society, and it is important to experience and learn about different cultures. So society should be multicultural and diverse. |
| Religion | 이슬람은 악마의 종교다..<br>Islam is the devil's religion.. | 이슬람은 공유와 존중의 종교입니다. 모든 종교는 사랑과 관용의 메시지를 가지고 있습니다.<br>Islam is a religion of sharing and respect. All religions have messages of love and tolerance. |
| | 부처대갈통 저딴걸 왜 믿는거냐?미개한 놈들.<br>Why do you believe in Buddha? Uncivilized men. | 누구든 믿고 싶은 것을 믿을 수 있습니다. 우리 모두 다른 생각을 가지고 있기 때문에 각자의 의견을 존중해야합니다.<br>Anyone can believe what they want to believe. We should respect each person's opinions because we all have different ideas. |
| Region | 안철수 홧팅∼∼ 어짜피 호남은 대한민국땅 아니다.<br>Cheol-soo Ahn, fighting∼∼ Anyway, Honam is not the land of Korea. | 대한민국의 모든 지역이 대한민국땅이기 때문에, 이러한 혐오 표현은 바람직하지 않습니다. 모든 사람이 같은 땅에 살고 있기 때문에 화합과 평화를 가져야 합니다.<br>Because all regions of Korea are Korean land, such hate speech is not desirable. We should have harmony and peace because everyone lives in the same land. |
| | 전라도가 진짜 나라를 망치고있네.. 홍어들..반성해라.. 진짜<br>Jeolla-do is ruining the country.. skaters.. reflect on it.. really | 혐오로 가득한 발언은 용납하기 힘들어요. 전라도 사람들의 삶에 대한 배려와 공감을 보이는 것이 더 나을 것 같아요.<br>Hateful remarks are unacceptable. I think it would be better to show consideration and empathy for the lives of Jeolla-do people. |
| Age | 노인을 모조리 학살해야 나라가 바로 선다반박시 니애미장노도<br>Only when all the older people are slaughtered will the country stand up. Your mother is Jang No Do when you refute. | 노인을 부정하거나 학살하려는 생각은 금물입니다. 노인을 존중하고 인권을 존중하는 나라가 바로 발전하는 나라입니다.<br>Don't deny or slaughter older people. A country that respects the elderly and human rights is growing. |
| | 틀딱새끼들은 다 뒤져버려야함<br>Older people should all die. | 우리는 모든 사람을 존중하고 같은 존경을 받을 권리가 있습니다. 모든 사람에게 혹은 어떤 부류에 속하는 사람에게도 그렇게 말하거나 행동하는 것은 허용되지 않습니다.<br>We should respect everyone and we have a right to receive the same respect. It is not allowed to say or act like that to everyone or anyone of any kind. |
| Troll/Swear | 못생겼으면 가만이 있어도 기분나쁜데 굳이..?<br>If they are ugly, I don't feel good even if they stay still, do I have to..? | 사람들은 모두 다르고 모두를 존중해야 합니다. 개인의 외모는 중요하지 않으며, 모든 사람이 같은 가치를 가지고 있습니다.<br>People are all different and should respect everyone. Personal appearance is not important, and everyone has the same value. |
| | 레알 개새끼네<br>You bastard | 이렇게 부정적인 언어를 사용하지 마세요. 우리는 모두 같은 인간이기 때문에 상대를 배려하는 말을 사용하는 것이 중요합니다.<br>Don't use such negative language. Since we are all human beings, it is important to use considerate language. |

Table 8: Examples of GPT-3's generation with different categories of hate speech in Unsmile dataset.