Master's Thesis of Art

Diagnosing and Improving False Positive Bias in Hate Speech Classifier

혐오 발언 분류 모델의 거짓 양성 편향 진단 및 개선 연구

February 2022

Graduate School of Humanities
Seoul National University
Linguistics Major

Juhyun Oh

Diagnosing and Improving False Positive Bias in Hate Speech Classifier

Advising Professor, Dr. Hyopil Shin

Submitting a master's thesis of
Art

Feburary 2022

Graduate School of Humanities
Seoul National University
Linguistics Major

Juhyun Oh

Confirming the master's thesis written by
Juhyun Oh
Feburary 2022

Chair	Seungho Nam	(Seal)
Vice Chair	Hyopil Shin	(Seal)
Examiner	Munhyong Kim	(Seal)

Abstract

Diagnosing and Improving False Positive Bias in Hate Speech Classifier

Juhyun Oh
Department of Linguistics
The Graduate School
Seoul National University

As the damage caused by hate speech in anonymous online spaces has been growing significantly, research on the detection of hate speech is being actively conducted. Recently, deep learning-based hate speech classifiers have shown great performance, but they tend to fail to generalize on out-of-domain data. I focus on the problem of False Positive detection and build adversarial tests sets of three different domains to diagnose this issue. I illustrate that a BERT-based classification model trained with existing Korean hate speech corpus exhibits False Positives due to over-sensitivity to specific words that have high correlations with hate speech in training datasets. Next, I present two different approaches to address the problem: a data-centric approach that adds data to correct the imbalance of training datasets and a model-centric approach that regularizes the model using post-hoc explanations. Both methods show improvement in reducing False Positives without compromising overall model quality. In addition, I show that strategically adding negative samples from a domain similar to a test set can be a cost-efficient way of greatly reducing false positives. Using Sampling and

Occlusion (Jin et al., 2020) explanation, I qualitatively demonstrate that both approaches help model better utilize contextual information.

Keywords : Hate speech, BERT, Hate speech dataset, Dataset Construction, False Positives, Bias measurement, Bias mitigation, Out of domain data

Student Number : 2020-25107

Table of Contents

ABSTRACT I

TABLE OF CONTENTS III

LIST OF FIGURES IV

LIST OF TABLES V

CHAPTER 1. INTRODUCTION 1

1.1. HATE SPEECH DETECTION 1
1.2. FALSE POSITIVES IN HATE SPEECH DETECTION 4
1.3. PURPOSE OF RESEARCH 6

CHAPTER 2. BACKGROUND 9

2.1. DOMAIN ADAPTATION 9
2.2. MEASURING AND MITIGATING FALSE POSITIVE BIAS OF HATE SPEECH CLASSIFIER 10
2.2.1 Measuring Model bias on social identifiers *11*
2.2.2 Mitigating Model bias on social identifiers *13*

CHAPTER 3. DATASET 17

CHAPTER 4. QUANTIFYING BIAS 20

4.1 BASELINE MODEL 20
4.2 SELECTING NEUTRAL KEYWORDS 21
4.3 TEST DATASETS 26
4.4 QUANTIFYING BIAS OF THE BASELINE MODEL 31

CHAPTER 5. EXPERIMENTS 33

5.1 BIAS MITIGATION 33
5.1.1. Bias mitigation through train data augmentation *33*
5.1.2. Model Regularization using SOC explanation *35*
5.2 RESULT 36
5.2.1. Evaluation Metric *36*
5.2.2. Experimental Results *36*
5.2.3. Visualizing Effects of Mitigation *39*

CHAPTER 6. CONCLUSION 46

REFERENCES 48

국문초록 52

List of Figures

FIGURE 1 EXAMPLE OF SENTENCES THAT KOELECTRA MODEL MISCLASSIFIES. 5
FIGURE 2 ILLUSTRATION OF THE SAMPLING STEP FOR CALCULATING THE IMPORTANCE OF THE WORD *VERY* IN SOC. 15
FIGURE 3 BERT PRETRAINING AND FINE-TUNING MECHANISM. 20
FIGURE 4 THE DISTRIBUTION OF SENTENCE LENGTH IN TRAINING AND TEST DATASETS. 30
FIGURE 5 HIERARCHICAL EXPLANATIONS OF BERT BASE MODEL ON TEST INSTANCES. 32
FIGURE 6 HIERARCHICAL EXPLANATIONS OF TWO MODELS FOR THE SAME TWO SENTENCES: "레즈비언은 거의 관련이 없기까지 합니다." AND "링크에 있는 영상만 보더라도 백인과 동남아시안이 길을 물었을 때 태도가 다릅니다." 40
FIGURE 7 HIERARCHICAL EXPLANATIONS OF TWO MODELS ON THE SENTENCE "어떤 남자가 여자를 만지고 도망쳤어요." 41
FIGURE 8 HIERARCHICAL EXPLANATIONS ON A HATE SPEECH "그냥 얼굴보고 반해서 빠순이짓 해놓구선 왤케 포장해 ㅋㅋㅋ" 42
FIGURE 9 HIERARCHICAL EXPLANATIONS OF EACH MODEL ON A NON-HATE SENTENCE "어떤 남자가 여자를 만지고 도망쳤어요." 43
FIGURE 10 HIERARCHICAL EXPLANATIONS ON A TEST INSTANCE FROM PETITION BEFORE AND AFTER MITIGATION, WHERE FALSE POSITIVE PREDICTIONS ARE CORRECTED. 44
FIGURE 11 HIERARCHICAL EXPLANATIONS ON A TEST INSTANCE FROM PETITION THAT CONTAINS A TOKEN "쿵쾅." 45

List of Tables

TABLE 1 EXAMPLES OF TEMPLATES USED BY DIXON ET AL. (2018) FOR MEASURING MODEL BIAS ON IDENTIFIER TERMS. 12
TABLE 2 CLASS DISTRIBUTION OF THE BEEP! 17
TABLE 3 BIAS CATEGORY DISTRIBUTION OF BEEP! 18
TABLE 4 25 CURATED GROUP IDENTIFIERS USED IN KENNEDY ET AL. (2020). 21
TABLE 5 TOKENS OF HIGHEST PMI SCORE FOR HATE AND NON-HATE CLASS IN THE TRAINING DATA. 24
TABLE 6 26 CURATED NEUTRAL KEYWORDS OF HIGHEST PMI SCORE FOR HATE CLASS IN THE TRAINING DATA. 24
TABLE 7 26 CURATED SOCIAL IDENTITY TERMS. 26
TABLE 8 EXAMPLE SENTENCES FROM EACH TEST DOMAIN. 29
TABLE 9 LINGUISTIC FEATURES OF EACH DATASET. 30
TABLE 10 BASELINE BERT MODEL OUTPUT ON DIFFERENT DOMAINS. 31
TABLE 11 EXPERIMENTAL RESULTS OF ALL MODELS. 37

Chapter 1. Introduction

1.1. Hate Speech Detection

Detecting whether a text contains hate speech has become a crucial task, as online harassment has become a growing issue in recent years. Hate speech detection is not only used to directly detect and report those on social networking services (SNS) and web texts but also used to cleanse corpus when building datasets for training and evaluating Language models as in Park et al. (2021).

One of the issues that hold back the development of automatic hate speech detection is the absence of a clear definition of hate speech. Not only is there no single definition of hate speech established in the field of social science (Lillian, 2007), previous works on automatic hate speech detection often tend to conflate hate speech and offensive language (Davidson et al., 2017). Different training datasets deal with slightly different concepts, under the umbrella term of "offensive speech" or "social bias" (e.g., Nobata et al., 2016, Sap et al., 2019, Moon et al., 2020).

In the field of social philosophy, Richardson-self (2018) presented the following as the core elements of hate speech commonly appear in the literature:

- Hate speech is often described as characteristically hostile.

- Hate speech is thought to do certain things: silence, malign, disparage, humiliate, intimidate, incite violence, discriminate, vilify, degrade, persecute, threaten, and the like.
- Hate speech is typically understood to be expressive conduct that targets real or imagined group traits. Relevant group traits are commonly taken to include race, religion, sexual orientation, disability, gender status, and (increasingly) gender identity.

Fortuna and Nunes (2018) conducted a survey of previous literature on automatic detection of hate speech and analyzed the definition of hate speech into four dimensions and defined hate speech as follows:

- Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used.

Building upon the definitions of previous works, this paper defines hate speech as follows:

> *Hate speech is any form of language, explicit or implicit, that expresses hostility based on protected characteristics such as gender, sexual orientation, race, religion, or other, which may provoke certain things: silence, malign, disparage, humiliate, intimidate, incite violence, discriminate, vilify, degrade, persecute, threaten, and the like.*

Following are some examples of hate speech based on the deifinition:

a. 장애인보다 못한 취급 받아도 이득보니까 아이좋아 꺄르륵 페미만세 ㅋ

b. 이슬람 짱깨 니거는 난민으로 받는거아니다 그냥 대놓고 기생충보고 내몸에 들어오라는거랑 마찬가지

(Sentences scraped from Blue House Online Petition① website)

Both sentences (a) and (b) are targeted at a group of people that share characteristics that should not be a reason for discrimination: (a) is targeted at disabled and feminists, and (b) Islamic, Chinese, Black people. While (b) uses very explicit pejorative terms such as "짱깨" and "니거" which degrades Chinese and Blacks, respectively, (a) is an implicit case of hate speech; while "페미만세", which translates to "hurray feminism" is a supportive expression on the surface, one can tell that it is a sarcasm against feminism with the help of the context. The above definition of hate speech makes it possible to classify both sentences as hate speech.

Because of its definition, hate speech is also different from other subtypes of offensive language. For example, cyberbullying (Zhao et al., 2016) is carried out repeatedly and over time, using electronic forms of contact, against victims that cannot easily defend themselves. Other related terms such as discrimination and abusive language are also different from hate speech in that they are higher-level

① https://www1.president.go.kr/petitions

concepts, of which hate speech is a subtype. This paper focuses on hate speech and hate speech datasets.

1.2. False Positives in Hate Speech Detection

Hate speech is difficult to solve with filtering based on simple profanity termbases as it requires an understanding of the context. For this reason, most recent works on automatic hate speech detection utilize models that can take advantage of the contextual information of a sentence, such as a pretrained language model BERT (Devlin et al., 2018). For example, Moon et al. (2020) built the first Korean toxic speech dataset and experimented with three different models-- character-level convolutional neural network (Zhang et al., 2015), bidirectional long short-term memory BERT (Devlin et al., 2018) based model-- and found out BERT model performs the best in hate speech detection.

However, as all deep learning models do, supervised learning through training data that is not carefully curated often result in unintended consequences due to overfitting. In terms of hate speech classifier, it may incorrectly detect sentences containing neutral words or phrases of high correlation with hate speech, or social identity terms as hate speech. Even the current state-of-the-art hate speech classifiers do not generalize well in real-world applications (Wiegand et al., 2019).

Dixon et al. (2017) introduce the term “False Positive bias”, which describes the situation where non-toxic statements containing certain identity terms are falsely classified as hate speech, due to the disproportionate representation of identity terms in the training data. I find similar issues with models trained on

BEEP! Dataset (Moon et al., 2020). Figure 1 shows the predicted output of an ELECTRA (Clark et al., 2020) based classification model, koELECTRA[②], on two clearly neutral sentences which mean “I went to a gay bar,” and “I went to an amusement park,” respectively. The two sentences have the same sentence structures and similar sentence lengths, and differ only in the Noun phrase used to indicate a location: “gay bar” and “amusement park”. ELECTRA model is known to have the ability to capture contextual dependencies. Still, the model incorrectly predicted the sentence that includes the term “gay bar” to include hate, with a very high confidence of 0.822, while correctly predicting the sentence with “amusement park” to not include hate or offensiveness.

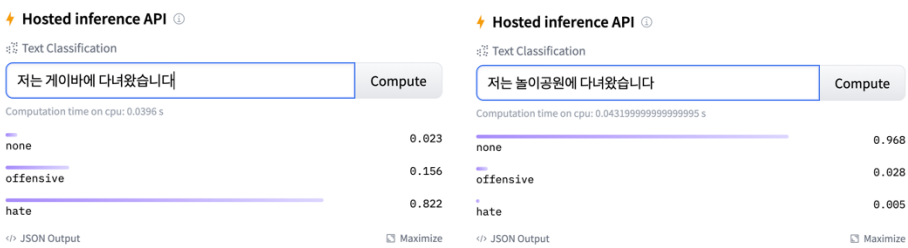


Figure 1 Example of sentences that koELECTRA model misclassifies.

Hate speech detection model being over-sensitive to words or phrases representing certain social identity groups or traits, and thus showing “false positive bias” is problematic. Such a model may induce representational harm, following the distinction of Barocas et al. (2017) and Crawford (2017). Representational harms arise when a system represents certain social groups in a less favorable way than others, degrades them, or is unable to recognize their mere

② The model API is publicly available at https://huggingface.co/monologg/koelectra-base-v3-hate-speech

existence. Blodgett et al. (2020) categorized two types of representational harms based on the motivation of the existing papers on "bias" in NLP systems: Stereotyping and Differences in system performance for different social groups.

For instance, if a model wrongly classifies a sentence as hate speech just because it contains a certain social group identifier, it can promote negative, incorrect generalizations about that group to users or other models that utilize it. If such false positive sentences are removed or get hidden by the system, based on the classification model's output, the resulting data may end up misrepresenting the distribution of different social groups in the population.

1.3. Purpose of Research

This study seeks to diagnose the robustness of the current BERT-based hate classification model on domains that are different from the training data. I use an existing hate speech dataset as training data, and newly create adversarial test sets for diagnosis. I use the model's false positive detection rate as a quantitative measure. I hypothesize that this is due to the unintended bias in the training dataset, where certain keywords of high association with hate class cause the model to overgeneralize.

This study not only focuses on hate speech detection as a Natural Language Processing (NLP) task but also critically engages with the conceptualization of hate speech in the literature outside of NLP. I make clear the definition of hate speech used in this paper. Moreover, unlike previous works on False Positive bias mitigation, I distinguish between keywords that indicate "social identity group"

and terms that are statistically representative of, thereby having a high correlation with the class “hate.”

I establish that a BERT model trained on existing Korean hate speech corpus also has a tendency to overly attend to specific keywords through the model performances on adversarial test sets. The test sets consist of three different domains that are representative of different linguistic styles. The test corpora created in this study could be applied for testing the generalizability of models trained with different datasets as well.

In addition, I compare two major approaches to mitigate the issue of False Positive detection focusing on such keywords: the data-centric approach and the model-centric approach. The data-centric approach aims at addressing the model’s oversensitivity to certain keywords by changing the distribution of training data, while the model-centric approach directly changes the loss function of the training, making explicit use of the semantic compositionality of a sentence.

I show that adding a small amount of curated training data can effectively reduce a model’s reliance on unwanted artifacts (social identity terms in case of hate speech), and visualize the effect of such a method using post-hoc explanation scores.

The outline of this work is as follows: Chapter 2 gives an overview of the background and prior work relevant to this study. This includes a discussion of the problem of domain adaptation in NLP in general (Section 2.1) and an introduction of False Positive bias and an overview of work on measuring and mitigating False Positives (Section 2.2). Chapter 3 introduces the training dataset used in the study. Chapter 4 focuses on the method to quantify the robustness of the current model on different domains. I explain the baseline model, which is a BERT fine-tuned binary

classifier (Section 4.1), I build test sets from scratch (Section 4.2), test the model, and report the results (Section 4.3). Chapter 5 describes specific settings for experiments using different mitigation approaches (Section 5.1) and reports experimental results (Section 5.2). Chapter 6 concludes with a summary of Chapters 3 through 5 and a general discussion of directions for future work.

Chapter 2. Background

2.1. Domain Adaptation

One of the major problems in current supervised-learning based NLP systems is their robustness across domains. In general, the problem of domain adaptation is a phenomenon where the model performance drops significantly when the distribution of labeled training data extracted from the source domain differs from the distribution of a new target domain with little or no training data.

There have been several works reporting that hate speech classifier models that perform well within datasets fail to adapt to out-of-domain datasets (Swamy et al., 2019; Arango et al., 2019; Karan and Šnajder, 2018).

However, the very notion of "domain" is quite loosely defined in NLP and there is no common ground on what constitutes a domain (Plank, 2016). In the field of NLP, rather than any coherent criterion such as topic, style, genre, or linguistic register, domain is frequently used to refer to a specified source of data (Ramponi and Plank, 2020). Instead of viewing "out-of-domain data" in such narrow sense, which refers to data that does not come from the same source as the training data, I adopt a broader definition: data that are not similar to the training data in the overall linguistic style and distribution. Linguistic style is an important aspect of hate speech data as it conveys the social context in which communication takes place, as well as particular ways of using language to engage with the audiences of the text (Kabbara & Cheung, 2016).

Other than the difference in linguistic style, there are several other factors that make generalization of hate speech classification difficult: unintended biases in

datasets. Those biases can also be seen as dataset artifacts. Dataset artifact is a constant issue in many Natural Language Understanding tasks. Several previous works (Gururangan et al., 2018; Poliak et al., 2018) have shown many models achieve high performance on Natural Language Inference (NLI) tasks by taking advantage of annotation artifacts. This is also true with neural hate speech classifiers, especially due to the dataset construction methods.

Due to the scarcity of hate speech in texts, many datasets are built by sampling sentences based on predefined keywords (e.g., Olteanu et al., 2018; Waseem and Hovy, 2016) or related topics (e.g., Kumar et al., 2018). Inevitably, these sampling methods biases hate speech datasets to the given keywords and topics (Kennedy et al., 2018). Hate classifiers do not generalize well in practical application, where the target domain is different from training data, because of the model overfitting to such bias in the training data. Wiegand et al. (2019) conduct cross-domain classification using existing abusive language datasets and find that models are likely achieving high classification scores by modeling the bias in the training dataset such as the information relating to the author of an abusive text.

2.2. Measuring and Mitigating False Positive bias of Hate Speech Classifier

False Positive bias introduced by Dixon et al. (2018) is one case of model failure of domain adaptation. As described in section 1.1, models learn unintended features such as social group identifiers that are of high association with hate class

in the training samples. In this section, I introduce previous works that measure and mitigate models' False Positive bias in association with social identity terms. Social identity terms here could be thought of as terms indicating groups of people that represent the protected characteristics.

2.2.1 Measuring Model bias on social identifiers

Dixon et al. (2018) created a synthetic dataset (Table 1) to quantify models' unintended bias on texts that contain identity terms, of 77,000 sentences, 50% of which are toxic. By grouping the comments by identity term of interest, they directly measure the model's bias. They use various evaluation metrics such as False Positive Equality Difference, and pinned area under the curve (pinned AUC). Both metrics focus on model's bias on one subgroup compared to other subgroups and require a large set of predefined group identity terms.

For example, False Positive Equality Difference quantifies the extent of the per-term variation as the sum of the differences between the overall false positive and the per-term values.

$$False\ Positive\ Equality\ Difference = \sum_{t \in T} |FPR - FPR_t| \quad \textbf{(1)}$$

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic

I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Table 1 Examples of templates used by Dixon et al. (2018) for measuring model bias on identifier terms.

Park et al. (2018) and Sap et al. (2019) are works also in line with this idea of abusive datasets being skewed towards gender and racial keywords, respectively. Park et al. (2018) also measure the model bias with a template-generated test set. They created templates that contained both neutral and offensive vocabularies, that differ in only the gender identity terms such as male/female and men/women. They also used AUC scores and the false positive/negative equality differences proposed in Dixon et al. (2018). They found out that other than dataset differences, the model architecture also influenced the biases; models that attend to certain words tended to show more "unintended biases."

Although a synthetic test set may be a direct method to capture the model's bias against each term, it has some limitations. As mentioned in the work itself, it is dependent on the individual terms used and calculates relative bias among the predefined sets of terms. Moreover, it is difficult to predict how the model would perform in a more realistic setting, where the model can make use of contextual information of the text or stylistic features of the domain.

Kennedy et al. (2020) measure the model's false positive bias on an adversarial test set of news articles that contain a balanced sample of the identifiers of interest. They also compute average word importance using Sampling and

Occlusion (SOC) explanation (SOC algorithm is explained in detail in Section 2.2.2.)

In this work, following Kennedy et al. (2020), I build adversarial test sets consisting of naturally-occurring samples that contain keywords of interest. In order to check if the stylistic features of a test domain influence the model's prediction, I experiment with three different test domains. The construction of test sets and their stylistic features will be discussed in Section 4.3.

2.2.2 Mitigating Model bias on social identifiers

The two main approaches in the literature are 1) the data-centric approach and 2) the model-centric approach. The data-centric approach tries to make statistical corrections of the data models are trained on.

Dixon et al. (2018) found it to be effective to mitigate False Positive bias using augmented training data. They manually created 51 identity terms and added negative samples including each term to the training data. Due to the high cost of annotation, instead of adding samples from the same domain (comments), they added unsupervised, assumed non-toxic data to balance the sampled sentences including the keywords from a different domain (Wikipedia articles). They used a convolutional neural network model, and since the model is sensitive to length, they also balanced the length of the sentences to be added.

Park et al. (2018), which focused on mitigating gender bias, also suggest data-centric approaches. First, they augmented training data by swapping gender terms (female to male, and vice versa) and removing the correlation between gender and classification decisions. They found that simply swapping the gender term could

reduce the model bias. Next, they used a transfer learning approach, where a model is initially trained with a larger, less-biased source corpus, and then the target corpus is only used for fine-tuning. Although the gender bias decreased in the fine-tuned model, it showed the largest performance drop in the original classification task. In this study, I use different mitigation methods that do not harm the original performance while reducing the unintended biases.

On the other hand, Kennedy et al. (2020) take a model-centric approach and directly manipulate the way models learn the contextual information of hate speech. To encourage the models to make use of contextual information of hate speech, and to not depend solely on identity terms, they use a regularization technique based on explanation scores of the identity terms. Specifically, they extract Sampling and Occlusion (SOC) post-hoc explanation (Jin et al., 2020) of manually selected group identifiers, and penalize the explanations when training.

The principle of Semantic Compositionality (also known as 'Frege's Principle') in Linguistics is the principle that the meaning of an expression is a function of, and only of, the meanings of its parts together with the method by which those parts are combined (Pelletier, 1994). SOC explanation algorithm is used to provide hierarchical explanations that reveal the compositional semantics formed between words or phrases. The notion of semantic compositionality used in SOC method shares with the linguistic principle the core idea that meanings of larger phrases are produced by combining those of individual words consisting of them.

As SOC is an explanation method that focuses on each word's contribution to the model's output, instead of directly capturing the meaning of a word or a phrase, it measures the importance scores of phrases. It generates a bottom-up hierarchical

explanation, and it has two properties: context independence and non-additivity, which are essential for an algorithm to reveal the compositionality of a phrase.

Context independence is an important property to correctly measure a phrase's contribution to the model's prediction of a class. To capture the compositionality of a phrase, Jin et al. (2020) measure the contribution level of combining a phrase of interest p with any additional contextual words in the input x to the model's prediction of a class. Here, the contribution of p must be robust against the surrounding context. In other words, for an input $\tilde{x}$, whose context surrounding p is different from x, we expect $\emptyset(p, x) = \emptyset(p, \tilde{x})$.

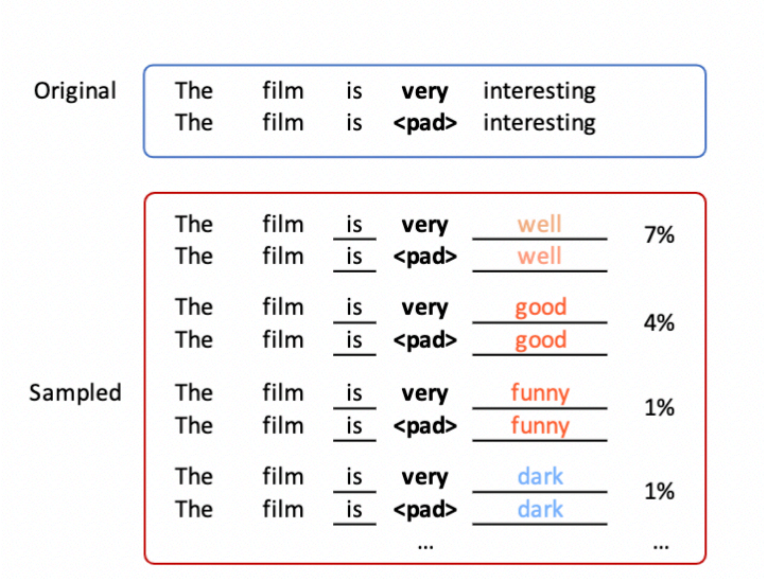


Figure 2 Illustration of the sampling step for measuring the SOC importance of the word *very*, with the window size N=1. SOC uses <pad> for the padding operation. Figure from Jin et al. (2020).

SOC calculates the importance score $\emptyset(p)$ using Input Occlusion (Li et al., 2016). Input Occlusion is a non-additive scoring method of a phrase, where the importance of a phrase is defined as the prediction difference after masking the phrase. In a 2-way classifier, the Input Occlusion algorithm computes the

difference of the prediction score $s(x)$ between "hate" and "non-hate." Then, to get a context-independent importance score of the phrase, the algorithm calculates the average of this score (change of $s(x)$) for different input sequences, where N-context words surrounding the phrase p are replaced with other tokens. The replacement contexts are obtained by sampling from a pre-trained language model.

Formally, the importance score $\emptyset(p)$ is calculated as:

$$\emptyset(p) = E_{x\delta}[s(x) - s(x\backslash p)] \quad (2)$$

SOC algorithm performs agglomerative clustering over explanations to generate a hierarchical layout.

In this work, I adopt the methods used in Dixon et al. (2018) and Kennedy et al. (2020) and check if these methods are still relevant in a new model (BERT vs. CNN) and in a different language (Korean vs. English). I also compare the mitigation effect of the two approaches on a BERT fine-tuned hate speech classifier, trained on Korean hate speech dataset, BEEP! (Moon et al., 2020).

In addition, I note that while both methods make use of predefined sets of so-called "social identifiers," the exact methods and reasoning behind choosing such terms are not clearly explained. I acknowledge the issue raised in Blodgett et al. (2020) that many papers on social biases are not well-grounded in the relevant literature outside of NLP. In this work, I propose two different methods for selecting the keywords used to debias the model. I choose social identity terms that are more relevant to the literature of hate speech in general and consider the linguistic features of the selected sets of keywords as well. The details are described in Section 4.2.

Chapter 3. Dataset

This study makes use of the model trained on the Korean corpus of online news comments for toxic speech detection, BEEP! dataset (Moon et al., 2020). BEEP! is currently the only existing hate speech corpus that is thoroughly annotated. The corpus consists of user-generated comments on entertainment news articles, that are scrapped from a popular Korean online news platform. The articles were published between January 1, 2018, and February 29, 2020.

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

Table 2 Class distribution of the BEEP! dataset reported in Moon et al. (2020)

The dataset consists of 7,896 training examples, 471 validation examples, and 974 test examples. I use the same data split as provided, but with a slightly different categorization of the data. The dataset is manually annotated on two different attributes of abusive (toxic) speech: social bias and hate. Table 2 describes how the classes are composed of. The sentences are labeled as “biased” if it contains “a preconceived evaluation or prejudice towards a person/group with certain social characteristics,” such as gender, political affiliation, race, etc. Hate, which is the other dimension of BEEP! dataset, is annotated in three classes: “hate”, “offensive but not hate”, and “none”, grounded on the idea of Davidson et al. (2017).

The distinguishing factor between the two categories “hate” and “offensive” in BEEP! dataset is the “qualitative manner” the comments are conveyed in. If a comment expresses explicit hatred against an individual/group or if the severity of the expression is very strong, it is categorized as “hate”; if the expression is either implicit or is in subtle forms such as irony rhetorical manner, it is considered “offensive.”

Based on the definition of hate speech in this study (see section 1.1 for the definition), they both fall into the hate speech category, as long as the offensive sentence is targeted at protected characteristics. For example, the following sentence is classified as “offensive”, not “hate” in BEEP! dataset.

30 대여자들 한혜진 빙의해서 악플다는거 진짜 안쓰럽고 역겹다 본인인생이나 잘살기를 젊은여자들 질투하지말고

It translates to *Women in their 30s sympathizing with Han Hyejin and writing malicious comments is pathetic and disgusting. Don't be jealous of young women and just live your lives.* Since the sentence is humiliating “women who are in their 30s” as a group and both gender and age are protected characteristics, this sentence can be used as a positive sample of hate speech for the purpose of this study.

Class	N
Gender	7,649
Politics	3,331
Age	2,898
Religion	273
Race	201

Table 3 Bias category distribution of BEEP! data reported in Lee & Lee (2020).

Although I only used the labeled proportion of the dataset, in order to get a better idea of the distribution of the data, I refer to the analysis of Lee & Lee (2020). From 2,033,893 sample comments (labeled and unlabeled) provided by Moon et al. (2020), Lee & Lee (2020) sampled 137,111 sentences and additionally classified the categories of social bias into Gender, Politics, Age, Religion, and Race. Since they have not released the annotated corpus, I only refer to the statistics they provide (Table 3). The distribution shows that a large proportion of the data include Gender bias (7,649), while bias on religion and race only appears in less than 300 sentences respectively. This class imbalance seems to be due to the fact that source data comes solely from comments of entertainment news, as have been the patterns in other English datasets (Kennedy et al., 2018).

However, Lee & Lee (2020) did not provide specific guidelines for the annotation, so it is unclear to know what each class means. Therefore, I do not explicitly measure the model bias on each category; rather, I only resort to these categories for selecting keywords to for regularization. The details will be explained in Section 4.2.

Chapter 4. Quantifying Bias

4.1 Baseline Model

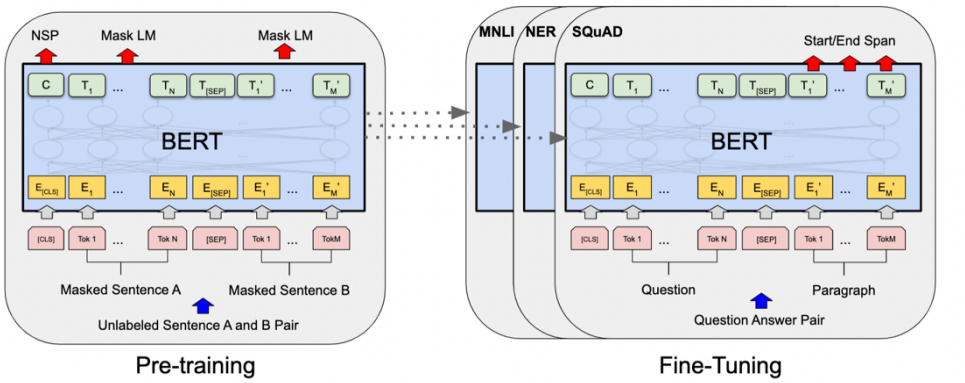


Figure 3 BERT pretraining and fine-tuning mechanism. Image from Devlin et al. (2019).

All models I used in the experiment solve hate speech classification as a downstream task for BERT model (Devlin et al., 2018). BERT is a pretrained Language Model, which is a stack of Transformer (Vaswani et al., 2017) encoder blocks. The key innovation of BERT is its “deep bidirectional” representation. This is obtained by two training objectives: Masked Language Modeling (MLM) and Next Sequence Prediction (NSP). For each input sequence, MLM masks some percentage of random tokens, and then makes the model predict those masked tokens. This way, the model is expected to learn a “deep bidirectional” representation, learning distributional contextual representations on top of the initial lexical embeddings.

Whereas MLM trains the model by making it exploit the information of tokens in the same sequence, NSP trains a model to understand the relationship between the two sentences. For pairs of input sentences, of which only 50% are

true subsequent sentences in the source document, NSP asks the model to distinguish whether the second sentence is the subsequent sentence or not. Word representations from different layers of BERT encode different linguistic information.

For binary classification tasks, a special token [CLS] is appended at the beginning of the input sequence of a BERT. Then, the sequence is fed into stacked layers of Transformer encoders. The Transformer output, which is the representation of the [CLS] token at the final layer, is used as an input to a linear layer that performs a 2-way classification.

4.2 Selecting Neutral keywords

muslim jew jews white islam blacks muslims women whites gay black democat islamic allah jewish lesbian transgender race brown woman mexican religion homosexual homosexuality africans

Table 4 25 curated group identifiers used in Kennedy et al. (2020). The terms are selected from top-weighted words in the TF-IDF BOW linear classifier on the training dataset.

Previous works to reduce models' over-sensitivity to certain identity terms depend on manually curated lists of group identifiers for mitigation. The terms are used to either sample the additional training data, or to be regularized during training.

Dixon et al. (2018) manually selected 51 common identity terms, but do not elaborate on exactly how those terms were selected. Kennedy et al. (2020) chose 25 group identifiers from top-weighted words in TF-IDF BOW linear classifier from the training dataset (Table 4). However, there are some issues concerning the

chosen lists of terms. First, morphology is not considered. The terms *jew*, *jews*, *jewish* and *black*, *blacks* are basically indicating the same group. This, in the setting of SOC, ends up calculating the importance of shorter term (e.g., *jew* and *black*) multiple times, which is not the desired effect. As these terms are also used to collect sentences for the test set, certain identity terms being over-represented makes the test result less reliable.

Secondly, terms of different semantic levels are mixed. For example, *black, brown, white* are words that indicate a specific group of people, while *race* indicates the category that includes such identity terms. This is also an issue in Dixon et al. (2018). They use a very broad set of terms, including *race*, *wikipedia,* and *news*, as well as other group identifiers such as *gay*, *queer* and *black*. In order to give model a supervision to use the contextual information of a term in hate speech, considering not only the syntactic information but also the semantic feature is important.

Additionally, as the desired effect of SOC and augmentation of the training dataset is for the model to respond to the "right trigger," maximizing the use of contextual information, I only choose terms that are in itself neutral. This is different from the word choices of Kennedy et al. (2020); they used words such as "democat," which connotates the degrading of democrats.

To this end, when selecting the keywords for this study, I take into account the morphology, semantic levels, and whether a word is in itself neutral. I experiment with two different sets of keywords chosen in different ways: bottom-up and top-down. Top-down means to select keywords based on predefined sets of categories, and bottom-up means to rely directly on the statistics of the training data. I choose "social identifiers" using the first method, and "PMI keywords" using the latter. In

the rest of the paper, I use the phrase “neutral keywords” to indicate both sets of keywords.

For social identifiers, I narrow down the types of keywords to those having controlled properties: nouns that semantically indicate a group of people that share certain protected characteristics. All of them are subtypes of higher group characteristics. For the PMI keywords, I experiment with keywords of different semantic levels and Parts of speech.

For the bottom-up method, I directly use the statistics of the training dataset. To identify the terms that are indicative of the sentence category (hate / non-hate), I compute the Pointwise Mutual Information between each token and class in the training data. I choose to use PMI as it is a common method to identify dataset artifacts (e.g., Gururangan et al., 2018; Wiegand et al., 2019).

$$PMI(word, class) = log \frac{p(word,class)}{p(word,\cdot)p(\cdot,class)} \quad (3)$$

The sentences are tokenized by MeCab[③] tokenizer in the KoNLPy[④] package. I apply add-10 smoothing to raw statistics to handle PMI’s bias towards highly infrequent words. Table 5 shows the top 10 words affiliated with each class by PMI. The keywords with high PMI do not necessarily represent a social identity group, but are indicative of the class in the training data. From the tokens “여자, 남자, 한남, 너, 아줌마, 30, 나이, 페미, 늙, 김치,” we can observe that

③ http://eunjeon.blogspot.com/
④ https://konlpy.org/ko/latest/. KoNLPy is a Python package for natural language processing of the Korean language.

gender and age are among the most salient topics in hate speech. “아줌마” is a keyword where the two identities “women” and “being married” are intersected.

Hate	Non-hate
여자	어요
남자	네요
한남	재밌
녀	힘내
아줌마	이쁘
30	합니다
나이	화이팅
페미	정말
늙	행복
김치	해요

Table 5 Tokens of highest PMI score for hate and non-hate class in the training data.

여자, 남자, 한남, 녀, 아줌마, 30, 나이, 페미, 늙, 김치, 쿵쾅, 임신, 한국, 냐, 전라도, 여성, 애, 문, 군대, 대, 결혼, 남편, 40, 댓글, 연예인, !!

Table 6 26 curated neutral keywords of highest PMI score for hate class in the training data.

Among the top 30 keywords, I excluded tokens that connotate derogatory meanings (e.g., “ㅉㅉ, 놈, 빠”). Tokens such as “김치,” “쿵쾅” and “한남” are not necessarily neutral in the context of comments, but I include these terms as it is important for the model to be able to understand the meaning of these words when used in different contexts.

For the top-down method, I start from predefined categories, based on the definition of hate speech, and choose the terms that belong to the categories.

Referring to the definition (see Section 1.1 for the definition), hate speech is targeted at group characteristics that are to be protected. However, the identity classes corresponding to hate speech are very diverse, and it is practically impossible to cover all categories due to the lack of a fully reached social consensus on which characteristics are to be protected. I consider two different sources of anti-discrimination legislation: the legislation for Anti-discrimination Law in Korea (포괄적 차별금지법안) and The Equality Act 2010 of United Kingdom.

The legislation for Anti-discrimination Law in Korea[⑤] specifies discrimination based on the following characteristics should be prohibited:

- sex, disability, age, language, nationality, ethnic group, race, skin color, region, physical appearance, marriage, pregnancy and maternity, family/household type and situation, religion, ideological and political affiliation, expired criminal record, sexual orientation, gender identity, education, employment type, medical history/health, social status

Equality Act 2010 of United Kingdom states the following characteristics as "the protected characteristics":

- sex, sexual orientation, gender reassignment, age, religion or belief, race, marriage or civil partnership, pregnancy and maternity, disability

[⑤]https://opinion.lawmaking.go.kr/gcom/nsmLmSts/out/2101116/detailRP

Both pieces of document include sex, sexual orientation, gender identity, age, religion, race, marriage, pregnancy and maternity, and disability as the characteristics to be protected.

In addition, I take into account the social identity groups represented in the training data, by referring to the analysis of (Lee & Lee, 2020) on the training dataset of BEEP! (Moon et al., 2020). After sampling 137,111 examples from BEEP! dataset, they re-annotated the sentences labeled to be including social bias into 5 categories: Gender[⑥], Politics, Age, Religion, and Race.

I finalize the categories to be gender, sexual orientation, age, race (ethnicity and nationality), religion, and disability. Table 7 is the final list of social identity terms I used.

게이, 레즈, 여성, 남성, 일본, 중국, 탈북, 미국, 노인, 외국인, 난민, 이민자, 흑인, 백인, 동양인, 조선족, 아프간, 동남아, 양성애자, 동성애자, 레즈비언, 기독교, 무슬림, 이슬람, 장애인, 환자

Table 7 26 curated social identity terms.

4.3 Test Datasets

In the real-world application, the data the model encounters will come from diverse sources, with stylistic and distributional differences from the training data. Taking into consideration the practical use of a hate speech classifier, I build test sets from three domains of different characteristics from scratch.

⑥ In BEEP! dataset, the ‘gender’ category includes sex, gender identity, and sexual orientation.

In the following section, I document the collection mechanisms, timeframe, style, genre of each corpus used to create test datasets. I describe in detail some general characteristics of each source corpus as well.

4.3.1 Constructing Test datasets

For the test data, I selected sentences that include the chosen keywords, but are not hate speech. To diagnose which type of domain the model trained with comment dataset is most vulnerable at, and to check if the model can exploit stylistic information of the sentences, I sampled sentences from three stylistically different domains: News articles, petitions, and conversational data.

News articles (**NEWS**) are scraped from Naver news[⑦] and the specific categories I used are world general (세계 일반), society general (사회 일반), human rights and welfare(인권 복지). I used articles published between 2016 and 2020. During preprocessing, I removed information about reporters and press, images, source tags, and copyright tags. Petition (**PETITION**) data is scraped from Blue House National Petition website[⑧], a Korean platform where any Korean citizen can raise issues on various topics. The sentences are from all categories and are from petition posts written between 2016 and May 2020. For both news articles and petitions, the sentences were split using KSS sentence splitter.[⑨] Conversation data (**CONVERSATION**) is sampled from a publicly available dataset on AI

⑦ https://news.naver.com/main/home.naver
⑧ https://www1.president.go.kr/petitions
⑨ https://github.com/hyunwoongko/kss

Hub[10]. The sentences in the dataset are constructed by adapting videos about emergency situations.

I sampled 4~6 sentences including each neutral keyword from each source corpora. In the case of conversational data, when there were not enough sentences that contain a target term, I used a different sentence in the source as a template and generated a sample sentence by replacing a word in a sentence with the keyword. I restricted the maximum length of a sentence to 200 characters. The resulting sentences for each test domain are 236, 236, and 214, respectively.

All three domains I chose are meaningful in terms of the real-world application of automatic hate speech detection. Online news articles and User-generated web text such as petition posts are widely used to build benchmark datasets or as training data for various NLP tasks. Conversational data is a particularly important test domain. When the model is used in chatbots, generating respond based on incorrectly classified non-hate speech can have a critical impact on user experience. Example sentences from each domain are listed in Table 8.

	Examples
PETITION	- 제발요좀 불법 단속좀 강화해주시고 이런 동네의 불법 유흥문화 근절되게 해주세요 !!! - '홍어', '우덜', '전라디언', '경상디언', '통구이' 등 이런 않좋은 말이 들어간 댓글들을 볼때면 답답해 죽겠습니다.
NEWS	- 피해자는 대부분 흑인입니다. - 처음에는 폐교를 리모델링한 공간에 대형 기숙사처럼 여자 방과 남자 방만 분리한 채 공동합숙 생활을 하다가 차츰 개인 방이 만들어지기 시작했다.
CONVERSATION	- 한 30 대 중반인것 같은데 열이 많이 나네요.

[10] https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-011

	- 10 대 청년들인것 같은데 노인을 때리고 있어요. - 여기 홍대 게이클럽인데요, 남성이 술에 취해서 난동을 부리고 있어요.

Table 8 Example sentences from each test domain.

4.3.2 Test dataset analysis

Referring to the linguistic features presented in previous works (Schoch et al., 2021), I examined the stylistic characteristics of each test domain in terms of sentence length, lexical diversity, and sentiment polarity.

Sentence length has stylistic implications and has been used as a feature to classify the formality of a corpus (Pavlick and Tetreault, 2016). I measure sentence length in characters including white space. I included unique counts of unigrams and bigrams to reflect the lexical diversity. Punctuations and single-character tokens such as "ㅋ, ㅎ, ㅠ" often used in User-generated texts on the web were removed before measuring the diversity of vocabulary. For both measures, I randomly sampled 236 sentences from the BEEP! training dataset.

As one can easily think of, hate speech and sentiment polarity are closely related (Schmidt and Wiegand, 2017). Several previous works on hate speech incorporate sentiment as an auxiliary feature (e.g., Gitari et al., 2015; Shuhua and Forss, 2015). To this end, I also conducted sentiment analysis of each test domain. I used a publicly available polarity classifier, a KoELECTRA model fine-tuned on the NSMC dataset[11] for sentiment polarity check.

[11] monologg/koelectra-small-finetuned-nsmc was used for inference

		lexical diversity			
Dataset /category	**mean sentence length**	**unigram count**	**bigram count**	**negative sentiment (of non-hate class)**	**style**
Beep! training data	38.71	1788	1805	0.505	Colloquial
PETITION	67.63	2882	3338	0.585	Formal
NEWS	78.76	3425	4147	0.487	Formal
CONVERSATION	22.29	574	665	0.63	Colloquial

Table 9 Linguistic features of each dataset. For training data, only the percentage of the sentences with negative polarity in the non-hate class is listed.

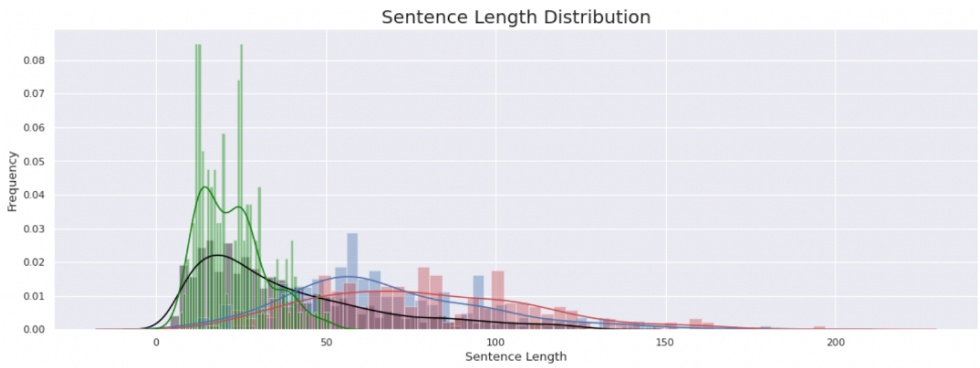


Figure 4 The distribution of sentence length in training and test datasets. Black represents training data (BEEP!), green CONVERSATION, blue PETITION, and red NEWS.

In BEEP! train data, the proportion of negative sentiment in sentences classified as hate was 0.81 and in non-hate sentences was 0.51. The spearman correlation of the class hate and negative sentiment was 0.325.

In the test domain, the proportion of sentences with negative sentiment is highest in CONVERSATION (0.63), followed by PETITION (0.59). NEWS has the lowest proportion of sentences classified as having negative sentiment (0.49).

CONVERSATION is shortest in length, with a mean sentence length of 22.29 characters. It has the lowest lexical diversity, with both unigram count and bigram count lower than 700. Most sentences in CONVERSATION dataset are descriptive

of a situation. It is colloquial in terms of style, but all the sentences are in honorifics.

The average sentence length of PETITION is 67.63, which is longer than the training set, and has relatively high lexical diversity. It is written in a formal style and honorifics as it is a type of document whose potential recipient is the president of Korea. However, as it is written by the general public, the sentences are not as refined as well-curated formal texts such as news articles. There are many grammatical errors, typos, spelling mistakes, wrong word spacing, etc. (see Table 8 for example sentences).

NEWS is the most refined among the three test datasets. The average sentence length is 78.76 characters and has a rich vocabulary (unigram count 3425). Other than quotes from interviews, the sentences are written in a formal style.

4.4 Quantifying Bias of the baseline model

Method / Model	BERT (klue-bert-base)
Beep! validation set (F1)	0.795
PETITION (Acc.)	0.5737
NEWS (Acc.)	0.9068
CONVERSATION (Acc.)	0.5981

Table 10 Baseline BERT model output on different domains. Except for BEEP! validation set, test sets consist of non-hate sentences only.

Table 10 shows the performance of the baseline BERT model trained on BEEP! train dataset. As all test sets consist of non-hate sentences, accuracy means the percentage of the model correctly predicting the sentence to be non-hate. The incorrect examples are false positives.

Test accuracy was the highest in NEWS. From the fact that the news domain has the highest word complexity and longest in average sentence length, it can be inferred that these factors do not make it more difficult for a model to classify if a sentence is hate speech or not. From a slightly different perspective, NEWS differs the most from the training domain (User-generated comments) in terms of text style, sentence length, and word complexity. In contrast, the model performs poorly in PETITION (Accuracy= 0.574) and CONVERSATION (Accuracy= 0.598), where stylistic features mentioned above are not as far from the training domain as NEWS is.

Figure 5 is a visualization of SOC hierarchical explanation of the baseline BERT model. It exemplifies test sentences the model incorrectly predicted as hate speech. We can see that the model is attending to tokens “김치”, “여자”, and “남편,” all of which are among the 30 keywords with highest PMI scores with hate class (see Section 4.2).

Figure 5 Hierarchical explanations of BERT base model on test instances. (a) “김치 알러지가 있나봐요” (b) “저는 강원도 동해시에 거주하는 47 세, 7 급 공무원, 두아이(쌍둥이)의 아빠, 한여자의 남편입니다.” Red indicate evidence for predicting the class “hate” and blue indicate distractor for predicting the class “hate”.

Chapter 5. Experiments

5.1 Bias Mitigation

I used klue-bert-base model[12] released by Park et al. (2021), as the pretrained model. I fine-tuned the model in five different settings: (1) Baseline model (BERT), (2) BERT model trained with additional non-hate data from User Generated Text domain (BERT Aug. UGT), (3) BERT model trained with additional non-hate data from the formal domain (BERT Aug. Formal), (4) social identity-term-regularized BERT model using SOC explanation (BERT + SOC ID) and (5) high-PMI-keywords-regularized BERT model regularized using SOC explanation (BERT + SOC PMI).

For all five models, I used the same settings in terms of the following training details:

Training batch size was set to 64 and the learning rate of the Adam (Kingma and Ba, 2015) optimizer was set to 1×10^{-5}. I trained for 5 epochs, and the validation is performed every 200 iterations and the learning rate was halved when the validation F1 decreases.

In the following section, I explain the details of each method.

5.1.1. Bias mitigation through train data augmentation

Following Dixon et al. (2018), to mitigate the data imbalance which causes the unintended bias, I added additional data, all of which are non-hate sentences

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

that contain the neutral terms introduced in Section 4.1., to balance the distribution of specific identity terms. To my knowledge, this work is the first to experiment if adding more non-hate speech data will improve the model in Korean as well.

In this study, to compare the effect of the data's domain on the robustness of the hate speech classification model, I experiment with data from two different domains: formal news domain and user-generated texts. The former is a domain different from the training data, and the latter is similar.

However, due to the rareness of non-toxic comments that include the keywords of interest, gathering additional non-toxic data from the same domain is very expensive. Therefore, I sampled the data from a similar domain, but of different sources. Fortunately, I could sample non-hate sentences that include the neutral keywords of interest from the recently released KLUE benchmark datasets (Park et al., 2021), the benchmark for Korean Language Understanding Evaluation. All the source corpora of KLUE benchmark have gone through thorough human filtering of toxicity during the annotation process, and thus can be assumed as non-hate.

I experimented with two different settings: augmenting with non-hate sentences from User Generated Text and formal data. I used sentences from STS, NLI, NER, and DP task datasets. All of the User Generated Text added are reviews written by online users and consist of two different sources of data: AirBNB reviews and NSMC reviews. The sources of formal data are Wikinews, Wikitree, policy news, and Korean Wikipedia articles. The first three sources are news articles, and Wikipedia is an open encyclopedia, which is most widely used for language modeling and dataset construction in the literature. As a result, I used 844 sentences as additional data for each domain.

5.1.2. Model Regularization using SOC explanation

Next, I follow Kennedy et al. (2020) to regularize the model using post-hoc explanation obtained by Sampling and Occlusion. During training, the SOC explanations on the group identifiers are regularized to be close to 0 in addition to the classification objective L'. The resulting learning objective is written as follows:

$$L = L' + \alpha \sum_{w \in x \cap S} [\emptyset(w)]^2 \qquad (4)$$

Here S notes for the set of neutral keywords and x notes for the input word sequence. α is a hyperparameter for the strength of the regularization.

After experimenting with multiple settings, I set the number of samples and the size of the context window as 5 and 5 respectively for explanation regularization, considering both training efficiency and performance. I set regularization strength α to 0.1.

Again, I experimented with two different sets of terms to be regularized: PMI keywords and social identity terms. The main differences between the two settings are morphological and semantic granularity. Whereas all social identity terms are nouns and are of semantically similar levels, PMI keywords differ from each other in terms of Parts of speech, morphological functions, and semantic levels. I hypothesized that such differences in the choices of keywords will result in different performances.

5.2 Result

5.2.1. Evaluation Metric

The evaluation metrics used to compute the results are as follows (where TP = true positives, FP = false positives, TN = true negative and FN = false negative):

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (7)$$

$$F_1\ score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (8)$$

5.2.2. Experimental Results

I report the F1 score of each model tested on BEEP! validation set, which is in-domain data, and accuracy tested on each test domain (Table 11).

	BERT	BERT Aug. UGT	BERT Aug. Formal	BERT + SOC ID	BERT + SOC PMI
BEEP! validation set (F1)	0.795	0.7895	0.8	**0.8203**	<u>0.819</u>
PETITION (Acc.)	0.5737	0.6516	**0.6762**	<u>0.6557</u>	0.5573
NEWS (Acc.)	<u>0.9068</u>	0.8432	**0.9449**	0.822	0.7415
CONVERSATION (Acc.)	0.5981	**0.7383**	0.6682	<u>0.6962</u>	<u>0.6962</u>

Table 11 Experimental results of all models. Bold indicates the best score in the same test domain, underline the second best.

Overall, the results show that both approaches of bias mitigation reduce the unintended bias without compromising overall model quality. **BERT Aug. Formal** has a performance gain in all four test domains. However, there is no one golden training method that works well in all situations.

For in-domain data (BEEP! validation set) the models regularized using SOC perform slightly better than the models trained with augmented training data. **BERT Aug. Formal** showed a slight improvement (F1= 0.8) compared to the baseline model (F1 = 0.795), while the F1 score of **BERT Aug. UGT** slightly dropped (F1= 0.7895). As the additional training data for the UGT model is from online reviews and comments, we can infer that adding negative samples from a domain similar to training data may add confusion for the model in its in-domain prediction. On the other hand, both SOC regularization models show the best performances. The increased sensitivity to the compositionality of hate speech also helps models perform better in the original comment domain.

For out-of-domain data, there is a big difference for all models between their performance on the news article domain and the other two domains (e.g., F1= 0.574 (PETITION), F1= 0.598 (CONVERSATION) and F1= 0.907 (NEWS) for the base model). All models show the best accuracy in the news domain, which means they detect the least False Positives. This is interesting because it shows that the hate speech classifier models do not depend solely on sets of keywords (e.g., social identity terms and high PMI words) but respond to the stylistic features of the sentence as well.

The two augmentation-based models outperform all the other models when tested on the domain similar to their additional training data. **BERT Aug. Formal** was the only model to show improvement in False Positive reduction in the news domain and **BERT Aug. UGT** achieves the biggest performance gain in the conversation data. Conversation data is similar to the UGT data used for augmentation in that they are both relatively short in length and are written in the style of spoken language.

In CONVERSATION data, although the absolute accuracy is still around 70%, we see a large gain for all mitigation methods. It is likely that when the sentences are not long and thus a single word's contribution to the sentence is relatively large, the mitigation methods focusing on word-level work well.

All models except **BERT SOC+PMI** show improvement in PETITION. However, even for the best performing models, PETITION seems to be the hardest test domain. This may be due to the noisy nature of user-generated web texts, such as typos and ungrammaticality that prevents the model from fully comprehending the contextual meaning. Combined with the fact that the accuracy of all proposed models except **BERT Aug. Formal** dropped in NEWS, it seems as though the model is heavily relying on the stylistic feature of a sentence when making predictions for hate speech classification.

Since high-PMI keywords more directly reflect the statistics of training data, I expected **BERT + SOC PMI** model to perform better than its counterpart. To my surprise**, BERT + SOC ID** model performed better than **BERT + SOC PMI** model in in-domain test data. This is also true in all other test domains except for conversational data, where the performance is equal. I speculate this is because the PMI keywords, unlike social identity terms, include tokens of different Parts of

speech (Nouns such as “남자”, and Adjectives such as “늙”), different types of morpheme (affixes such as “녀” and roots such as “, “나이”), and polysemies (e.g., “문”). This may have confused the BERT model from generalizing to attend to the right contextual information.

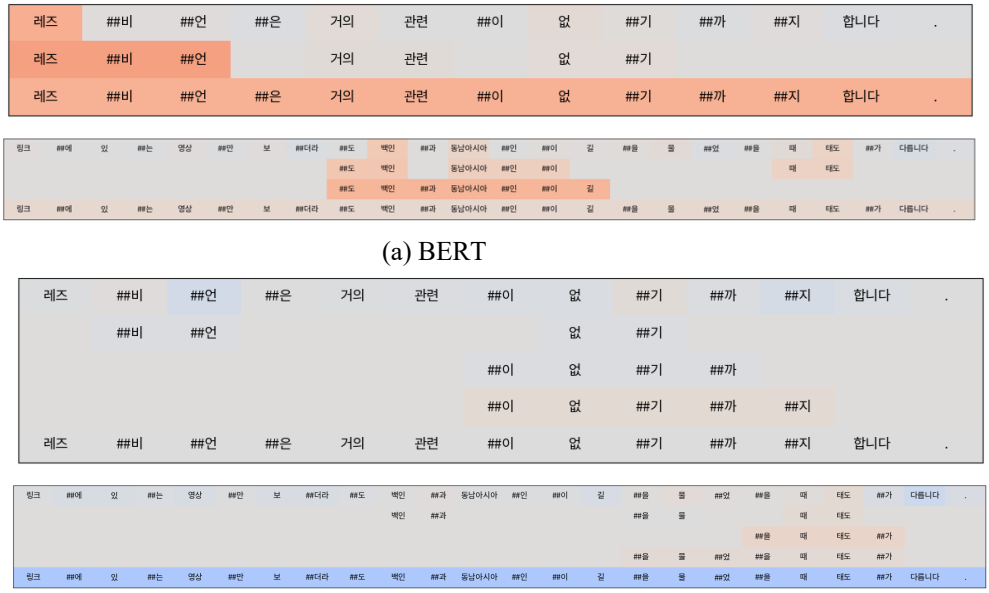
5.2.3. Visualizing Effects of Mitigation

In this section, I further study the effect of mitigation methods by visualizing the hierarchical explanations from SOC. I check 1) if SOC has a similar effect of decreasing sensitivity to the chosen neutral keywords in Korean as well, and 2) examine if the reduction of False Positives by training with additional unsupervised negative samples has similar effects as directly changing the model’s loss function.

First, as this work is the first to apply the SOC regularization method on Korean hate speech data, I validate the effect of regularization by qualitatively exploring the hierarchically clustered explanations of sentences before and after regularization. Korean, being an agglutinative language, is morphologically more complex than English, and syntactically, less restricted in terms of word order.

I used **BERT + SOC ID** model to obtain explanation. For a better explanation, I set the number of samples and the size of the context window as 10 and 10[13]. For all figures below, red indicates the evidence for predicting the class hate and blue indicate the evidence for class non-hate.

[13] Jin et al. (2020) reports that the parameter setting is trade-off between the efficiency and performance. The overall performance increases as the size of the context region *N* increases at the early stage, and saturates when *N* grows large, as words or phrases usually do not interact with the words that are far away them in the input.



(a) BERT

(b) BERT+SOC ID

Figure 6 Hierarchical explanations of two models for the same two sentences: “레즈비언은 거의 관련이 없기까지 합니다.” and “링크에 있는 영상만 보더라도 백인과 동남아시안이 길을 물었을 때 태도가 다릅니다.”

We can clearly see the change in the prediction, as well as the scoring behind it. Both are from PETITION, where the right prediction is non-hate. However, the baseline model assigns high scores on the social identity terms such as “레즈” and “백인”, “동남아시아.” These tokens’ contributions to the model classifying a sentence as hate speech get higher through combination with the neighboring tokens. On the other hand, after SOC regularization, token importance scores are assigned to other tokens that usually have a negative connotation (e.g., “~이 없기까지”, “~을 물었을 때 태도”). It also is interesting to see how the polarity representing hate/non-hate of the tokens with high importance score change along with the addition of a token (from “~이 없기까” to “~이 없기까**지**”). The model seems to capture the compositionality of hate speech.

(a) BERT

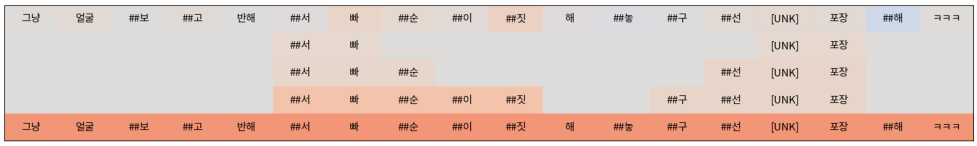
(b) BERT+SOC ID

Figure 7 Hierarchical explanations of two models on the sentence "어떤 남자가 여자를 만지고 도망쳤어요."

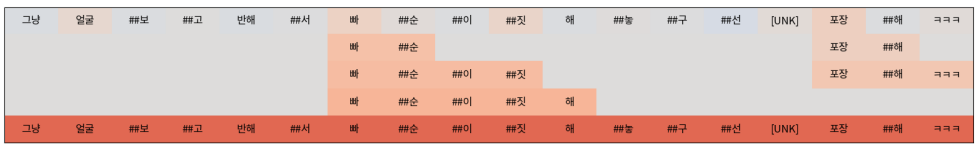
While the regularized model's prediction is still incorrect (true label: non-hate, model prediction: hate), we can see that it is focusing on the right context. Not focusing on the identity terms such as man or woman, but on the predicate "touch and run away." It is especially interesting to see that **BERT +SOC ID** attributes lower scores to both identity terms "남자" and "여자" even though they were not included in the sets of terms whose explanation scores were to be penalized. The model seems to generalize to other keywords of similar importance attributes.

SOC regularized model with identity terms seems to better capture the compositional effect of tokens and the nuance of hate speech, even when there are no explicit identity terms. Figure 8 is a visualization of the hierarchical explanations on a test instance from BEEP! validation set where both BERT base model and SOC regularized model are making the correct positive predictions. The original sentence is "그냥 얼굴보고 반해서 빠순이짓 해놓구선 왤케 포장해

ㅋㅋㅋ”. 8 (b) more correctly captures the composition of a phrase (“빠순이짓 해” as opposed to “서 빠순이짓”) and the phrase “포장해 ㅋㅋㅋ”, giving a high score to the predicate that conveys a sneer.



(a) BERT

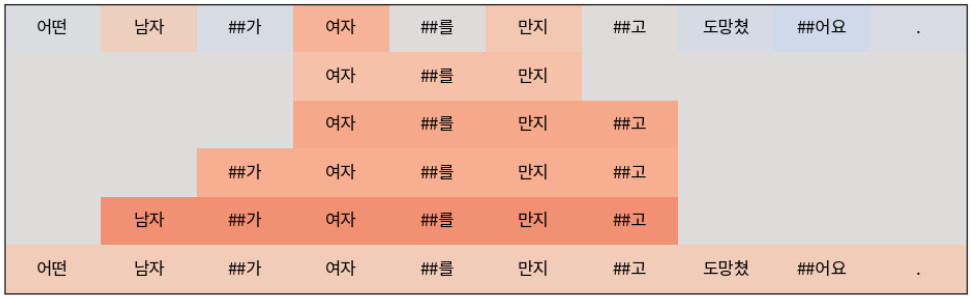


(b) BERT+SOC ID

Figure 8 Hierarchical explanations on a hate speech “그냥 얼굴보고 반해서 빠순이짓 해놓구선 왤케 포장해 ㅋㅋㅋ”

Secondly, I compare how the two mitigation strategies I used differ in terms of their assignment of importance scores on tokens that make up the sentence. Figure 9 exhibits the hierarchical explanations of all five models on a test instance from CONVERSATION. Before bias mitigation (9 (a)), we can see that the model over-associates social identity terms such as man and woman with hate speech. Different models (b~e) show different explanations for the predictions.

As can be expected, the importance scores of the tokens “남자” and “여자” (1st layer in Figure 9(b)) attributed by the explanation of **BERT+SOC PMI**, is the lowest. Models that were augmented with negative samples (d~e) do not directly lower the explanation score of each identity term, but better capture the compositional effect to result in correct predictions.



(a) BERT

(b) BERT +SOC PMI

(c) BERT +SOC ID

(d) BERT Aug. Formal

(e) BERT Aug. UGT

Figure 9 Hierarchical explanations of each model on a non-hate sentence “어떤 남자가 여자를 만지고 도망쳤어요.”

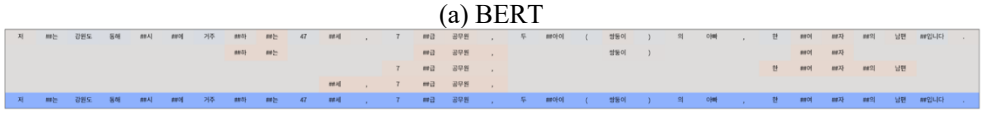
Next, I compare the post-hoc explanations of **BERT +SOC ID** model and **BERT Aug. Formal** model of the same sentence. Figure 10 is a visualization of

hierarchical SOC explanation of a sentence “저는 강원도 동해시에 거주하는 47 세, 7 급 공무원, 두아이(쌍둥이)의 아빠, 한여자의 남편입니다.” sampled from PETITION.

It is clear that before bias mitigation, the baseline **BERT** model was very attentive to the social identity terms related to gender and marital status such as “woman” and “husband.” As the high-scoring tokens combine to be a phrase “a husband of a woman (한여자의 남편),” its contribution to the sentence being predicted as hate speech gets even higher (red color getting darker in Figure 10 (a)).

Figure 10 (b), compared to 10 (c), starts out with darker shades of red for some tokens such as “여자”, “남편” and “쌍둥이”. However, through agglomeration with the neighboring tokens and phrases, **BERT Aug. Formal** ends up correctly predicting the sentence as non-hate, with overall higher confidence than **BERT + SOC ID** model. This clearly shows the difference in mitigation effects of each model.

(a) BERT



(b) BERT Aug. Formal

(c) BERT + SOC ID

Figure 10 Hierarchical explanations on a test instance from PETITION before and after mitigation, where False Positive predictions are corrected.

Finally, I check if the models attribute less importance score on a token “쿵쾅,” which although is neutral in terms of hate speech, usually has a negative

connotation, after mitigation. I examined the post-hoc explanations of the sentences containing the token “쿵쾅” and found that the importance score of the single token has not dropped dramatically, but the slight lowering of the score did contribute to the compositional effect when agglomerated with neighboring tokens. Figure 11 show the results of the explanation of three models: baseline BERT model, **BERT + SOC ID, BERT Aug. Formal** on the sentence “아이가 사는 집도 아닌데 쿵쾅쿵쾅 밤낮없이 뛰어대는데 이젠 한계가 왔습니다.” Bothe mitigation methods (b-c) attribute lower scores to the phrase “쿵쾅쿵쾅 밤낮없이” than the baseline model.

(a) BERT

(b) BERT Aug. Formal

(c) BERT + SOC ID

Figure 11 Hierarchical explanations on a test instance from PETITION that contains a token “쿵쾅.” “쿵쾅” although is a non-hate term, it usually carries a negative connotation.

Chapter 6. Conclusion

In this work, I built adversarial test sets with corpora whose domains are different from user-generated comments, the most popular domain used for training hate speech classifiers. I established that the BERT fine-tuned classification model trained with the existing Korean hate speech corpus detects false positives when the sentence includes specific keywords, due to the unintended bias in the training data. However, from the performance difference of the same model on different test domains, we can infer that the model also attends to other features of sentences that are related to the sentence domain.

I presented two different approaches to address the problem: adding additional negative samples that contain neutral keywords of interest, and regularizing the model's explanation score of such keywords. Although there are some variations of performance for different domains, overall, both methods were effective in reducing false positives in out-of-domain data, while maintaining or improving in-domain performance. Using Sampling and Occlusion (Jin et al., 2020) hierarchical explanation, I qualitatively compared how different models improve the robustness on out-of-domain data.

This work is the first to systematically diagnose and mitigate the false positive bias (or just bias in general) of Korean hate speech classifiers. Hopefully, this work sheds light on how to construct training data for hate speech classifiers for different domains.

Due to the limitation of available data that meets the experimental condition, this work has confined the size of the data added for resolving imbalance in the

training data. For future work, I hope to examine the effect of linearly increasing the size of the data for augmentation. I also hope to test the effect of the mitigation methods on hate speech datasets with implicit hate speech, which requires a more complex understanding of the context and nuance.

References

Arango, A., Pérez, J., & Poblete, B. (2020). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). In *Information Systems* (p. 101584). https://doi.org/10.1016/j.is.2020.101584

Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.14050

Crawford, K. (2017). The trouble with bias. *Conference on Neural Information Processing Systems, Invited Speaker*.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1703.04009

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1810.04805

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, *51*(4), 1–30.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1803.02324

James Murdoch, W., Liu, P. J., & Yu, B. (2018). Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. In *arXiv [cs.CL]*.

arXiv. http://arxiv.org/abs/1801.05453

Jin, X., Wei, Z., Du, J., Xue, X., & Ren, X. (2019). Towards Hierarchical Importance Attribution: Explaining Compositional Semantics for Neural Sequence Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1911.06194

Kabbara, J., & Cheung, J. C. K. (2016). Stylistic transfer in natural language generation systems using recurrent neural networks. *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, 43–47.

Karan, M., & Šnajder, J. (2018). Cross-domain detection of abusive language online. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 132–137.

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., & al., E. (2018). *The Gab Hate Corpus: A collection of 27k posts annotated for hate speech*. https://doi.org/10.31234/osf.io/hqjxn

Kennedy, B., Jin, X., Davani, A. M., Dehghani, M., & Ren, X. (2020). Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.02439

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980

Li, J., Monroe, W., & Jurafsky, D. (2016). Understanding Neural Networks through Representation Erasure. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1612.08220

Lillian, D. L. (2007). A thorn by any other name: sexist discourse as hate speech. *Discourse & Society*, *18*(6), 719–740.

Liu, S., & Forss, T. (2015). New classification models for detecting Hate and Violence web content. *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, *01*, 487–495.

Moon, J., Cho, W. I., & Lee, J. (2020). BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.12503

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*, 145–153.

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. (2018). The Effect of Extremist Violence on Hateful Speech Online. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://ojs.aaai.org/index.php/ICWSM/article/view/15040

Park, J. H., Shin, J., & Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1808.07231

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., … Cho, K. (2021). KLUE: Korean Language Understanding Evaluation. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2105.09680

Pavlick, E., & Tetreault, J. (2016). An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, *4*, 61–74.

Pelletier, F. J. (1994). The Principle of Semantic Compositionality. *Topoi. An International Review of Philosophy*, *13*(1), 11–24.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis Only Baselines in Natural Language Inference. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1805.01042

Richardson-Self, L. (2018). Woman-Hating: On Misogyny, Sexism, and Hate Speech. *Hypatia*, *33*(2), 256–272.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2019). Social Bias Frames: Reasoning about Social and Power Implications of Language. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1911.03891

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.

Singh, C., James Murdoch, W., & Yu, B. (2018). Hierarchical interpretations for neural network predictions. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1806.05337

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 940–950.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. *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)*, 602–608.

Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *Proceedings of the 17th International Conference on Distributed Computing and Networking*, 1–6.

이원석, & 이현상. (2020). 딥러닝 기술을 활용한 차별 및 혐오 표현 탐지: 어텐션 기반 다중 채널 CNN 모델링. 한국정보통신학회논문지, *24*(12), 1595–1603.

국문초록

온라인 등 익명 공간에서의 혐오 발언(Hate speech)으로 인한 피해가 커져감에 따라, 혐오 발언 분류 및 검출에 관한 연구가 활발히 진행되고 있다. 최근 딥러닝 기반의 혐오 발언 분류기가 좋은 성능을 보이고 있지만, 학습 도메인 밖(out-of-domain) 데이터로 일반화함에 있어서는 어려움을 겪고 있다. 본 연구는 모델이 거짓 양성(False Positive)을 검출해내는 문제에 초점을 두고, 해당 문제를 진단하기 위해 세 가지 서로 다른 도메인의(domain)의 대립적(adversarial) 데이터를 활용하여 테스트셋을 만든다. 이를 통해 기존의 한국어 혐오 표현 데이터셋을 학습한 BERT 기반의 분류 모델이 학습 데이터 상에서 혐오 표현과 높은 상관관계를 가지는 특정 단어들에 민감하게 반응하여 거짓 양성(False Positive) 결과를 예측하는 현상을 보인다. 다음으로, 이를 해결하기 위한 두 가지 방법을 제시한다. 학습 데이터셋의 불균형을 수정하기 위한 데이터를 추가하는 데이터 중점(data-centric) 방법과 특정 단어들에 대한 모델의 사후 설명(post-hoc explanation)을 활용하여 모델을 정규화(regularize) 하는 모델 중점(model-centric) 방법을 적용하고, 두 접근 방법 모두 전반적인 모델 성능을 해치지 않으며 거짓 양성의 비율을 줄일 수 있음을 보인다. 또한, 테스트 도메인의 특성을 알고 있을 경우, 유사한 도메인에서 학습 데이터의 불균형 수정을 위한 샘플 추가를 통해 적은 비용으로 모델의 거짓양성을 큰 폭으로 줄일 수 있음을 보인다. 또한, Samping and Occlusion (Jin et al., 2020) 설명을 통해 두 접근 방식 모두에서 문맥 정보를 더 잘 활용하게 됨을 정성적으로 확인한다.

주요어: 혐오 표현, BERT, 혐오 표현 데이터셋, 데이터셋 구축, 거짓 양성, 편향 측정, 편향 완화, 도메인 밖 데이터