How Does the Hate Speech Corpus Concern Sociolinguistic Discussions? A Case Study on Korean Online News Comments

Won Ik Cho
Dept. of ECE and INMC
Seoul National University
tsatsuki@snu.ac.kr

Jihyung Moon
Spekt
jihyung.moon@spekt.to

Abstract

Social consensus has been established on the severity of online hate speech since it not only causes mental harm to the target, but also gives displeasure to the people who read it. For Korean, the definition and scope of hate speech have been discussed widely in researches, but such considerations were hardly extended to the construction of hate speech corpus. Therefore, we create a Korean online hate speech dataset with concrete annotation guideline to see how real world toxic expressions concern sociolinguistic discussions. This inductive observation reveals that hate speech in online news comments is mainly composed of social bias and toxicity. Furthermore, we check how the final corpus corresponds with the definition and scope of hate speech, and confirm that the overall procedure and outcome is in concurrence with the sociolinguistic discussions.

1 Introduction

Hate speech is an issue that has permeated deeply into our daily life (ElSherief et al., 2018). Hate speech is often explicitly stated along with insulting expressions, and some of them are perceived as hateful or offensive just by incorporating social bias. Accordingly, public figures or underrepresented groups suffer from tremendous mental damage, while some experience depression or end their lives.

Regarding the hate speech in online spaces, discussions are divided into *definition* and *detection* (MacAvaney et al., 2019) point of view. *Definition* mainly concerns "*What the hate speech is*" in a deductive manner, while *detection* attacks the issue with more an inductive methodology. For Korean, the hate speech study has mainly been conducted in the sociolinguistics community regarding the definition, and these include the discussion on the appropriateness of expression "hate speech" itself (Hong et al., 2016), its scope (Kim, 2017b), and the legal issues around discrimination and insult (Park and Choo, 2017). In Hong et al. (2016), hate speech is defined as "an expression that discriminates/hates or instigates discrimination/hostility/violence for some social minority individual/group". To back up these studies, further discussions on the underrepresentedness of each group have also actively taken place (Kim, 2017a, 2018).

However, aside from the importance of such discussions, there is a gap between the theoretical definition of *hate speech* and real *hateful expressions* that appear in our lives (Davidson et al., 2017). From a detection perspective, the following questions are mainly discussed which are not easy to answer in a definition point of view: "Should a certain expression be regarded as *hateful* even if a majority of people do not feel offensive for the same sentence?", "What if for the pre-existing terms that a small group of people insists on its harm?", and "How about the toxic expressions that head the criminals?". If there is a consensus on these issues, collecting data to develop a model for hate speech detection would be more clear.

Most previous approaches on Korean online hate speech detection have been keyword-based that regards glossaries on profanity terms[1] (Kang, 2018; Park and Cha, 2018). It is also challenging to find cases of constructing a corpus referring to preceding researches in other cultural regions (Waseem and Hovy, 2016; Davidson et al., 2017; Basile et al., 2019). Therefore, to understand how hate speech is represented in the Korean online expressions and how the inductive analysis corresponds with the concurrent discussions, we should investigate the attributes of hate speech and construct a corpus in advance.

[1] https://github.com/doublems/korean-bad-words

Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 13–22
co-located with ICON-2021, December 16-19, 2021, Silchar, India.
https://doi.org/10.26615/978-952-94-5833-2_003

In light of this, we study on a hate speech corpus construction scheme that reflects the characteristics of the Korean expressions. Recent works on hate speech corpus have considered social bias as one of the hate components (Waseem and Hovy, 2016; Assimakopoulos et al., 2020) as the hypothesis that bias and hate are closely related (Boeckmann and Turpin-Petrosino, 2002) but hardly labeled it. Though Sanguinetti et al. (2018) performed the most decent work on Italian, we wanted to give more fine-grained analysis on the social bias and stereotype. Referring to the prior works, we create an annotation guideline for Korean entertainment news articles comments where hate speech issues have been prevalent in recent years (McCurry, 2019b,a), and construct annotated corpus through crowd-sourcing. Specifically, we describe bias and toxicity as two main attributes of hate speech and label each of them with three-fold categories.

Throughout the paper, we present the annotation guideline built upon the observation of the comments and corpus construction scheme based on crowd-sourcing. Then, we introduce the corpus characteristics and check whether our procedure and result are adequately accepted within the preceding hate speech studies. The contribution of our study to the field is as follows:

- We observe social bias and toxicity convey hate speech in Korean online news comments and build the hate speech annotation guideline, making the annotated corpus and guideline publicly available.[2]
- We conduct an analysis to find the correspondence of our inductive approach with the preceding sociolinguistic discussions.

2 Data

2.1 Language and Domain

The language of interest in this paper is the Korean online expressions, which are generally variations of written Seoul dialect. We target the news comments that show a lot of informal expressions that are difficult to face in the Sejong corpus (Kim, 2006) or Wikipedia.

For domain, we took a look at the violence of the entertainment news article comments, which triggered the comment overhaul in Korea (Yim, 2020). In entertainment news, not only a target is specified, but also the target is perceived as a representation of a certain group of people, so that many utterances including social bias may appear. Besides, the legal system to regulate hate speech for them has not yet been well settled, and above all, the mental suffering of figures targeted by hate speech is severe.

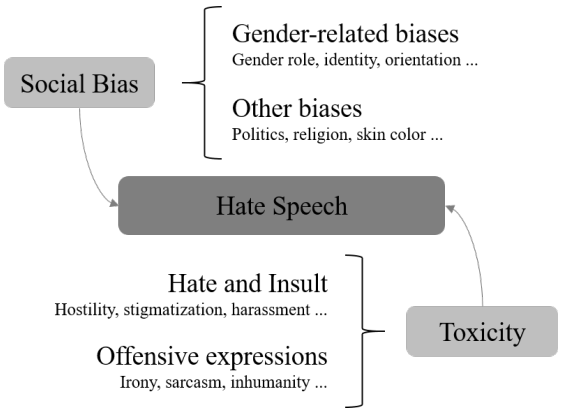

Figure 1: Hate speech attributes in our guideline

2.2 Crawling and Sampling

The articles were crawled for about two years from January 2018 to February 2020, namely the daily top 30 of the online portal news platform, yielding 23,700 articles and 10,403,368 comments.

The following pre-processing was executed to filter the comments.

1. Two articles are sampled daily to avoid bias in a certain period
2. Select the top 20 comments for each article based on the Wilson score (Wilson, 1927) for the downvote
3. Remove duplications
4. Remove (potentially ambiguous) single token comments
5. Remove long sentences over 100 characters

A total of 10,000 comments were collected during this process, along with the head and body of the article, upvotes and downvotes per comment, and a timestamp.

3 Guideline

To create a guideline for large-scale human annotation, three Korean speakers with a computational linguistics background, all familiar with the Korean online materials, read about 1,000 comments

[2]The corpus is disclosed in https://github.com/kocohub/korean-hate-speech with a dataset paper (Moon et al., 2020) and this study focuses more on the annotation guideline, examples, and analysis regarding socioloinguistic discussions.

together, and labeled simultaneously. In our observation, the hate speech that appears in entertainment news comments is judged in terms of "social bias" and "toxicity", as illustrated in Figure 1. Each attribute is identified as follows:

- **Existence of gender-related or other kinds of social bias**[3]
- **Amount of toxicity exposed by hateful, insulting, or offensive expressions**

The main difference between toxicity and bias is that the toxicity is determined based on its intensity (Davidson et al., 2017; Park and Choo, 2017), while for bias, the existence is relatively apparent compared to the former.

3.1 Social Bias

We observe which social bias is disclosed or implicated in each comment. Here, social bias refers to a hasty generalization, stereotyping, or prejudice that an individual/group with a specific social identity will display some characteristics or act in a biased way (Song et al., 2001; Keene, 2011; Blodgett et al., 2020). In the annotation process, the question was labeled in three slots, which are as follows.

3.1.1 Gender bias

Does gender-related bias exist explicitly or implicitly in the text? We considered gender-related biases as primary factors of the social bias, for their prevalence in the Korean online spaces[4] (Kim, 2017b; Lee and Park, 2019). The utterances corresponding to this category include gender role, sexual orientation and identity, and biases for gender-related ideas (Hong et al., 2016; Kim, 2017b). For example, '*Wife must be obedient to the husband's words*' and '*Gays will be vulnerable to disease*' belong to this category. Also, even when gender-related and other biases exist simultaneously in the text, we judged that the case belongs to this category.[5] Brief examples of the subcategories are provided in Korean with English translation, split with '/'. WARNING: This part contains contents that may offend the readers.

- **Prejudice on roles or abilities according to gender**:
  남자는 능력이고 여자는 외모지 *Beautiful women for capable men* / *여자는 집에서 살림하는게 최고, 남자가 부잔가? It's the best for women to stay at home and make beds. The dude's gotta be rich..?*

- **Gender and age**:
  남자가 나이많은 여자를 만날 이유가있나..? 1. 돈줄 2. 성욕해소 말고는.. 글쎄 결혼은 어린여자랑 하는게 맞지 *Why should guys see old women anyway..? 1. for money 2. for spunk release, maybe.. Young ladies are the marriage material*

- **Prejudice against groups with specific gender, sexual orientation, sexual identity, and gender-related ideas**:
  레즈비언들이 좋아할만한.. *Likely to appeal to Lesbians...* / 오래도살았네 게이가 *Talk about longevity for a gay man* / 성전환자가 여자냐? *Transie's a chick?* / 페미들 ㄹㅇ 토나온다 *Feminazis make me sick*

3.1.2 Other biases

Does there exist non-gender-related biases explicitly or implicitly in the text? The utterances of this category incorporate a bias for the factors regarding other social characteristics. In other words, we count the cases that extend an individual's characteristic to a property of a group, so that the speaker's bias towards the group is revealed. These factors include race, ethnicity, nationality, social background, political stance, skin color, religion, disability, age, appearance, richness, occupation, education, and military experience (MacAvaney et al., 2019). Examples of each are as follows.

- **Age**:
  나이에 걸맞게 쳐 놀아라. 이제 마흔이 넘었는데 언제까지나 귀여운 척 할래 ㅉ *Act your freaking age.. Isn't the age of forty a little too old to act all cute?*

[3] In Korean, there are various social deixes that differ according to gender or age, which are commonly used without pejorative purpose. However, there are social movements that try to change or normalize such terms in order to prevent probable offensiveness coming from being referred to as a specific class of gender or age. Since these discussions are ongoing, we could not take into account such sides in making up the annotation guideline. Nonetheless, we believe the change of language may have to be considered as a future improvement of this research.

[4] We first considered that the gender factor had been a central issue regarding hate crime in the previous studies (Jenness, 2003), but some other reasons are to be supplemented in Section 5.4.

[5] This is in concurrence with our decision to separately tag the gender-related bias, and also aims at making the problem a classification.

- **Region of origin**:
  스테이크에 와인한잔하면서 가족끼리 회의했겠지 상도랑 라도는 인생살면서 걸러야한다 *Topic on the table at the family gathering while wining and dining.. Gotta screen them out those Texans and Mississipians*[6]

- **Appearance**:
  돼지같은 것들은...다 이유가 있다 *A pig's a pig for a reason* / 이 외모면 퇴사해도된다 *She doesn't need a job when she looks like that*

- **Occupation**:
  한효주가 뭐가 아쉬워서 사람 많은 클럽에서 침까지 질질 흘리며 콧물까지흘리며 마약을 했을까. 이미지 너무 하락하겠다. *What was wrong with Hyo-Joo Han, for drooling in a crowded club and using drugs? the image will fall too much.* / 존못인데 연기자로서는... *Ugly face, but her acting's okay*

- **Political stance**:
  또 자한당 짓이냐?ㅋ *The Libertarian Party, again. . . ??*[7] / 역시 대깨문답네 *A Trumpist, indeed*[8]

- **When prejudice against a group is involved in judging an individual**:
  무도에서 조금 얻은 인기로 여자 만나고, 그러다 점점 일은 생겼고, 이제는 숨길수 없고 *Get some reputation from the Infinite Challenge*[9]*, then get some girls, then things get messed up, cannot hide* / 경수진 YG소속이네ㅉ YG는 무조건 믿고 거른다ㅉ *Damn, Gyeong Sujin is in YG? B-bye now! I detest all them YGs*[10]

3.1.3 None

The utterances in this category refer to the comments that do not correspond to the two categories above. In the text, prejudice against a group with specific social characteristics is not intervened to judge the group or the individual that belongs to it.

[6]These were originally the profanity terms that denote two main non-capital provinces of Korea, 상도 *Sangdo* and 라도 *Rado*, known to have opposed political stances.

[7]The original expression is 자한당 *Jahandang*, a previous conservative party of Korea.

[8]The original expression is 대깨문 *Daekkaemoon*, initially used for the supporters of the president in Korea, but now used as a swear term to stigmatize and ridicule them.

[9]A variety show of Korea.

[10]YG is a famous entertainment company in Korea, notoriously known for some crimes committed by the members.

3.2 Toxicity

The second attribute is how toxic the comment is, either considering the speaker's intention or the influence on the readers (Wulczyn et al., 2017). The degree of toxicity is a subjective measure, and it is difficult to avoid the annotation being influenced by the annotator's experience and linguistic intuition in this issue. However, to determine the boundary as precise as possible as in Davidson et al. (2017), we categorized the 'intensity' as follows.

3.2.1 Hate or Insult

Is there a strong hate or insult for the target of the article, related figures, or other people? Hate can be seen as an expression in which adversarial and aggressive views towards the aforementioned social characteristics are observed (Hong et al., 2016; Park and Choo, 2017), and insult is a mean expression that can seriously impair the social prestige of a specific figure or group (Kim, 2013).[11]

Here, hate is utilized as a bit different from the one in 'hate speech', which is a slightly more abstract concept. For instance, we can tell that the *toxic comments* are where the *hate speech* is displayed in online spaces, and in deciding some comments as *hate speech*, the attribute of *toxicity* might be taken into account if they contain *hate* or *insult*. Therefore, an expression may be categorized into this type for only including some swear words.

To make this clear, we checked the following properties.

- **Expression that can cause mental pain by severely criticizing or deterring an object**:
  노래실력 제일 거품인 색기 *The worst vocal ever* / 돼지 어쩌구였는데 지워짐 흑흑 ㅜㅜ *What a lardbag* / 노잼에 늙은성 괴들 처노는 방송 *This show is a total bore filled with old nip tucks*

- **Sexual harassment or objectification**:
  겨드랑이도 빨겠다 *I'd even lick her armpit* / 스타킹 벗겨서 발가락빨구시퍼용. *Wanna take off her stockings and suck on those toes* / 보팔 *Pussy chaser*

- **Comments containing hostile feelings toward individuals or groups based on the innate characteristics of individuals/groups**:

[11]Note that the definition here mainly follows the cognates of *hate* and *insult* in Korean articles, while also taking into account the global standard such as Facebook, Youtube, and Twitter.

성 정체성을 잃어가는 병자들이 많은 시대네...병은 고쳐야지 자랑이라고 떠들어 대나? *It is the age of sickos without any proper sexual identity. They need a cure, not a chance to brag about themselves* / 여기 성별에 댓글만봐도 한남 믿거할수있겠다 *I can just look at the gender label and the contents of the replies to pick out the worthless pricks*

- **Expressions intended to negatively stigmatize or define a specific individual/group**:
  홍윤화메갈 ?[12] *Hong Yoon Hwa a feminazi?* / 애국보수 산이의 음악행보룰 전폭 지지하는 바입니다. *STAN for San-E and his music career, the true republican nationalist* / 참 대단하다 탑게이 *Applause for the gay lord*[13]

- **Exhibition of the notorious factual events**:
  ㅋ 접대와 조작의 아이콘 아이즈원 엑스원 *IZONE X1,*[14] *THE ICON for manipulations and booty calls*

- **Comments showing hostility towards other users**:
  언제까지 반일 감정에 불탈래 막상 역사도 그렇게 모르는 개돼지들이 꼭 흥분하더라 *Till when will you be stuck with anti-Japaneseism Morons without any historical knowledge always bark the loudest*

- **Comment showing hostility towards the journalists who wrote the article**:
  기레기새끼.의식불명될때까지 쇠파이프로 대가리 깨야됨 *Newshounds*[15] *need to be bashed in the head to the point of unconsciousness* / 인턴기자라는 것이 인턴때부터 제목 좟같이 뽑아서 조회수 늘릴 꼼수를 부리고 있네. 아주 싹이 노랗다 못해 형광색일세. 기레기 꿈나무에 캬얄 퉷 *What a trashy title to come from an intern journo, just dying to get more views. It's too transparent I can even see through it. Here's a finger for the future newshound*

3.2.2 Offensive expressions

Does not reach Insults or hate, but contains aggressive and rude content? The toxicity of these utterances is less than that of hate or insult, but the contents can still make listeners feel offensive. It is expected to be represented by the following properties.

- **Ironic and rhetoric expressions**:
  짠내투어 멤버로 랩퍼 도끼를 추천합니다. 근검절약의 아이콘 이시더라고요 *Recommend Doki as a member for Salty Tour.*[16] *Heard he's the man of frugality*

- **Inhumane expressions**:
  ? 정말 좋아하는 배우 였는다... 가셔서 행복하시길 바라고 갈때가더라도 돈좀 주구가.. 그게 아니면 로또1등좀. *Really liked him/her as an actor/actress.. Well farewell, godspeed, and oh drop some money in my pocket? Or the lottery winning numbers?*

- **Cynical or guessing expression**:
  이분 빚투나오는거 아닌지.. 다갚으시고 집자랑 하신거겠죠.. *I'm afraid there will be a #ILoanedHimMoneyToo for the guy. You did pay all your debts before showing off your crib like that, right?* / 송사끝나서 후련한 마음에 동남아 좀 갔다고 뭐 문제라도 있음? *Work complete, you feel nice about it, so took up a trip to Southeast Asia. What's the big deal?*

- **Expressions that can make someone feel bad or demean them**:
  무슨 다들 작가들 납셨나봄ㅋㅋㅌㅌㅌㅋ 제발 보지마라 씨부릴거면 ㅋㅋㅋㅋ 각자 취향에 다른거지 난 좋음ㅇㅇ *Wow y'all must be the screenwriters lolololololol just stfu and don't watch it lolololol I like it, everyone's got different tastes. I like it.* / 누군데 얘네? *So who are they?*

- **Comment with no hate, but with abusive language such as swear words**:
  한채아를 감히... ㅅㅂ *Fuck, who dare did her?*

3.2.3 None

These refer to comments that do not meet the above toxicity. Even if there is criticism, it is judged as a tolerable opinion in case there is no offensive or rude content. Toxicity is hardly observable in the instances that belong to this.

[12] 메갈 *Megal* is a stigmatized term for the feminists in Korea.

[13] Originally 탑게이 '*Top Gay*', a term that a Korean homosexual entertainer first used to introduce himself.

[14] A Korean Idol group that has been suspected for their agency manipulating the voting system of TV Pro.

[15] The original expression is 기레기 *Giregi* which is a compound of *garbage* and *reporter*.

[16] 도끼 *Doki* is a rapper who is famous for showing off his richness, and 짠내투어 *Salty Tour* is a TV program that aims to travel with as little money as possible.

4 Annotation

The annotation guideline described above was primarily constructed through the analysis of the observed 1,000 comments. However, to help the annotators refer to it in tagging the large-scale corpus, utilization of the crowd-sourcing platform was inevitable. The three factors considered in this process are: 1) *whether the platform has a sufficient number of potential workers*, 2) *whether our guideline can be well taken into account in the annotation process*, and 3) *if the annotation can be performed by the workers that exhibit an expected ethical standard*. Based on these, we adopted *DeepNatural AI*[17] that accommodates a variety of participants, allows pilot study for the selection of the annotators, and supports the system for checking whether the feedback to the annotators is reflected in the resubmission.

Labeling was performed for each attribute through the pilot study and crowd-sourcing, and the decision was made with majority voting. For this, the tagging of three participants were guaranteed for each instance. Additional adjudication was conducted in cases where all three annotators tagged differently or the answers were significantly divided (e.g., when there was no choice in the middle area for the tagging over toxicity).

4.1 Pilot Study

In order for workers of the crowd-sourcing platform to participate in large-scale corpus construction, a pilot study must be performed to ensure the appropriateness of their labeling. We used randomly selected 1,000 comments that were not exploited to make up the guideline, to select the workers who understood our guideline well. The detailed checklist is as follows.

- Is the number of tagging performed more than a certain standard (e.g., 30)?
- Wasn't the omission of tagging too frequent?
- Was the tagging consistently done for challenging instances?
- Was the feedback on the rejected work well reflected in resubmission?
- Isn't the participant exhibiting particular criteria, for gender and other factors, that have significant gaps with the given guideline?

[17] https://app.deepnatural.ai/

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

Table 1: The composition of the constructed corpus.

4.2 Crowd-sourcing

We conducted the annotation of left 8,000 comments executed by eight selected participants. Unlike the pilot study that the authors reviewed, rejected, and accepted in a case-by-case manner for selecting the participants, the annotation of the participants was performed on the platform without further restriction.

5 Corpus

The final dataset comes from a total of 10,000 instances, namely those exploited in making up the guideline, the instances reviewed and accepted through pilot studies, and the rest 8,000, crowd-sourced through the annotation of the selected participants. In this process, 659 cases that did not reach the final agreement or was omitted by the participants were dropped.

5.1 Agreement and Performance

The agreement was calculated based on the corpus after adjudication, and based on this, an inter-annotator agreement (IAA) was calculated with Krippendorff's alpha (Krippendorff, 2011). At this time, the agreement on social bias was divided into a binary case that only checks the existence of gender-related bias and a ternary case that separately checks the existence of other biases. The task that detects the existence of a gender-related bias (*gender bias*) shows a relatively high agreement (0.765) compared to the other two cases, while the other two ternary tasks (*social bias* and *toxicity*) showed a moderate but slightly more uncertain label decision (0.492 and 0.496, respectively). The model performance using baseline deep learning architectures is provided in the original dataset paper (Moon et al., 2020), showing the best F1 score of about 0.63 for social bias and 0.58 for toxicity (ternary classification). The agreement and performance show the validity of the proposed corpus.

5.2 Composition

The composition of the whole corpus is as shown in Table 1. Overall, toxic instances occupy a larger volume than those which are not, while the portion of the instances with social bias is comparably smaller than their counterpart. However, it is hesitant to conclude that the toxic comments are more visible in the entertainment news domain than the comments with the bias, since we had collected the comments according to the portion of the downvote. Instead, it may more make sense to interpret that toxicity more influences on judging the comment, compared to the social bias factors which is usually implicated within.

5.3 Analysis

One of the points worth paying attention to is that most of the comments regarding gender-related or other biases are at least offensive or disclose hate/insult in general (toxic among the comments with gender-related bias: 93.76%, toxic among the comments with other bias: 90.42%). On the other hand, it was observed that the toxic comments were not necessarily the ones implicating the social bias.

However, another tendency is displayed in between each social bias type. We were able to discern from the results that the gender-related bias could boost the intensity of the toxicity. That is, we checked that in the comments with higher toxicity (namely hate and insult), the gender-related bias is disclosed about 40% more frequent than other biases (10.15% to 7.48%), while in less toxic (offensive) comments, the tendency is reversed (4.58% to 8.94%).

5.4 Why Gender-related?

The result in Section 5.3 is in concurrence with our premise in the guideline that the gender-related hate speech is more prevalent (Kim, 2017b; Lee and Park, 2019), which is assumed to be in connection with the cultural background (Kim and Lowry, 2005; Koh, 2008; Prieler, 2012). First, as stated in Section 3, considering that the corpus concerns entertainment news article comments which are often in less correlation with other political or social issues, our guideline placed more attention on the gender-related issues, which differs from the previous study that has considered stereotype (Sanguinetti et al., 2018) yet in a binary manner. We directly or indirectly recommended intolerance for gender-related content, for example, by categorizing them separately (social bias) or citing them as a representative example (sexual harassment and sexual insults).

Our approach does not contradict the current data-driven hate speech studies on other languages (Waseem and Hovy, 2016; Fortuna et al., 2019). Our policies were set on purpose since the gender-related factors can influence readers more universally than other contents (such as politics, religion, financial power, etc.), in that the properties are often innate and determine one's identity. That is why other identity factors such as nationality and ethnicity are also crucially investigated in international studies where multiculturalism plays a more significant role (MacAvaney et al., 2019). Those factors are to be further specified and developed in the future guideline.

6 Discussion

Our guideline aims at making the blurry boundary between hate speech and freedom of speech more explicit in order to attack the real-world problem. In this section, we extend how this categorization process can connect with "*which expressions are sociolinguistically defined as hate speech*". We refer to Hong et al. (2016), Kim (2017b), and Park and Choo (2017), where each mainly concerns the definition of hate speech, its target and scope, and the legal issues regarding freedom of speech.

6.1 Definition

Previous studies As stated earlier, in Hong et al. (2016), hate speech is defined as "an expression that discriminates/hates or instigates discrimination/hostility/violence for some social minority individual/group", which follows the National Human Rights Commission of Korea, closer to the European definition (No, 15). Accordingly, the types of hate speech fall into four categories: 1) discriminative bullying, 2) discrimination, 3) disclosed contempt/insult/threat, and 4) hate incitement, where 1-2) are in concurrence *social bias* and 3-4) with *toxicity* defined in this study. Hong et al. (2016) attempts to discern "*this is hate speech*" rather than "*what the hate speech is*", and we expand such factors to the process of corpus construction.

Our approach As described in the guideline, we take the social bias (stereotype, prejudice) and toxicity (hate, insult, contempt, threat) as main attributes, which comes from the typological definition in the deductive approaches (Hong et al.,

2016; Kim, 2017b). The results in Section 5 further suggests that social bias is likely to accompany the toxicity. This is also in concurrence with the discussion that 'discrimination', which is an act of making a distinction based on human identities, is a core factor of hate speech that can represented by bullying, contempt, threat, etc.

6.2 The Borderline of Hate Speech

Previous studies In a slightly distinguished view, Park and Choo (2017) focuses on clarifying the boundary of freedom of expression and hate speech, and aims to establish a principle that can regulate hate speech while minimizing the infringement of freedom. To be concrete, the actual legal cases are examined regarding insults or hate speech, considering which expressions are acknowledged as *violation*. Similarly as in the *definition*, Park and Choo (2017) finds it is challenging to define which expression is clearly illegal. However, Park and Choo (2017) emphasizes the freedom of speech should not be used to justify the attack towards minority or underrepresented groups, and any expressions that infringe on the dignity or personal values of others should be restricted.

Our approach In the previous study, the freedom of speech was taken into account in judging the hate speech (Park and Choo, 2017). In contrast, we made a decision on the toxicity assuming we were the target figure, for instance, how the expression may insult, offense, or mentally harm the addressee. However, since putting on one's shoe is difficult, three annotators' opinion on such perception were aggregated to make up the decision.

One challenging example was the comment that quotes a female celebrity as a sexually attractive figure. This may harass the ordinary female addressee, but since we had limited knowledge of how the target might perceive it, we had to leave it to the annotators' decision. This kind of ethical or social perception is highly dependent on linguistic intuition and experience, and it is also challenging to find well-defined ground truth. We attempted to carefully draw the borderline of biasedness and toxicity in the pilot study, crowd-sourcing, majority voting, and adjudication to guarantee freedom of speech while restricting the social harm (Park and Choo, 2017).

6.3 Minority

Previous study Kim (2017b) refers to the definition of Hong et al. (2016), and describes an *objectiveness* of minority and *mental harm* received by listeners/targets of hate speech. The focus here is whether an utterer denies the identity of the victim via hate speech. It emphasizes the importance of preventing potential victims from facing such violence in open space, and that the society-level response is urgent to this issue.

Our approach In our study, bias and toxicity towards even social dominants (males, the rich, and so on) were regarded as hate speech. Though this may not be harmonized with the previous study (Kim, 2017b), we argue that the mental harm triggered by the hate speech should not be masked out by whether the target is in the majority group or not. Also, the underrepresented group is not always fixed and can change by era.

More importantly, the justification on the bias and toxicity towards the privileged could make unexpected model bias. For instance, what if one indiscriminately insults a public figure just because s/he is rich or educated? How should we handle the inhumane reaction when the victim of a tragedy is male, as in "*Good man is a dead man*"? Again, confining the objectiveness of hate speech to certain minorities may not help detecting real "hateful expressions" from which the victims might suffer. This is also intertwined with the way we draw the borderline, and putting on one's shoe plays an important role here.

7 Conclusion

Throughout this study, we investigated which factors result in hate speech in an inductive way. The hate speech found in Korean online news comments contains either social bias, toxicity, or both. We built a guideline upon the findings and constructed a dataset to train a model that automatically detects them. Furthermore, we refer to the previous discussions on hate speech treated in sociolinguistics and journalism, to see how our approach is related to them and what the distinguished points scrutinized in our approach are.

As a follow-up study, we verified how effective this corpus is as input data for real-problem-solving machine learning model (Moon et al., 2020), and will check whether its detection performance is affected by (maybe biased) pre-trained language models. Also, we will investigate how the construction scheme of our corpus can be leveraged in other domain such as depressive online text detection

(Hämäläinen et al., 2021). Besides, we expect that by the release of this corpus, Korean hate speech research is to be diversified and that real online space might be detoxified.

Acknowledgments

Authors thank Jumbum Lee for creating the dataset together and appreciate Hyunjoong Kim for aiding the crowd-sourcing process. Also the authors are grateful for encouraging comments from two anonymous reviewers.

References

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. pages 5088–5097.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

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

Facebook. Facebook's policy on hate speech. https://www.facebook.com/communitystandards/hate_speech. Accessed: 2020-10-06.

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

Mika Hämäläinen, Pattama Patpong, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Detecting depression in Thai blog posts: a dataset and a baseline. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 20–25, Online. Association for Computational Linguistics.

Sung Soo Hong et al. 2016. *Study on the State and Regulation of Hate Speech*. National Human Rights Commission of Korea.

Valerie Jenness. 2003. Engendering hate crime policy: Gender, the "dilemma of difference," and the creation of legal subjects. *Journal of Hate Studies*, 2(1):73–92.

Seungche Kang. 2018. *A study on constructing dictionary for Korean hate speech classification: Focusing on online news comments*. Korea Advanced Institute of Science and Technology.

Sabrina Keene. 2011. Social bias: Prejudice, stereotyping, and discrimination. *The Journal of Law Enforcement*, 1(3):2–4.

Bo-Myung Kim. 2018. Late modern misogyny and feminist politics: The case of Ilbe, Megalia, and Womad. *Journal of Korean Women's Studies*, 34(1):1–31.

Doo Sang Kim. 2013. A study on the regulations of defamation and insult on cyberspace. *The Journal of Legal Studies*, 21(1):175–196.

Hansaem Kim. 2006. Korean national corpus in the 21st century Sejong project. In *Proceedings of the 13th NIJL International Symposium*, pages 49–54. National Institute for Japanese Language Tokyo.

Jinsook Kim. 2017a. # iamafeminist as the "mother tag": Feminist identification and activism against misogyny on Twitter in South Korea. *Feminist Media Studies*, 17(5):804–820.

Kwangok Kim and Dennis T Lowry. 2005. Television commercials as a lagging social indicator: Gender role stereotypes in Korean television advertising. *Sex Roles*, 53(11-12):901–910.

Sooah Kim. 2017b. Expression of hate and discrimination in the Korean language from a social viewpoint: Problem statement and improvement measures for hate and discrimination against social minorities. *New Korean Language-life*, 27(3):49–63.

Eunkang Koh. 2008. Gender issues and Confucian scriptures: Is Confucianism incompatible with gender equality in South Korea? *Bulletin of the School of Oriental and African Studies*, 71(2):345–362.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-reliability. *Computing*, 1:25–2011.

Young-Joo Lee and Ji-Young Park. 2019. Emerging gender issues in Korean online media: A temporal semantic network analysis approach. *Journal of Contemporary Eastern Asia*, 18(2):118–141.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Justin McCurry. 2019a. K-pop singer Goo Hara found dead aged 28. *The Guardian*.

Justin McCurry. 2019b. K-pop under scrutiny over 'toxic fandom' after death of Sulli. *The Guardian*.

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, Online. Association for Computational Linguistics.

ECRI General Policy Recommendation No. 15. on combating hate speech, 8 December 2015.

Da-Sol Park and Jeong-Won Cha. 2018. Semi-supervised learning for detecting of abusive sentence on Twitter using deep neural network with fuzzy category representation. *Journal of KIISE*, 45(11):1185–1192.

Mi-suk Park and Ji-hyun Choo. 2017. *The State of Hate Speech and The Response Measures*. Korean Institute of Criminology.

Michael Prieler. 2012. Gender representation in a Confucian society: South Korean television advertisements. *Asian Women*, 28(2):1–26.

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

Kwan Jae Song, Jae Chang Lee, and Young Oh Hong. 2001. Prejudices and discrimination toward social stigmatized groups. *Korean Journal of Psychological and Social Issues*, 7(1):119–136.

Twitter. Twitter's policy on hate speech. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Accessed: 2020-10-06.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

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

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

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

Youtube. Youtube's policy on hate speech. https://support.google.com/youtube/answer/2801939?hl=en. Accessed: 2020-10-06.