Supporting Online Toxicity Detection with Knowledge Graphs

Paula Reyero Lobo, Enrico Daga, Harith Alani

Knowledge Media Institute, The Open University, UK
{paula.reyero-lobo, enrico.daga, harith.alani}@open.ac.uk

Abstract

Due to the rise in toxic speech on social media and other online platforms, there is a growing need for systems that could automatically flag or filter such content. Various supervised machine learning approaches have been proposed, trained from manually-annotated toxic speech corpora. However, annotators sometimes struggle to judge or to agree on which text is toxic and which group is being targeted in a given text. This could be due to bias, subjectivity, or unfamiliarity with used terminology (e.g. domain language, slang). In this paper, we propose the use of a knowledge graph to help in better understanding such toxic speech annotation issues. Our empirical results show that 3% in a sample of 19k texts mention terms associated with frequently attacked gender and sexual orientation groups that were not correctly identified by the annotators.

Introduction

The detection of toxic language is an active area of study, as there is a need to protect specific demographic groups that have become common targets of online hate (Silva et al. 2016). Much recent research focused on the detection of toxic speech in a binary fashion (toxic, not toxic), using manually produced annotations of such corpora (Poletto et al. 2021). However, few annotation protocols have identified the toxicity targeted groups (Mathew et al. 2021; Sap et al. 2019b; Borkan et al. 2019), even though this information is crucial to address the bias that these systems have shown towards the content of the groups they are meant to protect (Hutchinson et al. 2020; Sap et al. 2019a; Dixon et al. 2018; Davidson et al. 2017).

One of the challenges with using human annotators to identify this information is that toxic texts often need interpretation, which could reduce consistency and reliability in existing benchmark datasets (Poletto et al. 2021). For example, annotators in the commonly used Jigsaw's Toxicity dataset (Borkan et al. 2019) considered *"Its a mental order called gender dysphoria, and who is we?"* to be about a mental illness, whereas gender dysphoria is distress experienced by many transgender people. Without more knowledge, annotators may miss many mentions of frequently attacked groups. In fact, none of the ten annotators who analysed the example above considered it to be related to the LGBT+ community.

Previous work has investigated some factors in data and models, such as the dialect in which toxicity texts are written (Kim et al. 2020) or the features included in the model (Kocoń et al. 2021), which could compromise human annotations. To our knowledge, this is the first work that uses background knowledge about two specific *toxicity targets*, namely gender (G) and sexual orientation (SO), to better understand the problem of missing target information in toxic speech annotations. In particular, we aim to investigate:

- To what extent does a domain knowledge graph help identifying toxic language missed by the annotators?

We use over 14k semantic concepts from the Gender, Sex, and Sexual Orientation (`GSSO`) ontology (Kronk and Dexheimer 2020) to identify texts with mentions of these groups in a toxicity corpus with mentions of other protected characteristics of race, disability, or religion, and no mention of a specific demographic group. We found that 3% in a sample of $\sim$ $19k$ texts mention terms about these groups but were not correctly identified as targets by the annotators. We release the code and data at: *https://zenodo.org/badge/latestdoi/461912754*

Related Work

Toxic language is highly contextual and its interpretation is subjective, which could reduce consistency and reliability (Arhin et al. 2021). Group language features (e.g., Black English) may be more likely to be labelled as abusive (Kim et al. 2020). Assigned labels may vary between annotators with different demographics and beliefs (Kocoń et al. 2021), and especially if they have been a target of hate (Olteanu, Talamadupula, and Varshney 2017; Kumar et al. 2021).

These example factors indicate that the identification of the targeted groups may not be correctly reported. Content containing mentions to targeted groups has a higher level of subjectivity, as it expresses more personal opinions and less factual information (Zhao, Zhang, and Hopfgartner 2021), which can complicate its correct detection by human annotators. Unlike previous work, we here examine the use of a domain ontology about gender and sexual orientation to support the annotation of such targets in toxic speech corpora.

Data and Methods

Our aim is to understand the problems of toxicity annotations by using domain knowledge about two specific groups to improve the identification of these groups in toxic speech. We analyse a subset of a large-scale toxicity corpus (`Jigsaw Toxicity 448k`) because it includes *identity labels*, i.e. tags representing specific demographic attributes that are assigned to each text if it refers to an individual or group with those characteristics (Borkan et al. 2019).

In this section, we present the use of an ontology to address our research question. First, we analyse toxicity samples from different demographic groups separately to compare to what extent the semantic concepts of the `GSSO` ontology can identify texts with mentions of G and SO groups with respect to other toxicity targets. We then compare the use of these semantic concepts in the different group samples to find whether representative concepts exist in each group. Finally, we used this information to identify toxicity texts related to G and SO that were missed by human annotators.

Data Preparation

The `Jigsaw Toxicity 448k` dataset contains *identity labels* corresponding to the following demographic groups:

- Gender (G): female, male, transgender, other gender,
- Sexual Orientation (SO): homosexual or gay or lesbian, bisexual, heterosexual, other sexual orientation,
- Religion (Re): christian, jewish, muslim, hindu, buddhist, atheist, other religion,
- Race (Ra): black, white, asian, latino, other race or ethnicity,
- Disability (D): physical, intellectual or learning, psychiatric or mental illness, other disability,
- None: if the text did not mention any demographic attribute.

We assign a text to a group if at least 50% or more of the annotators agreed to give at least one of the *identity labels* of that group. As shown in Figure 1, there is a significant size disparity between the different groups (e.g. sample G is approximately 7 times larger than sample SO). A sample representing all groups equally avoids having biased results from one group, i.e. identifying more mentions from one group than from the others due to the larger sample size. As disability (D) is the smallest group, we randomly draw a sample of the same size from the other groups, allowing us to analyse balanced group samples (i.e. samples of 5559 texts).

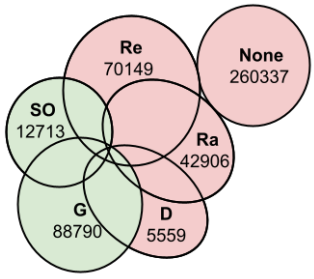


Figure 1: Demographic groups and their sample size in `Jigsaw Toxicity 448k` dataset: Gender (G), Sexual Orientation (SO), Religion (Re), Race (Ra), Disability (D), and not related to any (None).

Identifying Targeted Groups in Toxicity Texts

We search each group sample of the dataset for the appearances of concepts related to G and SO drawn from the `GSSO` ontology. If a concept is frequently found in one group sample but not in the others, we assume that it is representative of that specific group. We expect to find representative concepts particularly in the G and SO group samples, since we are using `GSSO` domain knowledge. Then, if a representative concept is found, we can assume that the given text is related to the G and SO identity groups.

The `GSSO` ontology has 14280 entities in a hierarchy system, extracted from encyclopedias, dictionaries, subject headings, and classification systems related to gender, sex, and sexual orientation (Kronk and Dexheimer 2020). It includes over 200 slang terms, 190 pronouns, and over 200 nonbinary and culturally specific gender identities.

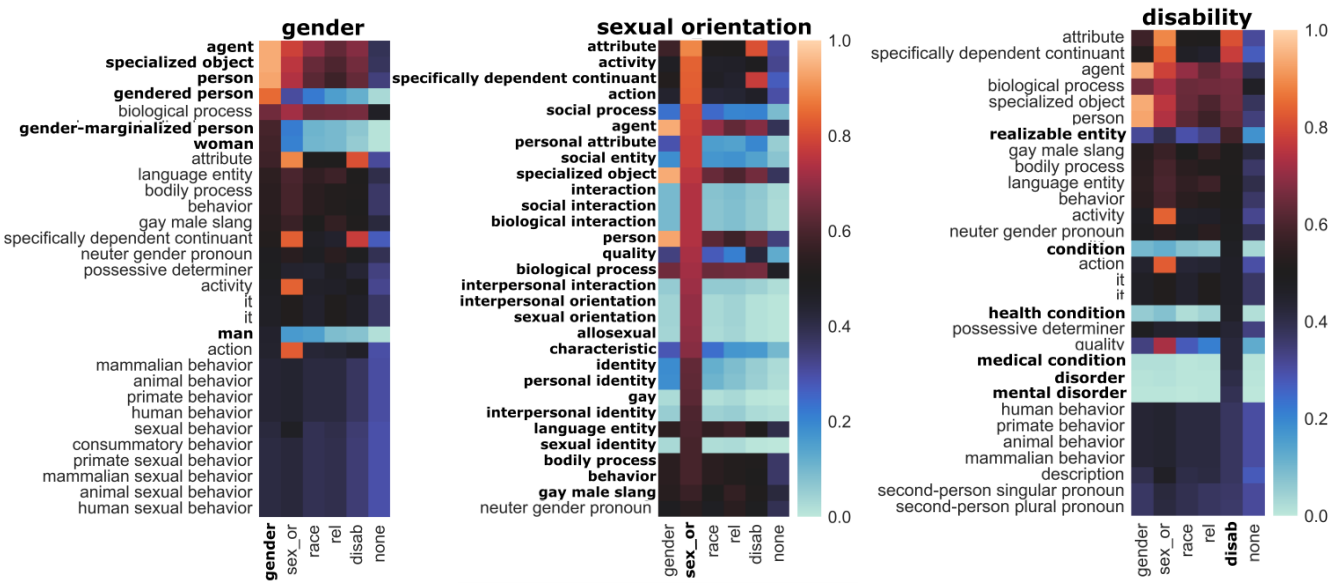
This domain knowledge is used to find G and SO representative concepts. First, we find *asserted* entities in the text by mapping entity labels or other properties (i.e. *alternate* or *short name*, *synonym*, *exact*, *broad*, *narrow* or *related synonym*, or *replaces*). Then, we use the ontology structure to reason over the semantic meaning of the entities directly mentioned in the text and obtain a list of *inferred* entities, such that for each class, we know all its superclasses, and for each individual, its types.

Once each text has its corresponding list of entities, we can find the most informative of each group by calculating their document frequency, such that for each entity, we divide its occurrence in the group sample by the total number of texts. By comparing the frequency distributions in other group samples, we can determine whether specific concepts are representative of a given group (i.e. if entities corresponding to those concepts are more frequent in that group than the others), or they are only the most common ones in that group sample. We show representative concepts (Figure 2) excluding the entities that appear in more than 50% of the texts in all groups (i.e. *common entities*): *object, entity, continuant, occurrent, independent continuant, process, information content entity, subject pronoun, object pronoun, third-person singular pronoun*.

Identifying Missing Target Annotations

The representative concepts of the G and SO group samples can be used to identify texts with mentions of these groups that annotators did not report (i.e. in the red area in Figure 1). This area corresponds to 18948 texts in our balanced group sample.

To do so, we compute a $metric$ that sums the average frequency of each entity in the interest group samples (i.e., of



(a) *Interest group samples (S^+): gender and sexual orientation*

(b) *Other group samples (S^-): disability*

Figure 2: Frequency analysis to identify targeted groups in toxicity texts based on representative GSSO concepts (in bold). We only show group samples which had representative concepts (i.e. with higher frequency than in the texts of other groups).

G and SO), minus the average frequency in the other groups, such that:

$$metric = \frac{\sum_{i=0}^{\breve{M}} \frac{\sum_{j=0}^{S^+} freq(i)_j}{S^+} - \frac{\sum_{j=0}^{S^-} freq(i)_j}{S^-}}{\breve{M}} \quad (1)$$

$freq(i)$ is the frequency of entity i in the j group sample. $S^+ \in G, SO$ are interest group samples, and $S^- \in Re, Ra, D, None$ the others. We divide by number of entities found in the text ($\breve{M}$) to normalize by text length.

Results

In this section, we present and explore our analysis findings along these two lines:

- The ontology is able to capture concepts that are especially representative in toxicity texts against SO groups, and to some extent G and D.
- These representative concepts can be used to identify toxicity texts with mentions of G and SO groups that human annotators did not correctly identify.

Figure 2 shows how terminology related to G and SO distributes in different group samples. Our first observation is that toxicity texts against SO groups have the highest number of representative concepts: *social process, personal attribute, social interaction, sexual orientation, allosexual* and *gay* appear in over 80% of SO texts, whereas these concepts never occur more than 20% in the other groups. This fact indicates the specificity of these concepts to the SO toxicity target.

The G and D group samples are the only other two groups with representative concepts. Entities most representative of G are *gendered person, gender-marginalised person*, and *woman*, since they occur in over 60% of G related texts compared to under 20% in the other groups. *Man* is also distinctive to G, but with a lower frequency (i.e. around 50%) that is similar to the representative entities in D texts: *condition, health condition, medical condition, disorder*, and *mental disorder*. The remaining groups do not have representative entities, as the most frequent entities have a higher frequency in the other group samples.

This background knowledge can be used to identify missing target information in the texts that according to human annotators did not contain any mention to G or SO. We use the $metric$ in Equation (1) to find such texts:

a) Using all the *asserted* and *inferred* entities in the text (*Score*).

b) Using only the *asserted* and *inferred* entities that represent the concepts of each group (*Score categories*), i.e. gender, sexual orientation, and so on.

Table 1 shows texts with a positive *Score* value. We find that only the text with *ID 6203311* was identified by 4 out of 10 annotators as related to SO. The other two examples were missed by the annotators, who did not relate *pansy* (a homophobic slur), *hon* (a transphobic slur), and *snowflake* (a derogatory term) to these groups.

However, we found limitations in this measure as it cannot identify intersections with other groups (i.e. mentioning concepts representative of other groups) and only identifies texts with few entities found or only dealing with G and SO topics. One example of these limitations is the text with *ID 7159592* in Table 1, as *serve* is the only entity found and happens to be slightly more frequent in the G and SO group samples.

To deal with these limitations, we give priority to the en-

ID	*Identifying missing target annotations using all entities*	*Score*	*Annotators*
5778476	*The only **snowflakes** I see are the insecure white supremacists and the Nazi **pansies**.*	0.0061 pansy, snowflake	**0/4**
6203311	*Me too Neko. Good Lord the **gay** brings out the worst in some 'TradCaths' understanding of Christianity. And yes, I am definitely making a distinction between Christians and self proclaimed traditional catholics.*	0.0055 gay, gay identity	**4/10** homosexual **1/10** other sexuality
7159592	*The Congressional Liberal Black Caucus is unfit to **serve**.*	0.0005 serve	**0/6**
7024852	*Most liberals are white **hon**.*	0.0002 hon	**0/4**

Table 1: Toxicity texts filtered using *all* entities in Equation (1), shown in the *Score* column. *Annotators* is the number of annotators who agreed to give each text a particular gender or sexual orientation label.

tities that fall under the concepts representing the targeted groups (*Score categories*) and identify 571 texts with terms related to G and SO. These are longer texts dealing with topics related to other groups. For example, the entities *gender, gestation* (pregnancy), *parturition* (birth), *control, parenthood, rape, abortion*, and *selection* are found in the following text, which deals with health and gender issues:

*"**Abortion** and Planned **Parenthood** is quite the misnomer as **abortion** should never be used as **birth control** or for **gender selection**, however, I do agree that **abortion** for **pregnancy** due to **rape** is necessary though."*

When we include all entities in the metric, it gives a value of -0.08. Texts tend to have more negative scores the more the number of representative concepts in other group samples increases (e.g. with disability-related entities):

*"The problem with your thesis is that you are treating this "**disorder**" like any other **mental disorder** where the cure is to rid yourself of the **symptoms**, **feelings**, urges, etc. Whereas the only reason the DSM and WHO list **gender dysphoria** as an **illness** is that it causes distress and **dysfunction**; however, the recommended cure by both **organizations** is to acknowledge the disconnect and support the **person** in transitioning. You, on the other hand, are also treating the remedy as a **delusion**. The DSM and WHO are not."*

Even though the following text contained fewer mentions of G and SO, it obtains a similar *Score* value (-0.11 and -0.12, respectively), due to additional entities being non-group related topics (e.g. media, family, elitism):

*"The progressive values are the ones the **elitist** left are slowly making norm, LGBTQ, **Gender neutrality**, break down of **families** and Government **control** of kids, open borders and the influx of Islam into every **nation**. One World Government is the international progressives goal. If you can't see this you are totally blinded by the lies of the **mainstream media**."*

In conclusion, this analysis allows us to observe 575 texts with mentions of G and SO which were not identified by human annotators, which correspond to the 3% of texts in the $\sim 19k$ sample. This variability could compromise the reliability of annotations, impacting the performance of the detection models (Pandey, Castillo, and Purohit 2019).

The number of inaccurate annotations could be even higher, as we have only presented an initial approach to exploit the GSSO domain knowledge. *Score* cannot prioritise the most relevant entities for G and SO, whereas *Score categories* is limited only to entities under the branches of the G and SO concepts. Thus, investigating the existence of additional representative entities could find other missing target annotations. The interplay between using these semantic concepts with current language models as constraints to provide a data representation more consistent with prior knowledge about these groups remains another critical open question for our future work.

Discussion and Conclusion

Focusing on gender and sexual orientation as a case study of toxic speech annotations, we show that the use of a domain ontology can support the identification of text with mentions of these groups that human annotations did not correctly identify. Due to the complexity of providing manual annotations of such corpora, giving background knowledge about commonly targeted groups can help reduce the variability of such annotations.

While this study is limited to toxic speech in a single platform, language and time, we identify limitations of annotation protocols that are widely used in the detection of toxicity and other related phenomena (e.g. hate speech, abusive and offensive language) (Fortuna, Soler-Company, and Wanner 2021). The use of ontologies to validate texts that have been checked and verified by human annotators or assist them during the annotation process is promising directions for building more consistent and reliable benchmarks. The use of knowledge of specific demographic groups can also help enrich and pre-process datasets where their information remains unidentified to avoid the risks associated with using "one-size-fits-all" models.

This work is only a small contribution to the current discourse about the use of semantic knowledge to create more unified and reliable training datasets, so we can better monitor the performance and prevent the discriminatory impact of AI systems on the groups they make decisions.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement no. 860630) for the project "NoBIAS-Artificial Intelligence without Bias".

References

Arhin, K.; Baldini, I.; Wei, D.; Natesan Ramamurthy, K.; and Singh, M. 2021. Ground-Truth, Whose Truth?–Examining the Challenges with Annotating Toxic Text Datasets. *arXiv e-prints*, arXiv–2112.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of the 2019 World Wide Web Conference*, 491–500.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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

Fortuna, P.; Soler-Company, J.; and Wanner, L. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3): 102524.

Hutchinson, B.; Prabhakaran, V.; Denton, E.; Webster, K.; Zhong, Y.; and Denuyl, S. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491–5501.

Kim, J. Y.; Ortiz, C.; Nam, S.; Santiago, S.; and Datta, V. 2020. Intersectional Bias in Hate Speech and Abusive Language Datasets. *arXiv e-prints*, arXiv–2005.

Kocoń, J.; Figas, A.; Gruza, M.; Puchalska, D.; Kajdanowicz, T.; and Kazienko, P. 2021. Offensive, Aggressive, and Hate Speech Analysis: From Data-centric to Human-centered Approach. *Information Processing & Management*, 58(5): 102643.

Kronk, C. A.; and Dexheimer, J. W. 2020. Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and workflow. *Journal of the American Medical Informatics Association*, 27(7): 1110–1115.

Kumar, D.; Kelley, P. G.; Consolvo, S.; Mason, J.; Bursztein, E.; Durumeric, Z.; Thomas, K.; and Bailey, M. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. In *Seventeenth Symposium on Usable Privacy and Security*, 299–318.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 14867–14875.

Olteanu, A.; Talamadupula, K.; and Varshney, K. R. 2017. The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. In *Proceedings of the 2017 ACM on Web Science Conference*, 405–406.

Pandey, R.; Castillo, C.; and Purohit, H. 2019. Modeling Human Annotation Errors to Design Bias-aware Systems for Social Stream Processing. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 374–377.

Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; and Patti, V. 2021. Resources and Benchmark Corpora for Hate Speech Detection: a Systematic Review. *Language Resources and Evaluation*, 55(2): 477–523.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019a. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2019b. Social Bias Frames: Reasoning about Social and Power Implications of Language. *Association for Computational Linguistics*.

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the Targets of Hate in Online Social Media. In *Tenth international AAAI conference on web and social media*.

Zhao, Z.; Zhang, Z.; and Hopfgartner, F. 2021. SS-BERT: Mitigating Identity Terms Bias in Toxic Comment Classification by Utilising the Notion of" Subjectivity" and" Identity Terms". *arXiv preprint arXiv:2109.02691*.