Annotating Targets of Toxic Language at the Span Level

Baran Barbarestani and **Isa Maks** and **Piek Vossen**
CLTL Lab, Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
{b.barbarestani, isa.maks, piek.vossen}@vu.nl

Abstract

In this paper, we discuss an interpretable framework to integrate toxic language annotations. Most data sets address only one aspect of the complex relationship in toxic communication and are inconsistent with each other. Enriching annotations with more details and information is, however, of great importance in order to develop high-performing and comprehensive explainable language models. Such systems should recognize and interpret both expressions that are toxic as well as expressions that make reference to specific targets to combat toxic language. We, therefore, create a crowd-annotation task to mark the spans of words that refer to target communities as an extension of the HateXplain data set. We present a quantitative and qualitative analysis of the annotations. We also fine-tune RoBERTa-base on our data and experiment with different data thresholds to measure their effect on the classification. The F1-score of our best model on the test set is 79%. The annotations are freely available and can be combined with the existing HateXplain annotations to build richer and more complete models.

1 Introduction

Communication through social media has exploded in the last decades. The ease of posting opinions and the relative anonymity of posters has also unleashed problematic communication that can take many different forms: offensive language, hate speech, discriminatory language, abusive language, cyberbullying, etc., which can be all captured under the umbrella term *toxic*. Such communication is often very complex and involves different values and perspectives. A comprehensive interpretation of such communication requires different aspects to be detected and combined, among which expressions that make a judgement or suggest negative implications and expressions that refer to targets such as a specific group of people or an individual belonging to such a group. An explainable system that can act as an automated moderator should be capable of "understanding" such phrases, reason over their content and bring specific aspects to posters' attention to explain what is wrong with a post and why it has to be, for example, removed by moderators (Kiritchenko and Nejadgholi, 2020). An explainable model not only produces the desired outputs, but also explains why such output are produced.

The Natural Language Processing community has started many initiatives to automatically detect and classify toxic language and created a plethora of datasets (Vidgen and Derczynski, 2020; Poletto et al., 2021). However, these data sets often address only one of the above-mentioned aspects. Furthermore, they use slightly different terminologies and definitions for annotation and their annotation guidelines lack compatibility, which makes it difficult to combine their annotations. Another problem is that annotation is often done at a global level, such as the whole sentence instead of specific phrases and tokens. Some recent initiatives have started to annotate specific spans within the text itself (Mathew et al., 2020; Pavlopoulos et al., 2021) but this is limited to toxic spans only.

Although previous studies did annotate the target community (e.g., women, Muslims, immigrants, etc.) at the post level, none of these studies marked the words that describe or refer to such a group. Being able to detect these phrases is, however, crucial to reason over who is targeted and how they are referenced. Furthermore, annotating references to targets separately from toxic spans makes it possible to also process larger contexts of communication, among which conversations where references to targets and toxic expressions may be dispersed over multiple posts. By building a framework where target spans are annotated, it is possible to train explainable models that not only tell which community groups are targeted in a piece of text, but

Proceedings of the 29th International Conference on Computational Linguistic, pages 43–51
October 12–17, 2022.

also indicate which words and phrases this decision is based on. This will make the model and its decision more understandable for end users.

In this study, we describe a crowd-annotation task to annotate such target spans. We used this framework to add target spans to the HateXplain data set (Mathew et al., 2020) and tested classification models by fine-tuning with different quality selections of data.

Our contributions are as follows:

- An explainable framework to combine different toxic data annotations has been discussed.
- A crowd-annotation task, which aims at the identification of target spans, i.e., sequences of targeting tokens that together refer to a target community, has been created.
- An existing data set has been extended (Mathew et al., 2020) with annotations of target spans.
- A preliminary quantitative and qualitative analysis of the annotations has been provided.
- Three RoBERTa-base models have been fine-tuned on our data and the results have been reported.

The paper is organized as follows. In Section 2, we summarize the related work on target annotations and position our work. In Section 3, we describe the resources that we use to sample data in order to obtain sufficiently diverse annotations from each source. Section 4 explains our annotation framework and the crowd-annotation task that we designed, while in Section 5 we describe the results of the annotation. In section 6, we report on the language models we fine-tuned with our data and explain the results. We conclude and discuss future work in Section 7.

Please beware that this paper may contain some examples of hateful content. This is strictly for the purpose of enabling this research and we seek to minimize the number of examples wherever possible. Please be aware that this content may be offensive and cause you distress, which is certainly not the intention of the authors of this article.

2 Related work

The number of studies on the automatic identification of hate speech and other forms of toxic language has rapidly increased in recent years. Several definitions for toxic language have been proposed and many different annotation schemes have been designed and applied.

Part of these annotation studies focuses on the target community that has been victimized by such language and acknowledges that the description of these targets is relevant in different ways for the automatic detection of toxic language. Early studies (De Gibert et al., 2018), (Davidson et al., 2017) presented this task as a binary task labeling data as hateful or not. In these studies, only toxic expressions targeting people were considered hateful. For example, according to the annotation guidelines of (De Gibert et al., 2018), an expression should be labeled as hate speech only if all of the three following conditions are met: (1) There is a deliberate attack. (2) The attack is on a specific group of people. (3) The motive for this attack regards aspects of the group's identity. Although the identity of the target group is decisive in determining whether an expression is considered hate speech or not, no details on this were annotated.

Another widely used annotation scheme (see e.g. (Basile et al., 2019), (Zampieri et al., 2020)) was developed by (Zampieri et al., 2019) who addressed the need for identifying more specific information about the target communities and therefore introduced several annotation layers as follows: (1) Determine whether the message is offensive. (2) If the message is offensive, determine whether it is targeting people or not. (3) If the message is targeting people, determine (a) whether the message is targeting an individual, or (b) whether the message is targeting a group or member of a group considered a unity due to the same ethnicity, gender or sexual orientation or any other common characteristic, or (c) whether the message is targeting other entities like an organization, a situation or an event.

Finally and most recently, several studies ((Mollas et al., 2022),(Kennedy et al., 2020), (Vidgen et al., 2021), (Ousidhoum et al., 2019)) have taken the target annotations one step further by providing the group aspect on which basis it was targeted (e.g. gender, race, national origin, disability, religion, sexual orientation, etc.) and by mentioning the specific target communities (e.g. Africans, immigrants, Muslims, homosexuals, politicians, etc.) This information would allow further research into differences in the framing of specific target communities and the building of classifiers that avoid bias in hate speech detection ((Shah et al., 2021))

or permit researchers to delve into issues related to such bias. Although all of these studies considered target detection in hate speech as challenging and important, none of them annotated target spans at the token level.

Our work builds on the already existing annotations of the HateXplain data set (Mathew et al., 2020) by adding such span annotations that refer to a target/ target community. In combination with the annotations already present in HateXplain, this allows us to train systems to detect both the phrases making reference to the targets as well as inferring the group aspect of these targets together with pointing at the phrases that represent the insulting content or judgement expressed about them.

3 Source data overview and sampling

HateXplain is the first hate speech data set that covers many aspects of toxic language (Mathew et al., 2020). Each post in this data set has been annotated from three different perspectives: 1) the three main classes: hate speech, offensive or normal 2) the target community (i.e., the community that has been the victim of hate speech/ offensive language in the post) 3) rationales that are the parts of a post based on which annotators have decided to label it as such. The annotations were carried out at the word and phrase levels except for the target information which was done at the utterance level. According to (Mathew et al., 2020), the data was collected from Twitter and Gab with a total of 9,055 and 11,093 samples, respectively.[1]

For this study, we added target spans to the already existing annotations in this data set. This means that, for each sample targeting a target community, we wanted to determine which tokens in that sentence referred to that target community. For this reason, we selected only those samples that a) were instances of offensive language or hate speech b) targeted only one target community c) at least 2 out of 3 annotators agreed on its target label and d) had more than two and fewer than 61 tokens. We had extracted the distribution graphs of sentences per number of words and noticed that there are very few sentences that had more than 60 words in our data set. Also, the more words a sentence has, the more complex it becomes. In addition sentences with fewer than 3 words seem to have not enough and useful information for analysis. That is why we selected only sentences whose number of words was within the range described. The reason why we chose sentences with only one target group was mainly to be make the task as easy and simple as possible for the crowd. Nonetheless, we later found that there were still a number of sentences that targeted more than one target group even though they were annotated in HateXplain as having only one target group. This is also referred to and explained in 5.4. As a result, a total of 6,445 samples were selected. From these, we selected 3,480 samples that were representative of different target communities and data sources, which constituted about 54% of the full sample set. The number of selected samples per target community are shown in Table 1. Only those target communities that appeared at least 10 times are shown in the table. The rest of them together with the "Other" target community are referred to as "Other target communities".

Target community	Gab	Twitter	Other	Total
Hispanic	160	2	0	162
Men	5	7	0	12
Homosexual	457	331	2	790
Islam	145	111	0	256
Arab	64	39	0	103
Refugee	105	83	1	189
Asian	82	68	0	150
Disability	3	29	0	32
African	355	278	0	633
Jewish	75	47	0	122
Caucasian	47	94	0	141
Women	225	385	0	610
Other target communities	150	130	0	280
Total	1873	1604	3	3480

Table 1: Data selected for annotation across target communities and sources (Gab, Twitter, Other)

To prepare the data set for the crowd-annotation task described in the next section, we sorted the data per target community and created separate batches of 30 samples. Each batch contained 29 samples from HateXplain and 1 attention check sentence randomly selected from a list of 60 sentences. These attention check sentences were manually generated and annotated, and were both targeting and toxic. Attention check sentences were included to inspect the crowd-annotators' annotation and make sure that they did not simply annotate the data samples haphazardly or even skip them. The total number of batches amounted to 120. The first 25 batches only contained "Hispanic" and Homosexual" target communities. From batch 26 onward, however, each batch contained a mix-

[1]However, we observed that only 9,027 samples were labelled with the source Twitter, resulting in 28 samples that were not identified.

ture from different target communities and sources (Gab and Twitter) to reduce imbalanced data bias.

4 Crowd annotation task

For the crowd-annotation task, annotators were provided with batches of 30 toxic sentences presented one-by-one, where each sentence was followed by a list of tokens with tick-boxes. They were asked to read each sentence carefully before determining who the target was and select the tokens that referred to that target. At the start, they received an instruction page with examples. They were instructed to choose all the words - including the pronouns - that referred to the target no matter how many times they appeared in the sentence. Also, complete noun phrases including any adjectives had to be selected. Below are a few examples that demonstrate how sentences should be annotated:

1. You Bangladeshi guys are f***ing gangsta.
   Target spans: You, Bangladeshi, guys

2. Just because we don't let groups of creepy Moroccan rapists into the club doesn't make us racist.
   Target spans: creepy, Moroccan, rapists

3. They do no shit other than going to the mosque.
   Target spans: They

4. Fuck the ignorant assholes from America, UK, and all other countries who try and soil the wonderful ties between all democratic nations.
   Target spans: the, ignorant, assholes, from, America, UK, and, all, other, countries

5. SHUT UP WHITE BOY!
   Target spans: WHITE, BOY

6. What's up nigger?
   Target spans: nigger

7. You spics are worthless.
   Target spans: You, spics

8. He is a stupid Muslim.
   Target spans: He, a, stupid, Muslim

9. I need a new gay hashtag.
   Target spans: gay

It is worth noting that there are certain words and phrases that are toxic and specifically used for referring to a target community, such as "spic" in 7, which generally refers to Hispanics, or "nigger" in 6, which refers to Black people. These words should be also annotated as target words although they are toxic as well, because they are ultimately used to address specific target communities.

We used LINGOTurk (Pusse et al., 2016) as the platform to design our task, which is a free and available crowdsourcing client/ server experiment management system. Furthermore, we used Prolific (Palan and Schitter, 2018), as the online platform for recruiting crowdworkers. The advantage of Prolific over other crowdsourcing platforms, such as Amazon Mechanical Turk, is that it provides more prescreening options, such as nationality, first language, ethnicity, political affiliation, socio-economic status, etc. Also, Prolific is not only limited to US participants. When recruiting annotators, we clearly described the aim of the study to them and explained what they had to do in detail. No specific sensitive information about annotators was stored. We also informed them beforehand that they should be aware of the inappropriate content of the sentences and they were not supposed to participate in this study if they were not comfortable with being exposed to such a language. Since the study was closely related to one's cultural understanding of the context and there were a lot of slang words and phrases used, we recruited only participants that met the following criteria:

- Both their nationality and country of birth had to be at least one of the following: United Kingdom (England, Wales, Scotland, Northern Ireland), United States, Ireland, Australia, Canada, Guyana, Jamaica, Liberia, New Zealand

- Their first, fluent, and primary language had to be English.

In order to determine the optimal number of annotators to recruit for each batch, we ran a test batch with 15 annotators and then extracted 10 random subsets, once with 5 and once with 10 annotators. mathtools

Following the CrowdTruth framework (Dumitrache et al., 2018), we used the Media Unit Quality Score (UQS) to analyze the collected results from different sets of annotators. UQS expresses the overall worker agreement over a so-called media unit. In our case, each token was considered to be a media unit with the binary classification as

either targeting (1) or non-targeting (0). In order to calculate the UQS, one needs to first calculate the average cosine similarity between all worker vectors, weighted by the worker quality *(WQS)* and annotation quality *(AQS)*. For more details on how each of these scores is calculated, please refer to (Dumitrache et al., 2018). The advantage of using UQS in comparison to other metrics for calculating the inter-annotator agreement is that CrowdTruth interprets both the disagreement among the annotators and the ambiguity of the token. The quality of an annotation is considered as the interaction between the quality of the annotator in terms of how often she/ he agrees with others as well as the complexity of the input data and set of annotation categories.

We calculated the UQS for the complete set with 15 annotators and each of its subsets including 10 and 5 annotators, respectively. Next, we took the average of the obtained results over all subsets. By doing so, we could test in which cases and with what number of annotators the results were more consistent. In Table 2, the average overall UQS, average UQS for targeting tokens, and average UQS for non-targeting tokens across the test batch with different numbers of annotators are given. Targeting tokens refer to the tokens the majority of annotators labeled as targeting while non-targeting tokens refer to the tokens labeled as non-targeting by the majority. Also, the standard deviations of the 3 metrics per subset are given. In the case where there were 15 annotators, the average was taken over the media units and not different subsets, since no subset was created in this case. The closer the UQS and standard deviation are to 1 and 0, respectively, the higher the quality is.

Number of annotators	15	10	5
Avg UQS	0.81	0.80	0.80
Avg UQS for targeting tokens	0.78	0.80	0.80
Avg UQS for non-targeting tokens	0.86	0.86	0.87
SD of Avg UQSs	0.17	0.18	0.23
SD of Avg UQSs for targeting tokens	0.16	0.14	0.16
SD of Avg UQSs for non-targeting tokens	0.12	0.12	0.14

Table 2: Comparison of the annotation quality with different numbers of annotators. Avg=average; SD = standard deviation;

As can be seen in Table 2, the differences between the values are quite marginal and, especially for 10 and 15 annotators, most values are the same. Therefore, we decided to recruit 10 annotators per batch to do the annotations.

To select 10 annotators within the Prolific platform, the above pre-screening criteria were applied to the total pool of annotators. After running each batch, we analyzed the data to make sure the annotation quality was good enough and annotators acted according to our instructions. In order to do so, we compared the performance of each annotator to that of other participants, validated the attention check sentences, and considered the time taken on the whole for each annotator to finish the task. We also validated the annotations of some other randomly selected sentences. Finally, we checked whether the data provided by each participant corresponded with their Prolific ID and if they had entered a completion code showing that they had completed the whole task. If annotators failed any of the above-mentioned criteria, their submissions were rejected and another annotator was recruited in their place. We added the IDs of rejected annotators to our blocklist after each batch, which would exclude them from the next batches. In the next section, the results will be described in more detail.

5 Annotated Data

5.1 Statistical analysis of the crowd labels

We ran the batches for several weeks on the Prolific platform to obtain 10 annotations per sentence, eliminating problematic annotators as explained above. Table 3 gives a numerical overview of the result of the crowd annotation. In total, 5,799 target spans were identified, of which 4,747 (82%) were single-token. Interestingly, Gab samples had more references to target communities (the average number of target spans per sample was 1,82) than tweets did (with 1,48 spans on average). Additionally, the target spans found in Gab were a bit longer (with 1,52 tokens per span on average) than those found in tweets (1,44 tokens on average). These numbers can be explained by the fact that the Gab samples were generally longer than tweets, having 24,8 tokens on average, whereas this number was 14,6 for Tweets. However, it also shows that the two data sources had different characteristics with respect to how they referred to target communities.

5.2 Gold data annotated by experts

To get an independent evaluation of the quality of the crowd annotation, we did an expert annotation on two batches (2 and 23) through the same

	Gab	Twitter	Total
nr of samples	1873	1604	3480
avg nr of tokens per sample	24,8	14,6	20,3
nr of target spans	3417	2378	5799
avg nr of spans per sample	1,82	1,48	1,66
avg tokens per span	1,52	1,44	1,5

Table 3: Annotation statistics

platform. The annotators were the authors of this paper (A1, A2, A3). We calculated the Cohen's kappa coefficient per pair of annotators. The results can be seen in Table 4.

		A1-A2	A2-A3	A1-A3
Batch 2	Percent agreement	0.90	0.90	0.91
	Kappa score	0.67	0.65	0.67
Batch 23	Percent agreement	0.87	0.89	0.89
	Kappa score	0.62	0.66	0.69

Table 4: Inter-annotator agreements among expert annotators (A1, A2, and A3) on the batches 2 and 23

The results show a reasonable agreement with kappa scores ranging from 0.62 to 0.69 across different annotators, different batches, and different classes (targeting vs. non-targeting). The percent agreement scores are above 87%. We discussed each case of disagreement and resolved these using predefined guidelines in order to have a fully adjudicated gold data set with expert annotations. The adjudicated annotations were used to determine the optimal settings for selecting the final label assigned to each token (see section 5.3).

5.3 Aggregating the crowd annotations

The expert annotation was used to determine the best threshold for selecting the labels of the tokens annotated by 10 annotators. We used the already explained UQS score per token as a basic measure to assign the final label and compared the results with the expert annotations for batches 2 and 23. We repeated this 3 times with different UQS thresholds: 50, 60, and 70. Tokens below these thresholds were considered as non-targeting since the crowd could not reach sufficient consensus. For the tokens with a UQS above these thresholds, we selected the label with the highest majority score, which could be either targeting or non-targeting. Table 5 shows the Kappa scores by comparing the crowd data with different UQS thresholds to the expert annotations.

We can see that the scores are lower than those for the expert annotations, as could be expected, but still reasonably high. Interestingly, we got the highest agreement using the lowest UQS threshold, which was 50. Therefore, we selected this threshold for creating a test set from the crowd annotations done on all batches.

Batch	UQS 50	UQS 60	UQS 70
2	**0.55**	0.48	0.37
23	**0.58**	0.56	0.45

Table 5: Kappa scores comparing the crowd data to the expert data with different UQS thresholds

5.4 Patterns of agreement and disagreement in the crowd annotations

We performed an error analysis in order to find which disagreement issues occurred most frequently and whether we could find any patterns. We illustrate the issues by making references to the examples below. The tokens marked with [T] are those that were annotated by a vast majority of annotators as targeting, tokens marked with [T*] were annotated by only a small number of annotators as targeting, which caused disagreement. The underlined tokens are those that were considered to be targeting by the authors of this paper.

1. Go back to France[T*] you[T*] anti-american[T] commie[T] socialist[T] traitor[T]
2. One must wonder how long she[T*] will do in jail[T*] on the taxpayers dime before they decide to toss her[T*] spic[T] ass[T] out; as for the guy[T*] castrate[T*] him[T*] and jail[T*] him[T*] for life
3. I[T*] am black[T*] with all white[T*] friends[T*], taco[T] man[T*]
4. the wall is not the final solution these[T*] beaners[T] will keep having babies[T*] keep balkanizing[T*] our society(T*)
5. i saw a[T*] couple[T*] of[T*] baums[T*] and[T*] steins[T*]
6. as if it only okay when you are a cute gay[T] boy

The main sources of disagreement can be summarised as follow:

Length of target span Not all annotators annotated the complete target span. Typically the beginning (cf. *these* in ex. 4 and *her* in ex. 2) or the end (cf. *man* in ex. 3) are missing.

Additional information about the target community It seems that some annotators annotated properties, descriptions and behaviours of the target communities, whereas these tokens are not references to the community, but describing them (cf. *France* in ex. 1; *jail* in ex. 3; *babies, balkanizing, society* in ex. 4)

Inconsistent identification of referring pronouns Pronouns that refer to the target community were often missed (cf. *you* in ex. 1; *she, her, him, him* in ex. 2). This pattern is further confirmed by the words listed in Table 6: the references with the highest agreement were ethnic slurs (right column), whereas the references with the lowest agreement were pronouns (left column).

Multiple candidate target communities Apparently, there was confusion among annotators when multiple communities were referred to. In ex.3, *black* and *white friends* were both incorrectly annotated as targeting, whereas no target community was targeted in this particular sample.

Different interpretations In many cases, annotators did not agree about whether a reference to a target community was toxic or not. For example, those who considered the expression 'Baums and Steins' (cf. ex. 5) to be ironic rather than offensive, did not label it as a targeting expression. xxxxxcbIn these cases it is not much possible to give the correct answers as these considerably depend both on the context and the annotator's individual perspective (cf. (Basile et al., 2021)).

No explicit target word There were cases where no target community was explicitly targeted, but because of the assumption that all sentences must be targeting (as explained in the instructions), annotators selected the existing community referred to in the sentence despite the absence of any obvious toxic reference to it (cf. ex. 6).

The analysis showed that toxic references to the target communities (such as *beaners, her spic ass*) were more easily identified than neutral ones such as *man* and the pronominal references. Moreover, it showed that annotations with a relatively low agreement required further analysis.

6 Automatic classification

After having obtained the labels for each token and having determined the best UQS threshold, we tested how well a language model could learn to detect the target spans and which UQS threshold for the training data would work best. Setting a high UQS threshold would give fewer data with a higher consensus, whereas a UQS threshold of 50, which had resulted in the highest Kappa score when the crowd annotations were compared with the expert annotations, would give us more targeting samples in the training data.

Low UQS	High UQS
You (206)	Nigger (357)
They (90)	Faggot (265)
The (77)	Bitch (211)

Table 6: Most frequent words targeting tokens: high vs. low agreement

To test this, we fine-tuned a pretrained language model for a token classification task to predict whether each token was targeting or non-targeting. In (Sharma et al., 2021), the performances of a number of language models for detecting toxic spans in a sentence were compared. The best-performing model (RoBERTa-base) had the highest F1-score on the test set with a value of 68.41%. Therefore, we chose RoBERTa-base as our pretrained model. For fine-tuning, we converted the data to the IOB (Inside-Outside-Beginning) format, which is widely used in token classification tasks (Evang et al., 2013).

We created a separate test set consisting of 20% of the whole data, but ensured that it was representative of all target communities and data sources. The test set was generated by setting the UQS threshold to 50, as this threshold had previously resulted in the highest agreement when the crowd annotations were compared with the expert annotations. For the training, on the other hand, we generated three different training sets with UQS thresholds of 50, 60, and 70, to test the effects on the predictions. All other hyperparameters and arguments remained the same in all three cases. Furthermore, we selected 10% of the training data as the validation set. The training set, test set, and validation set included each 2227, 696, and 557 samples (sentences), respectively. Arguments and hyperparameters used for the training are as follows: batch size=16; epochs=3; learning rate=2e-5; weight decay=0.01. To prepare the data for fine-tuning our models, they were tokenized using AutoTokenizer from Hugging Face[2].

[2] https://huggingface.co

During fine-tuning, evaluation was done at the end of each epoch. We batched our data with a data collator while using padding to make them all the same size. Each pad was padded to the length of its longest sample. We padded not only the inputs, but also the labels. We evaluated our model and its predictions on the test set with accuracy, precision, recall, and f1-score. After the predictions had been made, we needed to do some postprocessing. We picked the predicted index (with the maximum logit) for each token, converted it to its string label and ignored wherever we put a -100 label.

We repeated the training procedure with the three training sets, each generated with a different UQS threshold as described earlier. Table 7 shows the results on the test data, both overall and per class.

Overall, our model showed a good performance predicting the target spans. The scores for the dominant class "non-targeting" (0) were higher than the scores for the "targeting" class. The Weighted F1 scores ranged from 74 to 79% , which is significantly higher than the results for the toxic span detection task in (Sharma et al., 2021) although the tasks, data and annotations are different across these tasks. The best results were again obtained when the UQS threshold was set to 50.

UQS	Class	Recall	Precision	F1-score	Support
50	All	81%	78%	79%	
	0	96%	97%	96%	14404
	1	73%	73%	73%	2051
	2	75%	64%	69%	906
60	All	75%	81%	77%	
	0	97%	95%	96%	14404
	1	68%	76%	72%	2051
	2	58%	72%	64%	906
70	All	67%	82%	74%	
	0	99%	93%	96%	14404
	1	61%	76%	68%	2051
	2	42%	76%	54%	906

Table 7: Test results overall and per class when the UQS threshold on the training set is 50, 60 or 70; class 0 = non-targeting; class 1= targeting-beginning; class 2= targeting-inside

7 Conclusion

We presented an extension to the HateXplain data set with annotations for target spans using crowd-annotation. The extended data set will enable the community to train and test models that recognize not only toxic language, but also the referents that are targeted. This is essential for future systems that need to comment on "wrong" behaviour in possibly interactive settings, discussing who has been targeted by what aspect and what toxic comments are used against the targeted person or community.

We provided the guidelines and instructions with clear examples of what we meant by target in a toxic sentence. We collected expert-annotated data for two of the batches with reasonable agreement among annotators. We obtained crowd annotations for target tokens in 3,480 sentences that targeted one target community. We also analyzed frequent patterns observed in the annotations and provided a statistical overview of the collected annotations.

We fine-tuned three RoBERTa-base language models with our data and investigated how changing the UQS threshold would affect the results. Our best model resulted in an F1-score of 79% on the test set, which was higher than other works in the field of toxic span classification. All the required information regarding the data and models is available on our Github repository[3]. In future work, we will extend the data to multiple languages as well as to richer and longer contexts, such as in conversational settings, where toxic expressions and targets can be mentioned sparsely. We want to explore other language models and compare their results by changing the hyperparameters and training arguments. Also, we are keen to compare the predictions of these models to the crowd-annotations and perform some error analysis.

Acknowledgements

This research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/ or sponsors.

References

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider

[3]The link to our Github repository: https://github.com/cltl/Target-Spans-Detection

disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 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, pages 512–515.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080*.

Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence labeling for word and sentence segmentation. In *EMNLP 2013*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Koombs, Shreya Havaldar, G J Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Olmos, Adam Omary, Christina Park, Clarisa Wang, Xin Wang, and Morteza Dehghani. 2020. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Svetlana Kiritchenko and Isar Nejadgholi. 2020. Towards ethics by design in online abusive content detection. *CoRR*, abs/2010.14952.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex &amp Intelligent Systems*.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Stefan Palan and Christian Schitter. 2018. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

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

Florian Pusse, Asad Sayeed, and Vera Demberg. 2016. Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 57–61.

Darsh J Shah, Sinong Wang, Han Fang, Hao Ma, and Luke Zettlemoyer. 2021. Reducing target group bias in hate speech detectors.

Mayukh Sharma, Ilanthenral Kandasamy, and Wb Vasantha. 2021. Youngsheldon at semeval-2021 task 5: Fine-tuning pre-trained language models for toxic spans detection using token classification objective. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 953–959.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *CoRR*, abs/2004.01670.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.