

1 Data Overview

To understand the annotation variances caused by different concept definitions, we study three prevalently used conceptions of harmful content – hateful, offensive, and toxic. We select definitions frequently cited or used in the literature. For hateful we use Davidson et al.’s (2017) definition; for offensive we use a variation of Wiegand et al.’s (2018) definition highlighting components that differentiate it from hatefulness; and for toxic we use Perspective API’s definition (see Table 1 for the summary). We examine whether definitional dimensions characterize differences in how annotators label content as any of hateful, offensive, or toxic. In this section, we describe how we collected the comments to annotate, how we designed the annotation tasks, and how we analyzed the labels that annotators produced.

Concept	Definition
Hateful	“expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group” (Davidson et al. 2017)
Offensive	“contains hurtful, derogatory, or obscene comments” (Wiegand et al. 2018)
Toxic	“a rude, disrespectful, or unreasonable comment that is likely to make readers want to leave a discussion” (Wulczyn et al., 2017)

Table 1. Definitions of “hateful”, “offensive”, and “toxic” that we provided to annotators.

2 Data Collection

2.1 Comments

We assembled a collection of comments from Reddit, Twitter, and YouTube about engaging political news stories to assess annotators’ interpretations of different concepts. We focused on political news comments to emulate the conditions under which the popular toxicity detection model Perspective API was trained on (Wulczyn et al., 2017). We focused on engaging stories because those stories reflect real world human attention to some extent (e.g., more engaging stories receive more comments).

To identify popular news stories everyday, we partnered with a third-party company called NewsWhip (NewsWhip, 2021), which monitors social media posts containing URL links to news publishers on mainstream platforms (e.g., Twitter, Facebook) and tracks the engagement metrics (e.g. likes, shares, retweets) of each post. From the NewsWhip database, we queried the most engaging¹ 1,000 URLs shared on both Twitter and Facebook for each day of August 2021. We extracted the publishing timestamps, headlines, and summaries for a total of 51,747

¹ For Twitter, the engagement score sums up the number of tweets and retweets associated with a URL. For Facebook, the engagement score provided by NewsWhip sums up the number of likes, shares, reactions, and comments of all posts containing the specific URL.

URLs after removing duplicates. Following Bakshy et al.'s (2015) approach, we constructed a machine learning classifier to identify URLs about political news amounting to 24,219 news URLs total.

We searched for each news URL's appearance on Reddit, Twitter, and YouTube. Both Reddit API and Twitter API support search for a URL address. For Reddit, we obtained Reddit posts containing the target URL and used PRAW (Boe, 2012) to collect all comments under the posts. For Twitter, we only collected original tweets containing the target URL (i.e., we excluded retweets and replies) and used Twitter API (Twitter, 2023) to collect all replies under the original tweets. For YouTube, because many YouTube videos do not mention the URL address in the description, we searched for the story headline of the URL in the YouTube search bar and sent web requests to scrape all returned videos deemed relevant by YouTube. By using an embedding model *SentenceTransformers* (Reimers & Gurevych, 2019), we computed the cosine similarity between the URL headline embedding and video title embedding. When the cosine similarity was larger than 0.8², we interpreted that YouTube videos were discussing the same issues raised in the target URL. We scraped all comments under those matched videos.

We then filtered for only URLs with at least ten comments on all three platforms for a total of 1,287 news URLs with corresponding 6,554 Reddit posts, 265,632 original tweets, and 11,743 YouTube videos. We kept only comments responding to original posts with these news URLs within 24 hours to keep the discussion period consistent, resulting in 483,762 Reddit comments, 1,496,623 Twitter replies, and 2,718,404 YouTube comments.

HOT comments are generally rare (Ibrahim et al., 2018). To avoid annotator fatigue from receiving an excess of comments with no HOT characteristics to label, we used purposive sampling instead of random sampling to increase the prevalence of HOT comments in the samples we provided to annotators. Specifically, for each of the HOT concepts, we used a pre-trained machine learning model to assign a classifier score (ranging from 0 to 1) to each comment. We used Aluru et al.'s hate speech model³ to classify hatefulness (Aluru et al., 2020), Davidson et al.'s offensiveness model⁴ to classify offensiveness (Davidson et al., 2017), and Jigsaw's Perspective API⁵ to classify toxicity (Jigsaw, 2021). Higher classifier scores predict more annotators would label the comment as HOT. Next, for each concept on each platform, we binned the comments into ten strata (e.g., 0-0.1, 0.1-0.2, etc.) based on their classifier scores. Given our annotation budget, we sampled 40 comments from each of the ten strata. Our final dataset included with 1162 Reddit comments, 1154 Twitter reply tweets, 1165 YouTube

² One author randomly selected ten URL headlines, and manually annotated the relevance of all returned videos. Using the cosine similarity score as input and human annotation as output, a threshold of 0.63 yielded the best F1 score (0.77, precision=0.76, recall=0.79). In this task, we wanted to prioritize precision over recall, we thus used 0.8 as the threshold to determine video relevancy (precision=0.93, recall=0.32).

³ Hate speech detection model available at: <https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-english>

⁴ Offensiveness detection model available at: <https://github.com/t-davidson/hate-speech-and-offensive-language>

⁵ Toxicity detection model available at: <https://perspectiveapi.com/>

comments⁶. Figure 1 provides an overview of our comment collection process, with additional details on comment collection and sampling in the Appendix.

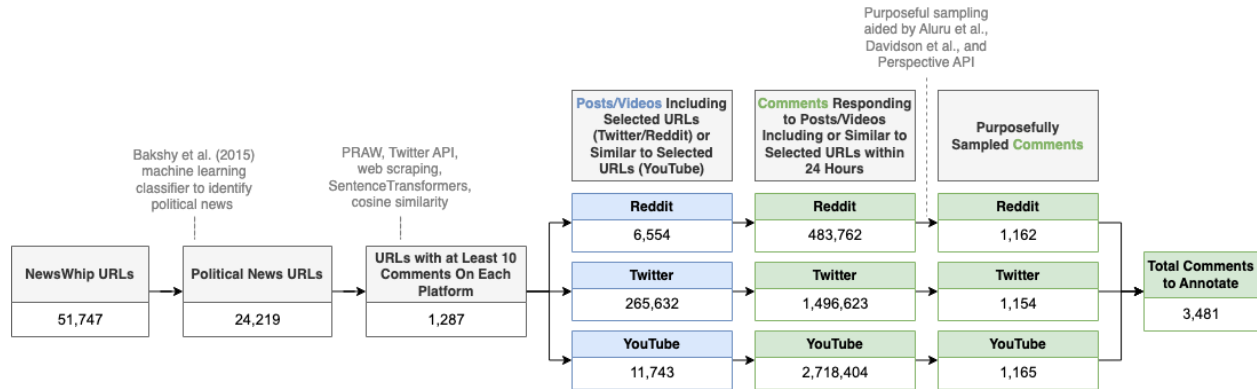


Figure 1. Social media comments data collection process.

2.2 Annotation Task

We recruited annotators on Amazon Mechanical Turk (MTurk). We required that annotators resided in the US, had completed at least 1000 Human Intelligence Tasks (HITs), and had at least a 98% HIT acceptance rate. We asked MTurk workers (annotators) who met these criteria to complete a qualification task. In this qualification task, we provided the concept definitions, labeling instructions (see Figure 2 for sample instructions), and three qualification questions (see Figure 3 for sample qualification task). Annotators had to answer all three questions correctly to be invited to the subsequent comment-labeling task. Once qualified, each annotator could label up to 100 comments.

Table 2 details the demographic characteristics of our annotator pool. Overall, our annotator pool was roughly equally split between male and female. Most were between age 30 and age 49, straight, White, and leaned toward Democratic party affiliation. This demographic distribution is not representative of a social media commenting population as a whole, and as such, it is possible that this distribution affects annotation outcomes. We account for possible rater effects in our statistical models.

⁶ One might expect 1,200 comments per platform (3 concepts x 10 stratum x 40 comments per stratum). We obtained annotated comments slightly smaller than 1,200 because the hate speech model rarely outputs very high classifier scores. We thus did not obtain 40 comments for either the 0.8-0.9 or the 0.9-1 strata for the hateful concept.


Instructions and Sample

In each HIT, you will be asked to label comments or tweets. One example tweet is shown below.

For each tweet/comment, you will be asked to assess whether or not it is **hateful**. You should label a tweet/comment as **hateful** if, in your judgment, it **"expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group."** Note that this requires **expression against a group**; simply saying "I hate you, John Doe." would not qualify as hateful under this definition.

Sometimes context can help you in making a judgment about the tweet/comment. If the tweet/comment was replying to another one, you can click the button that says, "Show tweet/comment this is replying to". You can also click on the button that says, "Show news article headline" in order to see the headline of the news article that prompted the conversation this comment is part of. Please use these only to understand the context of the tweet/comment you are judging; you should judge **only the original tweet/comment**, not the things it is responding to.

Tweet to Label



[Show tweet this is replying to](#)
 ↑ [Click here to see what the tweet is replying to](#)
 (required now; optional when you're doing real labeling tasks)
[Show news article headline](#) (temporary disabled)

Figure 2. Sample instructions.

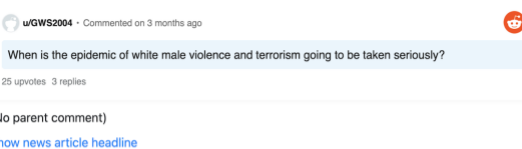
Practice Labeling

In each HIT, you will be asked to label comments or tweets.

For each tweet/comment, you will be asked to assess whether or not it is **toxic**. You should label a tweet/comment as toxic if, in your judgment, it is **"likely to make readers want to leave a discussion."**

Sometimes context can help you in making a judgment about the tweet/comment. If the tweet/comment was replying to another one, you can click the button that says, "Show tweet/comment this is replying to". You can also click on the button that says, "Show news article headline" in order to see the headline of the news article that prompted the conversation this comment is part of. Please use these only to understand the context of the tweet/comment you are judging; you should judge **only the original tweet/comment**, not the things it is responding to.

Comment to Label



Your Judgment

I cannot decide because the comment is deleted
 I don't know enough to judge this comment

↑ If you select any of these two options, this comment will disappear from this HIT and it's not counted into the 10 comments you will label.

Do you think this comment is toxic?

Yes

Do you think this comment is toxic?

Yes
 No

Explanations

You labeled this comment as toxic. We disagreed.

While some might disagree about whether such an epidemic exists or that it isn't taken seriously, we don't think the question as posed would drive people away from a conversation. Your judgment may differ!

Ratings From Others

Sorry, no one else has labeled this comment yet.

← Prev (#1 of 5 practice comments/tweets) Next →

Do you think this comment is toxic?

Yes
 No

[Get Feedback](#)

↑ [Click here to see ratings from others and explanations from experts.](#)
 When you do so, you **cannot change** your current rating on this page any more.
 Note that in the actually labeling HIT, you could only opt to see ratings from other crowdsourced workers.

← Prev (#1 of 5 practice comments/tweets) Next →

Figure 3. Sample qualification task.

Demographic Characteristic	No. Annotators	Demographic Characteristic	No. Annotators
Sex		Sexual Orientation	
Female	285	Straight	521
Male	311	Gay or lesbian	15
Non-binary	10	Bisexual	58
Unknown	2	Unknown or skip	14
Age		Race/Ethnicity	
18-29	95	White	471
30-39	238	Black	45
40-49	148	Asian	26
50-59	81	Latino or Hispanic	21
>60	44	Unknown or mixed	45
Unknown or skip	2	Political Affiliation	
		Lean Democrat	382
		Independent	81
		Lean Republican	145

Table 2. Demographic characteristics of our annotator pool.

We provided annotators with definitions of all three concepts and asked them to label comments for the presence of each. Figure 4 is a screenshot of a comment provided to an annotator, and Figure 5 is a screenshot of the task provided to annotators. Annotators were allowed to navigate to previous questions and change their answers. However, for each comment, they could change their answer only once. Each comment received five annotator labels, and we targeted a \$15 hourly rate for annotators.

Comment to Label

@60Minutes So he sold COVID-19 vaccine? Clutching my pearls...NOT!

Comment in Context

Literally a good idea !

@60Minutes So he sold COVID-19 vaccine? Clutching my pearls...NOT!

Title of Video that Started this Discussion

[Florida Governor Ron DeSantis confronted over Publix-COVID vaccination deal](#)

Figure 4. Sample comment provided to annotator for review prior to labeling HOT concepts.

I cannot decide because the comment is deleted

I don't know enough to judge this comment

↑ If you select any of these two options, this tweet will disappear from this HIT and it's not counted into the 10 tweets you will label.

Do you think this tweet is hateful?

Yes

No

Do you think this tweet is offensive?

Yes

No

Do you think this tweet is toxic?

Yes

No

Figure 5. Annotation task provided to annotators.

3 References

- Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep Learning Models for Multilingual Hate Speech Detection. In *arXiv [cs.SI]*. arXiv. <http://arxiv.org/abs/2004.06465>
- Bakshy, E., Messing, S., & Adamic, L. A. (2015). Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132.

<https://doi.org/10.1126/science.aaa1160>

Boe, B. (2012). *PRAW: The Python Reddit API Wrapper — PRAW 7.7.0 documentation*.

PRAW. <https://praw.readthedocs.io/en/stable/>

Davidson, T., Warmley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Eleventh International AAAI Conference on Web and Social Media*. <https://doi.org/10.1609/icwsm.v11i1.14955>.

Ibrahim, M., Torki, M., & El-Makky, N. (2018). Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 875–878.

<https://doi.org/10.1109/ICMLA.2018.00141>

Jigsaw. (2021). *Perspective API*. Perspective API. www.perspectiveapi.com

NewsWhip. (2021, November 26). *Real-time media monitoring*. NewsWhip.

<https://www.newswhip.com/>

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 3982–3992. <http://arxiv.org/abs/1908.10084>

Twitter. (2023). *Developer Platform*. Twitter. <https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction>

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. *Proceedings of the GermEval 2018 Workshop*.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina. In *Proceedings of the 26th International Conference on World Wide Web*. <https://doi.org/10.1145/3038912.3052591>