Hate Speech and Counter Speech Detection: Conversational Context Does Matter

Xinchen Yu
University of North Texas
xinchenyu@my.unt.edu

Eduardo Blanco
Arizona State University
eduardo.blanco@asu.edu

Lingzi Hong
University of North Texas
lingzi.hong@unt.edu

Abstract

Hate speech is plaguing the cyberspace along with user-generated content. This paper investigates the role of conversational context in the annotation and detection of online hate and counter speech, where context is defined as the preceding comment in a conversation thread. We created a context-aware dataset for a 3-way classification task on Reddit comments: hate speech, counter speech, or neutral. Our analyses indicate that context is critical to identify hate and counter speech: human judgments change for most comments depending on whether we show annotators the context. A linguistic analysis draws insights into the language people use to express hate and counter speech. Experimental results show that neural networks obtain significantly better results if context is taken into account. We also present qualitative error analyses shedding light into (a) when and why context is beneficial and (b) the remaining errors made by our best model when context is taken into account.

1 Introduction

The advent of social media has democratized public discourse on an unparalleled scale. Meanwhile, it is considered a particularly conducive arena for hate speech (Caiani et al., 2021). Online hate speech is prevalent and can lead to serious consequences. At the individual level, the victims targeted by hate speech are frightened by online threats that may materialize in the real world (Olteanu et al., 2018). At the societal level, it has been reported that there is an upsurge in offline hate crimes targeting minorities (Olteanu et al., 2018; Farrell et al., 2019).

There are two common strategies to combat online hate: disruption and counter speech. Disruption refers to blocking hateful content or users. To scale this strategy, researchers have proposed methods to identify hate (Waseem and Hovy, 2016; Davidson et al., 2017; Nobata et al., 2016). While these interventions could de-escalate the impact of hate speech, they may violate online free speech (Mathew et al., 2019). Additionally, attacks at the micro-level may be ineffective as hate networks often have rapid rewiring and self-repairing mechanisms (Johnson et al., 2019). Counter speech refers to the “direct response that counters hate speech” (Mathew et al., 2019). It has been shown to be more effective in the long term than disruption in theoretical and empirical studies (Richards and Calvert, 2000; Mathew et al., 2020). Identifying hate and counter speech in natural conversations is critical to understand effective counter speech strategies and the generation of counter speech.

Most corpora with either hate speech (Hate) or counter speech (Counter-hate) annotations do not include conversational context. Indeed, they annotate a user-generated comment as Hate or Counter-hate based on the comment in isolation (Davidson et al., 2017; Waseem and Hovy, 2016; Mathew et al., 2019; He et al., 2021). Therefore, systems trained on these corpora fail to consider the effect of contextual information on the identification of Hate and Counter-hate. Recent studies have shown that context affects annotations in toxic-
ity and abuse detection (Pavlopoulos et al., 2020; Menini et al., 2021). We further investigate the effect of context on the task of identifying Hate and Counter-hate. Table 1 shows examples where a comment, denoted as Target, is Hate, Neutral or Counter-hate depending on whether the preceding comment, denoted as Parent, is taken into account. In the top example, the Target goes from Neutral to Hate when taking into account the Parent: it becomes clear that the author is disparaging short men. In the bottom example, the Target goes from Hate to Counter-hate as the author uses offensive language to counter the hateful content in the Parent. This is a common strategy to express counter speech (Mathew et al., 2019).

In this study, we answer the following questions:
1. Does conversational context affect if a comment is perceived as Hate, Neutral, or Counter-hate by humans? (It does.)
2. Do models to identify Hate, Neutral, and Counter-hate benefit from incorporating context? (They do.)

To answer the first question, we create a collection of (Parent, Target) Reddit comments and annotate the Targets with three labels (Hate, Neutral, Counter-hate) in two independent phases: showing annotators (a) only the Target or (b) the Parent and the Target. We limit context to the parent comment. While the full conversation could provide additional information, it is also known to affect annotators’ stance (Dutta et al., 2020) and introduce biases. We find that human judgments are substantially different when the Parent is shown. Thus the task of annotating Hate and Counter-hate requires taking into account the context.

To answer the second question, we experiment with context-unaware and context-aware classifiers to detect if a given Target is Hate, Neutral, or Counter-hate. Results show that adding context does benefit the classifiers significantly.

In summary, the main contributions of this paper are:
1. A corpus with 6,846 pairs of (Parent, Target) Reddit comments and annotations indicating whether the Targets are Hate, Neutral, or Counter-hate; (b) annotation analysis showing that the problem requires taking into account context, as the ground truth changes; (c) corpus analysis detailing the kind of language people use to express Hate and Counter-hate; (d) experiments showing that context-aware neural models obtain significantly better results; and (e) qualitative analysis revealing when context is beneficial and the remaining errors made by the best context-aware model.

2 Related Work

Hate speech in user-generated content has been an active research area recently (Fortuna and Nunes, 2018). Researchers have built several datasets for hate speech detection from diverse sources such as Twitter (Waseem and Hovy, 2016; Davidson et al., 2017), Yahoo! (Nobata et al., 2016), Fox News (Gao and Huang, 2017), Gab (Mathew et al., 2021) and Reddit (Qian et al., 2019).

Compared to hate speech detection, few studies focus on detecting counter speech (Mathew et al., 2019; Garland et al., 2020; He et al., 2021). Mathew et al. (2019) collect and hand-code 6,898 counter hate comments from YouTube videos targeting Jews, Blacks and LGBT communities. Garland et al. (2020) work with German tweets and define hate and counter speech based on the communities to which the authors belong. He et al. (2021) use a collection of hate and counter hate keywords relevant to COVID-19 and create a dataset containing 359 counter hate tweets targeting Asians. Another line of research focuses on curating datasets for counter speech generation using crowdsourcing (Qian et al., 2019) or with the help of trained operators (Chung et al., 2019; Fanton et al., 2021). However, synthetic language is rarely as rich as language in the wild. Even if it were, conclusions and models from synthetic data may not transfer to the real world. In this paper, we work with user-generated content expressing hate and counter-hate rather than synthetic content.

Table 2 summarizes existing datasets for Hate and Counter-hate detection. Most of them do not include context information. In other words, the preceding comments are not provided when annotating Targets. Context does affect human judgments and has been taken into account for Hate detection (Gao and Huang, 2017; Pavlopoulos et al., 2020; Menini et al., 2021; Vidgen et al., 2021). Gao and Huang (2017) annotate hateful comments in the nested structures of Fox News discussion threads. Vidgen et al. (2021) introduce a dataset of Reddit comments with annotations in 6 categories taking into account context. However, the inter annotator agreement is low (Fleiss’ Kappa 0.267) and the

---

1 The examples in this paper contain hateful content. We cannot avoid it due to the nature of our work.

2 Code and data available at https://github.com/xinchenyu/counter_context
Table 2: Comparison of corpora with hate (above dashed line) and counter-hate annotations (below dashed line, some also include hate). Vidgen et al. (2021) is the only one considering counter-hate and context, but they only have 220 instances of counter-hate. Numbers between parenthesis indicate the number of counter-hate instances.

3 Dataset Collection and Annotation

We first describe our procedure to collect (Parent, Target) pairs, where both Parents and Targets are Reddit comments in English. Then, we describe the annotation guidelines and the two annotation phases: showing annotators (a) only the Target and (b) the Parent and Target. The two independent phases allow us to quantify how often context affects the annotation of Hate and Counter-hate.

3.1 Collecting (Parent, Target) pairs

In this work, we focus on Reddit, a popular social media site. It is an ideal platform for data collection due to the large size of user populations and many diverse topics (Baumgartner et al., 2020). We use a list of hate words to retrieve Reddit conversations to keep the annotation costs reasonable while creating a (relatively) large corpus of counter speech. We start with a set of 1,726 hate words from two lexicons: Hatebase and a harassment corpus (Rezvan et al., 2018). We remove ambiguous words following ElSherief et al. (2018). To collect (Parent, Target) pairs, we use the following steps. First, we retrieve comments containing at least one hate word (comment w/ hateword). Second, we create a (Parent, Target) pair using comment w/ hateword as Target and its preceding comment as Parent. Third, we create a (Parent, Target) pair using comment w/ hateword as Target and its preceding comment as Parent. Third, we create a (Parent, Target) pair using comment w/ hateword as Target and its preceding comment as Parent. Lastly, we remove pairs if the same author posted the Parent and Target. We retrieve 6,846 (Parent, Target) pairs with PushShift (Baumgartner et al., 2020) from 416 submissions. We also collect the title of the discussion from which each pair is retrieved.

3.2 Annotation Guidelines

To identify whether a Target is Hate, Neutral, or Counter-hate, we crowdsource human judgments from non-experts. Our guidelines reuse the definitions of Hate by Ward (1997) and Counter-hate by Mathew et al. (2019) and Vidgen et al. (2021):

- **Hate**: the author attacks an individual or a group with the intention to vilify, humiliate, or incite hatred;
- **Counter-hate**: the author challenges, condemns the hate expressed in another comment, or calls out a comment for being hateful;
- **Neutral**: the author neither conveys hate nor opposes hate expressed in another comment.

**Annotation Process** We chose Amazon Mechanical Turk (MTurk) as the crowdsourcing platform. We replace user names with placeholders (User_A

number of Counter-hate instances is small (220). Moreover, both studies use contextual information without identifying the role context plays in the annotation and detection. Pavlopoulos et al. (2020) allow annotators to see one previous comment to annotate Wikipedia conversations. They find context matters in the annotation but provide no empirical evidence showing whether models to detect toxicity benefit from incorporating context. Menini et al. (2021) re-annotate an existing corpus to investigate the role of context in abusive language. They found context does matter. Utilizing conversational context has also been explored in text classification tasks such as sentiment analysis (Ren et al., 2016), stance (Zubiaga et al., 2018) and sarcasm (Ghosh et al., 2020). In this paper, we investigate the role of context in Hate and Counter-hate detection.
and User_B) owing to privacy concerns. The annotations took place in two independent phases. In the first phase, annotators are first shown the Parent comment. After a short delay, they click a button to show the Target and then after another short delay they submit their annotation. Delays are at most a few seconds and proportional to the length of the comments. Our rationale behind the delays is to “force” annotators to read the Parent and Target in order. In the second phase, annotators label each Target without seeing the preceding Parent comment. A total of 375 annotators were involved in the first phase and 299 in the second phase. There is no overlap between annotators thus we eliminated the possibility of biased annotators remembering the Parent in the second phase.

**Annotation Quality** Crowdsourcing may attract spammers (Sabou et al., 2014). For quality control, we first set a few requirements for annotators: they must be located in the US and have a 95% approval rate over at least 100 Human Intelligence Tasks (HITs). We also block annotators who submit more than 10 HITs with an average completion time below 5 seconds (half the time required in our pilot study). As the corpus contains vulgar words, we require annotators to pass the Adult Content Qualification Test. The reward per HIT is $0.05.

The second effort is to identify bad annotators and filter out their annotations until we obtain substantial inter-annotator agreement. We collect five annotations per HIT and compute MACE (Hovy et al., 2013, Multi-Annotator Competence Estimation) for each annotator. MACE is devised to rank annotators by their competence and adjudicate disagreements based on annotator competence (not the majority label). Then, we use Krippendorff’s α (Krippendorff, 2011) to estimate inter-annotator agreement: α coefficients at or above 0.6 are considered substantial (above 0.8 are considered nearly perfect) (Artstein and Poesio, 2008). We repeat the following steps until α ≥ 0.6:

1. Use MACE to calculate the competence score of all annotators.
2. Discard all the annotations by the annotator with the lowest MACE score.
3. Check Krippendorff’s α on the remaining annotations. Go to (1) if α < 0.6.

The final corpus consists of 6,846 (Parent, Target) pairs and a label assigned to each Target (Hate, Counter-hate, or Neutral). The ground truth we experiment with (Section 5) is the label obtained taking into account the Parent (first phase). The second phase, which disregards the Parent, was conducted for analysis purposes (Section 4). We split the corpus into two subsets: (a) Gold (4,751 pairs with α ≥ 0.6) and (b) Silver (2,095 remaining pairs). As we shall see, the Silver pairs are useful to learn models.

### 4 Corpus Analysis

**Does conversational context affect if a comment is perceived as Hate or Counter-hate?** Yes, it does. Table 3 presents the percentage of labels that change and remain the same depending on whether annotators are shown the Parent of the Target comment (with and without).

| Example | With | Without |
|---------|------|---------|
| Parent: That chick needs a high-five in the face with a chair. Damn her for making us look bad! | Hate | Neutral |
| Target: A brick is more effective. | Counter | Neutral |
| Parent: If I knew her I would sh*t in her mailbox. | Target: The poor mail carrier in that neighborhood doesn’t deserve that. | Counter | Neutral |
| Parent: Go watch your incest porn on your own time. | Target: You’re a sick person. | Counter | Hate |

Table 3: Confusion matrix (percentages) showing annotation changes depending on whether annotators are shown the Parent of the Target comment.

| Example | With | Without |
|---------|------|---------|
| Parent: : You’re a sick person. | Counter | Hate |

Table 4: Examples of Target comments whose labels change depending on whether annotators are shown the Parent of the Target comment (with and without).

The second phase, which disregards the Parent, was conducted for analysis purposes (Section 4). We split the corpus into two subsets: (a) Gold (4,751 pairs with α ≥ 0.6) and (b) Silver (2,095 remaining pairs). As we shall see, the Silver pairs are useful to learn models.
Table 5: Linguistic analysis comparing the *Titles*, *Parents* and *Targets* in Counter-hate and Hate *Target* comments. Number of arrows indicate the p-value (t-test: one: p<0.05, two: p<0.01, and three: p<0.001). Arrow direction indicates whether higher values correlate with Counter-hate (up) or Hate (down). A check mark (✓) indicates that the test passes the Bonferroni correction.

| Textual factors                          | Title p-value | Parent p-value | Target p-value |
|------------------------------------------|---------------|----------------|----------------|
| Total tokens                             | ↓↓            | ↑↑↑            |                 |
| Question marks                           |               | ✓              |                |
| 1st person pronouns                      |               |                | ↑↑↑            |
| 2nd person pronouns                      | ↑↑↑           | ✓              |                |
| Sentiment and cognitive factors          |               |                |                |
| Profanity words                          | ↑↑↑           | ✓              |                |
| Problem-solving words                    |               |                |                |
| Awareness words                          | ↑↑↑           |                |                |
| Negative words                           |                | ↑↑↑           |                |
| Disgust words                            |               |                |                |
| Enlightenment words                      |               |                |                |
| Conflicting words                        | ↓↓↓           | ✓              |                |

Counter-hate with context to shed light on the differences between the language people use in hate and counter speech. We combine the set of hate words with profanity words to check for profanity words. We analyze sentiment and cognitive factors using the Sentiment Analysis and Cognition Engine (SEANCE) lexicon, a popular tool for psychological linguistic analysis (Crossley et al., 2017). Statistical tests are conducted using unpaired t-tests between the groups, of which the *Targets* are Counter-hate or Hate (Table 5). We also report whether each feature passes the Bonferroni correction as multiple hypothesis tests are performed. We draw several interesting insights:

- Questions marks in *Target* signal Counter-hate. They are often rhetorical questions.
- Fewer 1st person pronouns (e.g., I, me) and more 2nd person pronouns (e.g., you, your) in the *Parent* signal that the *Target* is more likely to be Counter-hate. This is due to the fact that people tend to target others in hateful content.
- High profanity count in the *Parent* signals that the *Target* is Counter-hate, while high profanity count in the *Target* signals Hate.
- More words related to awareness, enlightenment and problem-solving in the *Target* signal Counter-hate.
- When there are more negative words in the *Parent*, the *Target* tends to be Counter-hate. *Targets* labeled as Counter-hate contain fewer negative and disgusting words.

https://github.com/RobertJGabriel/google-profanity-words-node-module/blob/master/lib/profanity.js
5 Experiments and Results

We build neural network models to identify if a Target comment is Hate, Counter-hate, or Neutral. We randomly split Gold instances (4,751) as follows: 70% for training, 15% for validation and 15% for testing. Silver instances are only used for training.

Neural Network Architecture We experiment with neural classifiers built on top of the RoBERTa transformer (Liu et al., 2019). The neural architecture consists of a pretrained RoBERTa transformer, a fully connected layer (768 neurons and Tanh activation), and another fully connected layer (3 neurons and softmax activation) to make predictions (Hate, Counter-hate, or Neutral). To investigate the role of context, we consider two textual inputs:

- the Target alone (Target), and
- the Parent and the Target (Parent_Target).

We concatenate the Target and the Parent with the [SEP] special token. We conduct multiple runs of experiments, which show consistent results. The hyperparameters and other implementation details are presented in the Appendix. We also experiment with models that take the title of the discussion as part of the context, but it is not beneficial.

We implement two strategies to enhance the performance of neural models:

**Blending Gold and Silver** We adopt the method by Shnarch et al. (2018) to determine whether Silver annotations are beneficial. There are two phases in the training process: m blending epochs using all Gold and a fraction of Silver, and then n epochs using all Gold. In each blending epoch, Silver instances are fed in a random order to the network.

The fraction of Silver is determined by a blending factor \( \alpha \in [0..1] \). The first blending epoch is trained with all Gold and all Silver, and the amount of Silver to blend is reduced by \( \alpha \) in each epoch.

**Pretraining with Related Tasks** We also experiment with several corpora to investigate whether pretraining with related tasks is beneficial. Specifically, we pretrain our models with existing corpora annotating:

1. hateful comments: hateful or not hateful (Qian et al., 2019), and hate speech, offensive, or neither (Davidson et al., 2017);
2. sentiment: negative, neutral, or positive (Rosenthal et al., 2017);
3. sarcasm: sarcasm or not sarcasm (Ghosh et al., 2020); and
4. stance: agree, neutral, or attack (Pougué-Biyong et al., 2021).

5.1 Quantitative Results

We present results with the test split in Table 6. The majority baseline always predicts Neutral. The remaining rows present the results with the different training settings: training with the Target or both the Parent and Target; training with only Gold or blending Silver annotations; and pretraining with related tasks. We provide here results pretraining with the most beneficial task, stance detection, and present additional results in the Appendices. Blending Gold and Silver annotations requires tuning \( \alpha \). We did so empirically using the training and validation splits, like other hyperparameters. We found the optimal value to be 0.3 when blending Silver (+ Silver rows) and 1.0 when blending Silver and pretraining with a related task (+Silver + Related task rows).

As shown in Table 6, blending Gold and Sil-
6 Qualitative Analysis

When is adding the context beneficial? When does our best model make mistakes? To investigate these questions, we manually analyze the following:

- The errors made by the Target only network that are fixed by the context-aware network (Trained with Parent_Target, Table 7).
- The errors made by the context-aware network pretrained on related task (stance) and blending Silver annotations (Parent_Target+Silver+Related task, Table 8).

When does the context complement Target?
We analyze the errors made by the network using only the Target that are fixed by the context-aware network (Trained with Parent_Target). Table 7 exemplifies the most common error types.

The most frequent type of error fixed by the context-aware model is when there is Lack of information in the Target (48%). In this case, the Parent comment is crucial to determine the label of the Target. In the example, knowing what the author of the Target refers to (i.e., a rhetorical question, "Women can hover?") is crucial to determine that the Target is humiliating women as a group.

The second most frequent error type is Negation (27%). In the example in Table 7, taking into account the Parent allows the context-aware network to identify that the author of the Target is scolding the author of Parent and thus countering hate.

Nobata et al. (2016) and Qian et al. (2019) have pointed out that sarcasm and irony make detecting abusive and hateful content difficult. We find evidence supporting this claim. We also discover that by incorporating the Parent comment, a substantial amount of these errors are fixed. Indeed, 19% of
errors fixed by the context-aware network include sarcasm or irony in the Target comment.

Finally, the context-aware network taking into account the Parent fixes many errors (8%) in which the Target comment is Hate despite it does not contain swear words. In the example, the Target is introducing additional hateful content, which can be identified by the context-aware model when the Parent information is used.

When does the best model make errors? In order to find out the most common error types made by the best model (context-aware, Parent_Target+Silver+Related task), we manually analyze 200 random samples in which the output of the network differs from the ground truth. Table 8 shows the results of the analysis.

Despite 27% of errors fixed by the context-aware network (i.e., taking into account the Parent) include negation in the Target, negation is the most common type of errors made by our best network (28%). The example in Table 8 is especially challenging as it includes a double negation.

We observe that Rhetorical questions are almost as common (27%). This finding is consistent with the findings by Schmidt and Wiegand (2017). In the example, the best model fails to realize that the Target is hateful, as it disdains the author of Parent.

Swear words are present in a substantial number of errors. Wrongly predicting a Target without swear words as Counter-hate or Neutral accounts for 8% of errors, and wrongly predicting a Target with swear words as Hate accounts for another 8% of errors. As pointed out by Davidson et al. (2017), hate speech may not contain hate or swear words. And vice versa, comments containing swear words may not be hateful (Zhang and Luo, 2019).

Finally, we observe Intricate text in 7% errors. Our best model identifies the Target (“I have lost all respect for her.”) as Hate probably because by identifying the agreeing stance on the Parent. Indeed, the author of Target expresses his/her attitude without vilifying others. Hence, the ground truth label is Neutral.

7 Conclusions and Future Work

Conversational context does matter in Hate and Counter-hate detection. We have demonstrated so by (a) analyzing whether humans perceive user-generated content as Hate or Counter-hate depending on whether we show them the Parent comment and (b) investigating whether neural networks benefit from incorporating the Parent. We find that 38.3% of human judgments change when we show the Parent to annotators. Experimental results demonstrate that networks incorporating the Parent yield better results. Additionally, we show that noisy instances (Silver data) and pretraining with relevant datasets improves model performance. We have created and released a corpus of 6,846 (Parent, Target) pairs of Reddit comments with the Target annotated as Hate, Neutral or Counter-hate.

Our work have several limitations. First, we only consider context as the parent comment. While considering additional context might be sometimes beneficial, doing so would require careful design to not bias annotations (Dutta et al., 2020). Our research agenda includes exploring reliably strate-
gies to consider more context and identify which parts are most important. Second, people may have different opinions about what constitutes hate and counter speech due to different tolerances in online aggression. We obtained the ground truth according to annotators’ reliability (MACE scores), which may lead to controversial samples falling in the Silver set and thus being considered only for training (not for testing). Finally, the keywords sampling used to create our corpus may introduce biases. Despite we partially mitigate the issue by considering hateful comments in both the Parent and Target, community-based sampling (Vidgen et al., 2021) could be applied in our future work.

8 Ethical Considerations

We use the PushShift API to collect data from Reddit. Our collection process is consistent with Reddit’s Terms of Service. The data are accessed through the data dumps on Google’s BigQuery using Python.

Reddit can be considered a public space for discussion which differs from a private messaging service (Vidgen et al., 2021). Users consent to have their data made available to third parties including academics when they sign up to Reddit. Existing ethical guidelines state that in this situation explicit consent is not required from each user (Procter et al., 2019). We obfuscate user names as User_A or User_B to reduce the possibility of identifying users. In compliance with Reddit’s policy, we would like to make sure that our dataset will be reused for non-commercial research only.

The Reddit comments in this dataset were annotated by annotators using Amazon Mechanical Turk. We have followed all requirements introduced by the platform for tasks containing adult content. A warning was added in the task title. Annotators need to pass the Adult Content Qualification Test before working on our tasks. Annotators were compensated on average with $8 per hour. We paid them regardless of whether we accepted their work. Annotators’ IDs are not included in the dataset.

Acknowledgements

This work was supported by Research Seed Grant of the UNT College of Information. We would like to thank the anonymous reviewers for their insightful comments and suggestions.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. Comput. Linguist., 34(4):555–596.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020, pages 830–839. AAAI Press.

Manuela Caiani, Benedetta Carlotti, and Enrico Padoan. 2021. Online hate speech and the radical right in times of pandemic: The Italian and English cases. Javnost - The Public, 28(2):202–218.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. Behavior research methods, 49(3):803–821.

Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. Information Processing & Management, 57(2):102085.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018, pages 42–51. AAAI Press.
Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3226–3240, Online. Association for Computational Linguistics.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In Proceedings of the 10th ACM Conference on Web Science, WebSci ’19, page 87–96, New York, NY, USA. Association for Computing Machinery.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. ACM Comput. Surv., 51(4).

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 102–112, Online. Association for Computational Linguistics.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Mure-san. 2020. A report on the 2020 sarcasm detection shared task. In Proceedings of the Second Workshop on Figurative Language Processing, pages 1–11, Online. Association for Computational Linguistics.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, page 90–94, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Nicola F Johnson, R Leahy, N Johnson Restrepo, Nicolas Velasquez, Ming Zheng, P Manrique, P Devkota, and Stefan Wuchty. 2019. Hidden resilience and adaptive dynamics of the global online hate ecology. Nature, 573(7773):261–265.

Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability. https://repository.upenn.edu/asc_papers/43/. Accessed: 2021-02-08.

Yinhao Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2):1–24.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Raiagaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019, pages 369–380. AAAI Press.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 14867–14875. AAAI Press.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2):153–157.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. CoRR, abs/2103.14916.

Chikashi Nobata, Joel R. Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, pages 145–153. ACM.

Alexandra Oltzanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. The effect of extremist violence on hateful speech online. In Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018, pages 221–230. AAAI Press.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In
Robert D Richards and Clay Calvert. 2000. Counterspeech 2000: A new look at the old remedy for bad speech. *BYU L. Rev.*, page 553.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

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.

Kenneth D Ward. 1997. Free speech and the development of liberal virtues: An examination of the controversies involving flag-burning and hate speech. *University of Miami Law Review*, 52(3):733–792.

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, San Diego, California. Association for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augustin. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.

### A Annotation Interface

We show a screenshot of the annotation interface in Figure 2.

### B Detailed Results

Table 9 presents detailed results complementing Table 6 in the paper. We provide Precision, Recall and
Figure 2: Screenshot of the annotation interface. The left panel displays the instructions and examples. The right panel displays the Parent and the Target to be annotated.

|                  | Hate       | Counter-hate | Neutral | Weighted Average |
|------------------|------------|--------------|---------|------------------|
|                  | P          | R            | F1      | P                | R            | F1      | P        | R        | F1       | P          | R        | F1       |
| Majority Baseline| 0.00       | 0.00         | 0.00    | 0.51             | 1.00         | 0.67     | 0.26     | 0.51     | 0.34     |
| Trained with ... |            |              |         |                  |              |          |          |          |          |
| Target           | 0.56       | 0.55         | 0.56    | 0.41             | 0.36         | 0.38     | 0.67     | 0.71     | 0.69     | 0.58       | 0.59     | 0.58     |
| + Hate_Twitter   | 0.58       | 0.53         | 0.55    | 0.46             | 0.07         | 0.12     | 0.61     | 0.88     | 0.72     | 0.57       | 0.6        | 0.54     |
| + Hate_Reddit    | 0.57       | 0.52         | 0.55    | 0.44             | 0.32         | 0.37     | 0.64     | 0.75     | 0.69     | 0.58       | 0.59     | 0.58     |
| + Sentiment      | 0.59       | 0.47         | 0.53    | 0.00             | 0.00         | 0.00     | 0.59     | 0.92     | 0.72     | 0.45       | 0.59     | 0.50     |
| + Sarcasm        | 0.59       | 0.51         | 0.55    | 0.50             | 0.04         | 0.08     | 0.59     | 0.51     | 0.55     | 0.57       | 0.58     | 0.51     |
| + Stance         | 0.56       | 0.55         | 0.56    | 0.51             | 0.41         | 0.45     | 0.68     | 0.74     | 0.71     | 0.61       | 0.61     | 0.61     |
| Trained with ... |            |              |         |                  |              |          |          |          |          |
| Parent_Target    | 0.55       | 0.62         | 0.59    | 0.52             | 0.38         | 0.44     | 0.68     | 0.72     | 0.70     | 0.61       | 0.62     | 0.61     |
| + Hate_Twitter   | 0.49       | 0.64         | 0.56    | 0.29             | 0.13         | 0.18     | 0.66     | 0.73     | 0.7       | 0.53       | 0.57      | 0.54     |
| + Hate_Reddit    | 0.55       | 0.64         | 0.59    | 0.48             | 0.33         | 0.39     | 0.69     | 0.73     | 0.71     | 0.61       | 0.62      | 0.61     |
| + Sentiment      | 0.53       | 0.59         | 0.56    | 0.40             | 0.23         | 0.29     | 0.68     | 0.77     | 0.72     | 0.57       | 0.60      | 0.58     |
| + Sarcasm        | 0.56       | 0.54         | 0.55    | 0.45             | 0.09         | 0.15     | 0.62     | 0.86     | 0.72     | 0.56       | 0.60      | 0.54     |
| + Stance         | 0.55       | 0.66         | 0.60    | 0.54             | 0.43         | 0.48     | 0.71     | 0.70     | 0.71     | 0.63       | 0.63      | 0.63     |

Table 9: Detailed results (P, R, and F) predicting whether the Target is Hate, Neutral or Counter-hate when the input is only the Target or the Parent_Target. These results are using RoBERTa and pretrained with each related task. This table complements Table 6 in the paper.

|                  | Hate       | Counter-hate | Neutral | Weighted Average |
|------------------|------------|--------------|---------|------------------|
|                  | P          | R            | F1      | P                | R            | F1      | P        | R        | F1       |
| Majority Baseline| 0.00       | 0.00         | 0.00    | 0.51             | 1.00         | 0.67     | 0.26     | 0.51     | 0.34     |
| Trained with ... |            |              |         |                  |              |          |          |          |          |
| + Silver + Related Task | 0.56 | 0.54 | 0.55 | 0.48 | 0.46 | 0.47 | 0.67 | 0.71 | 0.70 | 0.60 | 0.60 | 0.60 |
| Mean (SD)        | 0.04       | 0.05         | 0.01    | 0.01             | 0.05         | 0.03     | 0.01     | 0.04     | 0.01     | 0.00       | 0.01      | 0.01     |
| Trained with Parent_Target | 0.55 | 0.6 | 0.59 | 0.51 | 0.49 | 0.50 | 0.72 | 0.72 | 0.72 | 0.64 | 0.64 | 0.63 |
| + Silver + Related Task | 0.03 | 0.04 | 0.01 | 0.04 | 0.05 | 0.02 | 0.02 | 0.06 | 0.02 | 0.00       | 0.01      | 0.01     |
| Mean (SD)        | 0.03       | 0.04         | 0.01    | 0.04             | 0.05         | 0.02     | 0.02     | 0.06     | 0.02     | 0.00       | 0.01      | 0.01     |

Table 10: Detailed results (P, R, and F) predicting whether the Target is Hate, Neutral or Counter-hate when the input is only the Target or the Parent_Target. The results are using both Silver and pretraining on related tasks. We experiment with multiple runs using different random seeds and report the mean scores and their standard deviation.

weighted F1-score using each related task for pre-training when the input is Target and Parent_Target
Table 11: Hyperparameters used to fine-tune RoBERTa individually for each training setting. We accept default settings for the other hyperparameters as defined in the implementation by Pruksachatkun et al. (2020).

| Setting                  | Epochs | Batch size | Learning rate | Dropout |
|--------------------------|--------|------------|---------------|---------|
| Target                   | 5      | 16         | 1e-5          | 0.5     |
| + Silver                 | 2      | 16         | 1e-5          | 0.5     |
| + Related task           | 2      | 8          | 1e-5          | 0.5     |
| + Silver + Related task  | 4      | 16         | 1e-5          | 0.5     |

Table 10 presents the mean scores of Precision, Recall and weighted F1-score and their standard deviation when we use both Silver data and pretraining on related tasks with different random seeds. The results are consistent with the findings in our study: adding the Parent does improve the performance compared to the system that does not (0.63 vs. 0.60).

C Hyperparamters to Fine-tune the Systems

The neural model takes about half an hour on average to train on a single GPU of NVIDIA TITAN Xp. We use an implementation by Pruksachatkun et al. (2020) and fine-tune the RoBERTa (base architecture; 12 layers) (Liu et al., 2019) model for each of the four training settings. For each setting, we set the hyperparameters to be the same when the textual input is Target and Parent_Target respectively. Hence we only report tuned hyperparameters for each setting when the input is Target in Table 11.