Hate Speech and Counter Speech Detection: Context Does Matter

Anonymous ACL submission

Abstract

Hate speech is plaguing the cyberspace along with user-generated content. Adding counter speech has become an effective way to combat hate speech online. Existing datasets and models target either (a) hate speech or (b) hate and counter speech but disregard the context. This paper investigates the role of 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.

Parent	As an average height male, idgaf how tall you are, if that's your issue then spend the money and get a better seat, or just f**king make the seat selection online to get more space.
Target	Found the short guy!
-*Target* is Neutral if considering only *Target*. -*Target* is Hate if considering *Parent* and *Target*.	
Parent	I deal with women all day with my job and this is how they are - extremely stupid, hate-filled, bizarre and they appreciate nothing.
Target	Maybe you're an a**hole if they treat you like that?
-*Target* is Hate if considering only *Target*. -*Target* is Counter-hate if considering *Parent* and *Target*.	

Table 1: Reddit comments (*Target*s) deemed to be Hate, Neutral, or Counter-hate depending on whether one takes into account the previous comment (*Parent*).

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 of 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).

Two types of strategies have been implemented or studied to combat online hate: disruption and counter speech. Disruption refers to blocking hateful content or users temporally or permanently on a platform. To make the solution scalable, automated detection algorithms have been invented 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 to some extent, 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-repair mechanisms (Johnson et al., 2019). Counter speech refers to the "direct response that counters hate speech" (Mathew et al., 2019), which is considered a remedy to address hate speech. It has been supported by theoretical and empirical studies to be more effective in the long term (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 thus automatically generate counter speech against hate speech.

Most corpora with either hate speech (Hate) or counter speech (Counter-hate) annotations do not include the 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; Ziems et al., 2020). 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 toxicity 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[1] 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 people. 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 focus on 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 seperate tasks: showing annotators (a) only the *Target* or (b) the *Parent* and the *Target*. 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 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:[2] (a) a corpus with 6,846 pairs of (*Parent*, *Target*) Reddit comments and annotations indicating whether each *Target* is Hate, Neutral, or Counter-hate; (b) analysis of annotations showing that the problem requires taking into account context, as the ground truth changes otherwise; (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.

[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 anonymous_GitHub_link

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 like 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; Ziems et al., 2020; Garland et al., 2020). Mathew et al. (2019) collect and hand-code 6,898 counter hate comments from YouTube videos targeting Jews, Blacks and LGBT communities. Ziems et al. (2020) use a collection of hate and counter hate keywords relevant to COVID-19 and create a dataset containing 359 counter hate tweets targeting Asians. Garland et al. (2020) work with German tweets and define hate and counter speech based on the communities to which the authors belong. 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 *Target*s. Context does affect human judgments and has been taken into account for Hate detection (Gao and Huang, 2017; Vidgen et al., 2021; Pavlopoulos et al., 2020; Menini et al., 2021). Gao and Huang (2017) annotate hateful comments in the nested structures of 10 Fox News discussion threads. Vidgen et al. (2021) introduce a dataset of

Authors	Source	Size	Labels	Context?	Counter?
Waseem and Hovy (2016)	Twitter	1,607	Sexism/Racism/Normal	✗	✗
Davidson et al. (2017)	Twitter	24,783	Hate/Offense/Neither	✗	✗
Nobata et al. (2016)	Yahoo!	2,000	Hate/Derogatory/Profanity/Clean	✗	✗
Mathew et al. (2021)	Gab	1,1093	Hateful/Offensive/Normal	✗	✗
Gao and Huang (2017)	Fox News	1,528	Hateful/Non-hateful	preceding comment	✗
Qian et al. (2019)	Reddit	22,324	Hate/Non-hate	full conversation	✗
Pavlopoulos et al. (2020)	Wikipedia	20,000	Toxic/Non-toxic	preceding comment	✗
Menini et al. (2021)	Twitter	8,018	Abuse/Non-abuse	preceding comment	✗
Mathew et al. (2019)	YouTube	13,924	Counter/Non-counter	✗	✓
Ziems et al. (2020)	Twitter	2,400	Hate/Counter-hate/Neutral	✗	✓
Ours	Reddit	6,846	Hate/Counter-hate/Neutral	preceding comment	✓

Table 2: Comparison of corpora with hate and counter-hate annotations. We are the first to study the role of context (parent comment) in the annotation and detection of hate and counter-hate in social media conversations (Reddit).

Reddit comments with hate annotations taking into account context. 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). To our knowledge, we are the first to investigate the role of context in Hate and Counter-hate detection.

3 Dataset Collection and Annotation

We first describe our procedure to collect (*Parent*, *Target*) pairs, where both *Parents* and *Targets* are Reddit comments. 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 start with a set of 1,726 hate words from two lexicons: Hatebase[3] 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 ($\text{comment}_{\text{w/ hateword}}$). Second, we create a (*Parent*, *Target*) pair using $\text{comment}_{\text{w/ hateword}}$ as *Target* and its preceding comment as *Parent*. Third, we create a *(Parent, Target)* pair using $\text{comment}_{\text{w/ hateword}}$ as *Parent* and each of its replies as *Target*. 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 in order to keep the annotation costs reasonable while creating a (relatively) large corpus. We also collect the discussion title for each pair.

[3]http://hatebase.org/

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 call 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 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. Then, we use MACE (Hovy et al., 2013, Multi-Annotator Competence Estimation) and Krippendorff's α (Krippendorff, 2011). MACE is devised to rank annotators by their competence and recover adjudicate labels grounded on annotator's competence (not the majority label). Krippendorff's α estimates 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 $\alpha \geq 0.6$:

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 $\alpha < 0.6$.

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 $\alpha \geq 0.6$) and (b) Silver (2,095 remaining pairs). As we shall see, the Silver pairs are useful to learn models.

		Without *Parent*		
		Hate	Counter-hate	Neutral
With	Hate	57.4	8.4	34.2
	Counter-hate	18.7	26.2	55.1
	Neutral	9.7	8.1	82.2

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

Example	With	Without
Parent: That chick needs a high-five in the face with a chair. Damn her for making us look bad! *Target*: A brick is more effective.	Hate	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 4: Examples of *Target* comments whose labels change depending on whether annotators are shown the *Parent* of the *Target* comment (with and without).

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*, i.e., the context. Many *Target* comments that are perceived as Hate or Counter-hate become Neutral (34.2% and 55.1% respectively) when the *Parent* is provided. More surprisingly, many *Target* comments are perceived with the opposite label (from Hate to Counter-hate (8.4%) or from Counter-hate to Hate (18.7%)) when the *Parent* comments are shown.

We show examples of label changes in Table 4. In the first example, annotators identify the *Target* ("A brick is more effective.") as Neutral without seeing the *Parent*. In fact, a female is the target of hate in the *Parent*, and the author of *Target* replies with even more hatred (and the ground truth label is Hate). In the second example, the *Target* alone is insufficient to tell if it is Counter-hate. When annotators see the *Parent*, however, they understand

	Title		*Parent*		*Target*	
	p-value	Bonferroni	p-value	Bonferroni	p-value	Bonferroni
Textual factors						
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	↓↓↓	✓				

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). Tests that pass the Bonferroni correction are marked with a check mark.

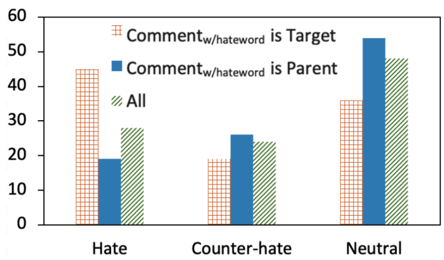

Figure 1: Label distribution in *Targets* depending on whether comment$_{w/\ hateword}$ is the *Parent* or the *Target*.

the author of *Target* counters the hateful content in the *Parent* by showing empathy towards the mail carrier. In the last example, the *Target* alone is considered Hate because it attacks someone by using the phrase "sick person". When the *Parent* is shown, however, the annotators understand the *Target* as calling out the *Parent* to be inappropriate.

Label distribution and linguistic insights Figure 1 shows the label distribution for all pairs (rightmost column in each block) and for pairs in which comment$_{w/\ hateword}$ (i.e., the comment containing at least one hate word) is the *Parent* or *Target*. The most frequent label assigned to *Target* comments is Neutral (49%) followed by Hate (28%) and Counter-hate (23%). While *Target* comments containing a hate word are likely to be Hate (45%), some are Counter-hate (19%) with context.

We analyze the linguistic characteristics of *Titles*, *Parents* and *Targets* when the *Targets* are Hate or 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[4] to count the profanity words. We analyze the components of linguistic features 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). As we are performing multiple hypothesis tests, we also report whether each feature passes the Bonferroni correction. We draw several interesting insights:

- Questions Marks in *Target* signal Counter-hate. We observe that people are inclined to use rhetorical questions as a way to counter hateful comments.
- 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 instead of themselves 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.

[4] https://github.com/RobertJGabriel/google-profanity-words-node-module/blob/master/lib/profanity.js

	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	0.00	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
+ Silver	0.58	0.55	0.57	0.44	0.42	0.43	0.69	0.72	0.70	0.60	0.61	0.61
+ Related task	0.56	0.55	0.56	0.51	0.41	0.45	0.68	0.74	0.71	0.61	0.61	0.61
+ Silver + Related task	0.55	0.56	0.56	0.49	0.53	0.51	0.67	0.69	0.70	0.61	0.61	0.61
Trained with Parent_Target	0.56	0.62	0.59	0.52	0.38	0.44	0.68	0.72	0.70	0.61	0.62	0.61
+ Silver†	0.58	0.57	0.57	0.49	0.51	0.50	**0.72**	**0.71**	**0.72**	0.63	0.63	0.63
+ Related task†	**0.55**	**0.66**	**0.60**	0.54	0.43	0.48	0.71	0.70	0.71	0.63	0.63	0.63
+ Silver + Related task‡	0.55	0.65	0.60	**0.54**	**0.52**	**0.53**	0.74	0.68	0.71	**0.64**	**0.64**	**0.64**

Table 6: Results obtained with several systems. We indicate statistical significance (McNemar's test (McNemar, 1947)) with respect to the model trained with the *Target* only using neither Silver nor pretraining on related tasks as follows: † indicates $p < 0.05$ and ‡ indicates $p < 0.01$. Training with the *Parent* and *Target* coupled with blending Silver annotations and pretraining with stance corpora yields the best results. The supplementary materials detail the results pretraining with all related tasks we consider.

5 Experiments and Results

We build neural network models to identify if a *Target* comment is Hate, Counter-hate, or Neutral. We 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 and Training 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 with 768 neurons and Tanh activation, and another fully connected layer with 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 report hyperparameters as well as other implementation details in the supplementary materials. We also experiment models that take the title of a 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 α 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 the supplementary materials provide detailed results pretraining with all the related tasks.

Blending Gold and Silver annotations requires tuning the α factor. We did so empirically using the training and validations splits, like any other hyperparameters. We found the optimal value to be 0.3 when blending Silver and 1.0 when utilizing both strategies.

As shown in Table 6, blending Gold and Silver annotations obtains better results by a small margin (Target: 0.61 vs. 0.58; Parent_Target: 0.63

Error Type	%	Example	Parent_Target	Target
Lack of information	48	*Parent*: Women can hover..? *Target*: No, they can't, but for some reason they keep trying and it gets sh*t everywhere.	Hate	Neutral
Negation	27	*Parent*: It's a joke you pu**y. *Target*: I don't see sexism as a joke, especially on a site dedicated to calling out sexism.	Counter-hate	Neutral
Sarcasm or irony	19	*Parent*: You must have been a real baller banging out those eighth graders as a High School senior. *Target*: Glad to see you have no rational argument left except childish jokes. We're done here pal.	Counter-hate	Hate
Hate without swear words	8	*Parent*: Name a dildo 'misogyny' so you can *literally* internalize it. *Target*: lol. Misogyny can already turn me on so that's a good idea.	Hate	Neutral

Table 7: Most common error types made by the *Target* only network (Target) that are fixed by the context-aware neural network (Parent_Target).

vs. 0.61). We also find that models pretrained for stance detection obtain better results than pretrained with other datasets. Pretraining with stance detection data benefits models trained without context (Target: 0.61 vs. 0.58) and models with context (Parent_Target: 0.63 vs. 0.61). These results indicate that these models have successfully transferred knowledge about stance between *Parent* and *Target* into the task of detecting whether the *Target* is Hate, Counter-Hate or Neutral.

From the results obtained when using neither of the two strategies, we observe: First, using the *Target* alone obtains much better results than the majority baseline (0.58 vs. 0.34). In other words, modeling the *Target* alone allows the network to identify *some* instances of Hate and Counter-hate despite the ground truth requires the *Parent*. Second, incorporating the *Parent* comment is beneficial (0.61 vs. 0.58). The difference is statistically significant when we in the meanwhile blend Silver or pretrain with related tasks (0.63 vs. 0.58).

Finally, the network pretrained with stance detection task first and then blending Silver in the training achieves the best performance (Parent_Target+Silver+Related task: 0.64). This result is statistically significant ($p < 0.01$) compared to the model trained with *Target* without blending Silver and pretraining with related tasks.

6 Qualitative Analysis

When is adding the context beneficial? When does our best model make mistakes? To investigate these questions, we perform a qualitative analysis. In particular, we answer the following questions:

- The errors made by the *Target* only network (Trained with Target) 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 manually analyze the errors made by the network using only the *Target* (Trained with 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 *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 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, 17% of errors fixed by the context-aware network include sarcasm or irony in the *Target* comment.

Error Type	%	Example	Ground Truth	Predicted
Negation	28	*Parent*: Those damn f**king white males, ruining it for everyone else. I'm going to a corner to process my guilt. *Target*: Don't forget male isn't a gender, it's a disease.	Hate	Counter-hate
Rhetorical question	27	*Parent*: Men are the ones that made inequality. *Target*: Do you get paid to be a dumba** in the internet?	Hate	Counter-hate
Hate without swear words	8	Parent: Circumcision is good for men. *Target*: Cut off the clitoris of women and cut of their breasts because of breast cancer then.	Hate	Neutral
Non-hate with swear words	8	*Parent*: <I wonder if feminists ever consider that? No. They are b**ches incapable of empathy. *Target*: This is the sh*t that gets screen capped and spread around to give this sub a bad name.	Counter-hate	Hate
Intricate text	7	*Parent*: Ah it's this again, f**king her and her cronies. *Target*: I have lost all respect for her.	Neutral	Hate

Table 8: Most common errors made by the best context-aware network (predictions by Parent_Target+Silver+Related task) compared to the ground truth.

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 also the reason for a substantial number of errors. In particular, 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 at all. 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 network considers the *Target* ("I have lost all respect for her.") to agree with the hateful *Parent*, thus it is predicted as Hate in the final example. Indeed, the author of *Target* expresses his/her attitude without vilifying others. Hence, the ground truth label is Neutral.

7 Conclusions and Future Work

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 have also shown that noisy instances (Silver data) and pretraining with relevant datasets can improve 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. As part of our future work, we plan to include broader context, such as all previous comments of a *Target*. Also, we observe a few counter hate replies in our dataset containing hate words. Our research agenda also includes investigating the effect of different types of counter speech and which type leads to the de-escalation of hate.

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.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

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.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

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 Muresan. 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.

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.

Yinhan 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 Rajgaria, 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 Olteanu, 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. DEBAGREEMENT: A comment-reply dataset for (dis)agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Rob Procter, Helena Webb, Pete Burnap, William Housley, Adam Edwards, Matthew L. Williams, and Marina Jirotka. 2019. A study of cyber hate on twitter with implications for social media governance strategies. In *Proceedings of the 2019 Truth and Trust Online Conference (TTO 2019), London, UK, October 4-5, 2019*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. 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 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive twitter sentiment classification using neural network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 215–221. AAAI Press.

Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 33–36, New York, NY, USA. Association for Computing Machinery.

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.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counterhate in social media during the COVID-19 crisis. *CoRR*, abs/2005.12423.

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

A Ethical Considerations

We use the PushShift API to collect data from Reddit[5]. Our collection is consistent with Reddit's Terms of Service. The data are accessed through the data dumps on Google's BigQuery using Python[6].

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 guidance indicates that in this situation explicit consent is not required from each user (Procter et al., 2019). We encrypt the users as User_A or User_B to avoid identification of users. In compliance with Reddit policy, we would like to make sure that our dataset will be reused for non-commercial research only[7].

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 Adult Content Qualification Test before working on our tasks. Annotators were compensated on average with 8 US$ per hour, we paid them whenever we accept their annotations or not. Annotators' IDs are not included in the dataset following the same principle to avoid profiling.

[5]https://pushshift.io/api-parameters/

[6]https://pushshift.io/ using-bigquery-with-reddit-data/

[7]https://www.reddit.com/wiki/api-terms

B Annotation Interface

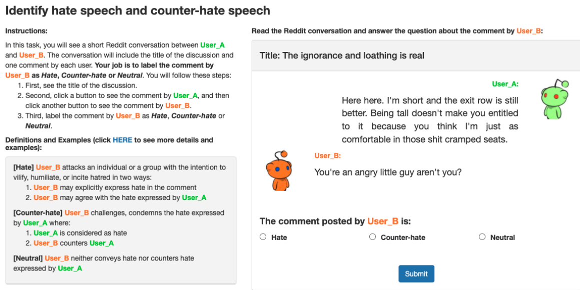

Figure 2: Screenshot of the annotation interface. The left panel displays the instructions and examples. The right panel displays the *Parent* and the *Target* to be annotated.

C Detailed Results

Table 9 presents detailed results complementing Table 6 in the paper. We provide Precision, Recall and weighted F1-score using each related task for pretraining when the input is Target and Parent_Target respectively in Table 9.

D Hyperparamters to Fine-tune the System for Each of the Training Settings

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 Phang et al. (2020) and fine-tune 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 10.

	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	0.00	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.

	Hp-1	Hp-2	Hp-3	Hp-4
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: Hyperparameters used to fine-tune RoBERTa individually for each training setting. Hp-1, Hp-2, Hp-3 and Hp-4 refer to the number of epochs, training batch size, learning rate and dropout used in the training procedure. We accept default settings for the other hyperparameters when we used the implementation by Phang et al. (2020).