# Pinpointing Fine-Grained Relationships between Hateful Tweets and Replies

**Abdullah Albanyan,[1] Eduardo Blanco[2]**

[1] University of North Texas
[2] Arizona State University
abdullahalbanyan@my.unt.edu, eduardo.blanco@asu.edu

## Abstract

Recent studies in the hate and counter hate domain have provided the grounds for investigating how to detect this pervasive content in social media. These studies mostly work with synthetic replies to hateful content written by annotators on demand rather than replies written by real users. We argue that working with naturally occurring replies to hateful content is key to study the problem. Building on this motivation, we create a corpus of 5,652 hateful tweets and replies. We analyze their fine-grained relationships by indicating whether the reply (a) is hate or counter hate speech, (b) provides a justification, (c) attacks the author of the tweet, and (d) adds additional hate. We also present linguistic insights into the language people use depending on these fine-grained relationships. Experimental results show improvements (a) taking into account the hateful tweet in addition to the reply and (b) pretraining with related tasks.


Figure 1: Hateful tweet (top) and two replies. The first reply disapproves of the hateful tweet and provides a justification. On the other hand, the second reply approves of the hateful tweet and adds additional hate.

## 1 Introduction

The rapid growth of social media platforms and the shield of anonymity have enabled online hate speech to proliferate. The Committee of Ministers of the Council of Europe defines hate speech as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin."[1] In recent surveys,[2] 83% of participants reported that they had encountered online hate speech, and LGBT youth, Muslims, immigrants, and women were the top 4 targets. 36.5% of participants felt personally threatened or offended by online hate speech, and 38.5% of those who encountered online hate speech reacted and replied to counter the hateful content.

These statistics along with other surveys (Fernandes et al. 2014) show the widespread presence of hate speech online. In an attempt to address this issue, the European Commission allied with Facebook, Microsoft, Twitter and YouTube to implement a "Code of conduct on countering illegal hate speech online" (European Commission 2019). Instagram, Snapchat and Dailymotion joined the alliance in 2018, Jeuxvideo.com in 2019, and TikTok in 2020. These companies are reportedly investing millions in manual moderation of hate speech yearly (Seetharaman 2018).

Identifying hate speech and blocking such content can be eased by automatic classifiers (Zampieri et al. 2020). Another strategy is to counter hate speech with new content in an attempt to redirect the conversation away from hate speech (Section 2), as counter hate replies can neutralize the effect of online hateful content and it is considered an effective alternative to blocking or removing such content (Gagliardone et al. 2015). In this paper, we investigate hate and counter hate speech in Twitter. We work with hateful tweets and replies posted by *real* Twitter users, and study the relationships between them beyond whether the reply counters the hateful tweet. Consider the example in Figure 1. The hateful tweet[3] uses irony to criticize Michelle Obama's physical appearance. The first reply counters the hateful content and provides an alternative opinion or justification. On the other hand, the second reply uses sarcasm to introduce additional hate (not being pretty vs. being a male). Note that (a) countering hate does not require a justification (e.g., *You are wrong. This comment is offensive*) and may include an



[1] https://rm.coe.int/1680505d5b

[2] https://rm.coe.int/1680700016

[3] Despite tweets are public content once published, we remove author information to preserve a degree of privacy.

attack towards the author of the hateful tweet (e.g., *You must be dumb or blind!*), and (b) agreeing with the hateful tweet need not introduce additional hate (e.g., *You nailed it!*).

Recent studies on hate and counter hate speech from a computational perspective (Section 2) primarily work with synthetic content (e.g., crowd workers or domain experts write counter hate messages on demand). In contrast, this study focuses on analyzing hateful tweets and their replies as written by real users. We believe that working with tweets from real users is crucial to better study this problem. The main contribution of this paper are:[4] (a) a corpus of 5,652 replies to hateful tweets published by real users and annotated with fine-grained relationship information (whether the reply counters the hate, provides a justification, attacks the author of the hateful tweet, or introduces additional hate); (b) linguistic analysis shedding light into the language used in the replies; (c) experimental results showing modest improvements considering both the hateful tweet and the reply as well as pretraining with related tasks; and (d) qualitative analysis describing when it is harder to perform any of the four classification tasks.

## 2 Previous Work

Hate speech in user-generated content has received substantial attention in recent years (Fortuna and Nunes 2018). The boundaries sometimes intersect with free speech (Howard 2019), and the very definition of hate speech is not agreed upon. Indeed, previous work targets hateful, toxic, abusive, and offensive language among others (Fortuna, Soler, and Wanner 2020; Vidgen et al. 2019).

Early work studied hate speech in user comments on Yahoo! Finance and News (Warner and Hirschberg 2012; Djuric et al. 2015; Nobata et al. 2016). Wikipedia conversations are another popular domain (Cécillon et al. 2020; Wulczyn, Thain, and Dixon 2017; Hua et al. 2018; Karan and Šnajder 2019), as well as Reddit (Qian et al. 2019; van Rosendaal, Caselli, and Nissim 2020) and Twitter (Jha and Mamidi 2017; Waseem 2016; Founta et al. 2018). In this paper, we work with Twitter for several reasons. First, Yahoo! disabled comments in 2020.[5] Second, unlike Reddit and Wikipedia users, Twitter users share their thoughts about a wide range of events ranging from mundane (e.g., having dinner, exercising, the weather) to world events (e.g., important elections, breaking news) in (almost) real time.

Several previous efforts on hate speech in Twitter approach the problem as a binary classification task (Waseem and Hovy 2016; Burnap and Williams 2015). More distinctions include identifying offensive but not hateful content (Davidson et al. 2017; Malmasi and Zampieri 2018) and determining if the target is a group or individual (Basile et al. 2019). The above works have shown that the presence of offensive words alone does not necessarily mean that the content is hateful—the context around them is important. Holgate et al. (2018) study swear words in Twitter and point out that they are often used to emphasize or express positive emotions. These previous efforts target hateful and offensive content at the tweet level. Unlike them, we study the relationships between hateful tweets and their replies.

Beyond single tweet classification, tweet popularity—likes, retweets, number of replies, etc.—has been studied using account information (Matsumoto et al. 2019; Fiok et al. 2020) and linguistic information (Wang, Chen, and Kan 2012). Hate and counter hate have also been studied beyond single tweets. Tekiroğlu, Chung, and Guerini (2020) argue for automatic generation and ranking of counter hate replies followed by manual validation. Garland et al. (2020) work with German tweets authored by members of self-reported hate and counter speech groups. Pavlopoulos et al. (2020) study the toxicity of replies in Wikipedia conversations taking and not taking into account the parent comment. Mathew et al. (2020) target hateful tweets containing the lexical patterns *I hate <target>* and their replies. Finally, Qian et al. (2019) crowdsource counter hate interventions and propose models to generate them. Their interventions are synthetic (i.e., written by crowdworkers on demand rather than social media users countering hate spontaneously) and as a result generic (e.g., *Use of this language is not tolerated and it is uncalled for*). Chung et al. (2019) provide a large-scale multilingual corpus with hate and counter hate messages. Yet both hate claims and counter hate interventions are also synthetic: they were written by experts. Our work complements these previous works. First, we consider tweets by any user as opposed to those by certain groups. Second, we work with tweets expressing hate and counter hate without imposing lexico-syntactic patterns. Third, we work with natural user-generated counter hate rather than synthetic. Finally, we go beyond identifying tweets expressing counter hate and also determine whether they include a justification, attack the author of the hateful tweet, or add additional hate.

## 3 A Corpus of Hateful Tweets and Replies

We create a new corpus of hateful tweets and replies annotated with fine-grained information beyond whether the reply counters the hateful tweet. The creation process builds upon corpora targeting hateful tweets. Unlike previous studies (Section 2), our corpus allows us to (a) quantify how often *real* replies counter hateful content and (b) identify what language real users use to do so. We argue this is more sound than working with synthetic interventions to counter hate from crowd workers or experts.

**Selecting Hateful Tweets and Replies** In order to select a sizable amount of hateful tweets and their replies, we followed two strategies. The first strategy consists in collecting all hateful tweets and replies from the only two corpora that include tweet identifiers according to Madukwe, Gao, and Xue (2020). Other corpora include the text in tweets but not identifiers, making it impossible to retrieve the original tweet and any replies using the Twitter API. Specifically, we considered the tweets labeled as *racist* or *sexist* by Waseem and Hovy (2016), and those labeled as *abusive* or *hateful* by Founta et al. (2018). This strategy only resulted in 652 (hateful tweet, reply) pairs since many tweets have been deleted by the authors and are no longer available.

[4] Corpus and implementation available at https://github.com/albanyan/hateful-tweets-replies

[5] https://bit.ly/2WJ5gz8

The second strategy allowed us to collect additional (hateful tweet, reply) pairs. Using the tweets labeled as *hate* or *offensive* in HateSpeech (Davidson et al. 2017), we query the Twitter search engine and retrieve 8,127 similar tweets with 72k replies. Then, we discard (hateful tweet, reply) pairs if:

1. The tweet is not labeled *hate* or *offensive* by the classifier by Davidson et al. (2017). The result is 6,213 hateful tweets with 50k replies.
2. The tweet does not share at least two tokens with the tweet used in the search engine. The result is 5,530 hateful tweets with 33k replies.
3. The reply is shorter than four tokens. The result is 3,755 hateful tweets with 18k replies.
4. The reply is a retweet. The result is 3,019 hateful tweets with 14k replies.

We designed these filters to maximize the likelihood that the result is hateful tweets (Filters 1 and 2; note that the search engine will always return some tweets even if the overlap is minimal) with meaningful replies (Filters 3 and 4). The last step in the selection process is to manually validate the result of the four filters. We did so until we identified 5,000 (hateful tweet, reply) pairs (310 hateful tweets with 5,000 replies) with the second strategy.

The size of our corpus combining both strategies is 5,652 (hateful tweet, reply) pairs. The number of replies per hateful tweet ranges from 1 to 167.

**Annotating Relationships between Hateful Tweets and Replies** In addition to identifying whether a reply to a hateful tweet counters the hate, we include finer-grained information. Our annotation process includes three steps.

The first step is to identify whether the reply is **counter hate**. We consider that the reply is counter hate if it disagrees with the hateful tweet explicitly or implicitly. For example, we consider counter hate calling into question the veracity of the content of the hateful tweet (e.g., I wonder where you are getting those facts from) or asking rhetorical questions (e.g., Are you the expert here?).

If the reply is counter hate, we ask two additional questions. First, we ask whether it provides a **justification**. We define a justification as any counterargument or reason that opposes the hateful content. For example, denying the content of the hateful tweet or generic replies (e.g., This kind of comment does not help the conversation) are not considered justifications. On the other hand, providing reasons why the hateful tweet is false or alternative scenarios are considered justifications (see examples in Figure 1 and Table 1). The second question to further characterize counter hate replies is whether the author of the reply **attacks the author** of the hateful tweet. We consider attacks in a broad sense, including making fun of, calling into question beliefs, or making derogatory comments regarding any protected class. Attacking the author of a hateful tweet is arguably a form of hate speech. We reserve for future work the effectiveness (or lack thereof) of countering hate speech with hate speech. The work presented here is limited to characterizing the kinds of counter hate observed in real social media communications.

If the reply does not counter the hateful tweet, we distinguish between tweets that simply agree with the hateful tweet and those that include **additional hate**. Agreeing with a hateful tweet alone is hate speech, but sharing additional hateful content is arguably worse.

| | |
|---|---|
| *Hateful Tweet 1*: this b**ch think she in I Am Legend LMAOOO <URL> | |
| *Reply*: @user This lady obviously has issues and instead of being calm and attempt to get her to stop, you antagonize and film her? This could have ended bad if she breached the window and it would all be on camera. | |
| Counter Hate? Yes | Justification? Yes |
| Attacks Author? No | Additional Hate? n/a |
| *Hateful Tweet 2*: on my way to f**k your b**ch. | |
| *Reply*: @user But my b**ch is your mamma | |
| Counter Hate? Yes | Justification? No |
| Attacks Author? Yes | Additional Hate? n/a |
| *Hateful Tweet 3*: Sad how people using Odell's name for views. These hoes need to get a job. Leave the man alone. | |
| *Reply*: @user I agree man, women are worthless and should just stay in the kitchen | |
| Counter Hate? No | Justification? n/a |
| Attacks Author? n/a | Additional Hate? Yes |

Table 1: Three examples of hateful tweets and replies from our corpus, and annotations after adjudicating disagreements. Annotations include four binary questions: whether the reply (a) is *counter hate*, (b) provides a *justification*, (c) *attacks the author* of the original (hateful) tweet, and (d) adds *additional hate*.

We believe answering these four questions can help make decisions towards limiting hate speech in social media. For example, replies to a hateful tweet that counter the hate with a justification are (presumably) more likely to be effective and could be highlighted. Additionally, hateful content that has been countered with a justification instead of an attack on the author of the hateful tweet may represent less potential harm. Likewise, a hateful tweet followed by several replies that not only do not counter the hate but include additional hate ought to be addressed in order to prevent harm associated with spreading hate speech.

**Examples** Table 1 provides real examples from our corpus. The first hateful tweet makes fun of a woman in distress (shown in the picture linked in the modified URL and explained in the reply). The reply counters the hateful content and provides a sound *justification*: all things considered, the reaction could be justified. The reply also calls into question the actions of the author of the hateful tweet (filming vs. helping), but the reply does not *attack the author*. For illustration purposes, here is a reply that is counter hate, does not provide a justification, and does not attack the author of the hateful tweet: *She seems fine to me. The world would be a better place if we were more kind to each other.*

Similar to the first example, the reply to the second tweet

| | Observed (%) | Cohen's $\kappa$ |
|---|---|---|
| Counter Hate? | 89.3 | 0.64 |
| Justification? | 88.9 | 0.71 |
| Attacks Author? | 87.7 | 0.75 |
| Additional Hate? | 89.3 | 0.66 |

Table 2: Inter-annotator agreements. We provide the observed agreements (percentage of answers annotators agreed on) and Cohen's $\kappa$. $\kappa$ coefficients between 0.6 and 0.8 are considered *substantial* agreement, and above 0.8 (nearly) perfect (Artstein and Poesio 2008).

uses sarcasm to counter the hateful content. This reply, however, does not provide a *justification*. Instead, the author of the reply attacks the author of the hateful tweet (anonymized Twitter handle @user) by turning the tables.

Finally, the third hateful tweet uses sexist and offensive language to diminish some women. In this case, the reply not only condones the hateful content, but introduces *additional hate* by making more derogatory and sexist comments. For illustration purposes, here is a reply that does not counter the hateful tweet either but does not introduce *additional hate*: *That's right, they need to get a job*.

**Annotation Process and Inter-Annotator Agreements** The questions and definitions above were refined iteratively after conducting pilot annotations. The annotation interface showed both the hateful tweet and reply, and guided the annotators to answer the questions. In order to avoid issues displaying tweets (e.g., special characters, emojis, links or pictures no longer available) the interface showed a screenshot of the tweets as displayed in the Twitter website.

Two graduate students who are active in social media platforms independently annotated the 5,652 (hateful tweet, reply) pairs. While social scientists could be a good choice as annotators, we believe regular social media users is a sound choice. Our rationale is that we are interested in the perceptions regular social media users have about hate and counter hate speech. Table 2 shows the inter-annotator agreements. The observed agreements are almost 90% across all questions, meaning that they disagree approximately in 1 out of 10 answers. The Cohen's $\kappa$ coefficients show that the annotation process resulted in *substantial* agreement; coefficients above 0.8 would indicate (nearly) perfect agreement (Artstein and Poesio 2008). We note that Cohen's $\kappa$ are higher when annotating whether a reply that is counter hate includes a *justification* or *attacks the author* of the hateful tweet. This is due to the fact that replies that include these two characteristics tend to do so explicitly. On the other hand, countering hate and providing additional hate often uses sarcasm and other implicit, nuanced language. After each annotator completed all the individual annotations, they discussed the instances in which they disagreed and adjudicated the final label.

## 4 Corpus Analysis

Table 3 presents the percentages of each label (no and yes) for each question. The majority of replies to hateful tweets

| | %No | %Yes |
|---|---|---|
| Counter Hate? | 80 | 20 |
| Justification? | 79 | 21 |
| Attacks Author? | 62 | 38 |
| Additional Hate? | 78 | 22 |

Table 3: Label percentages per question after adjudicating disagreements. The label distribution is biased towards *No* with all questions.

(80%) does not counter the hateful content, and it is somewhat rare for these replies to add additional hate (22%). Similarly, the replies that counter the hateful tweet rarely include a justification (21%). On the other hand, it is more common for them to include an attack to the author of the hateful tweet (38%). While our corpus has medium size (5,652 (hateful tweet, replies) pairs), this analysis provides evidence that user-generated counter hate speech in Twitter may not be well thought out. First, it rarely includes a justification. Second, it often includes an attack towards the author of the hateful content, which could be considered hate speech itself (despite one could argue that countering hate with more hate may be justified).

**Linguistic Insights** We analyze the replies in our corpus from a linguistic perspective in order to shed light into what kind of language people use to reply to hateful tweets (Table 4). The linguistic features we analyze build upon previous work and capture characteristics of the reply by itself or differences between the hateful tweet and the reply. We count the number of tokens with spaCy (Neumann et al. 2019) after removing URLs. We consider a token to be a misspelling if it does not appear in the Brown corpus (Francis and Kucera 1979), or lexicons of Twitter abbreviations[6] and bad words.[7] We collect a lexicon of negation cues from CD-SCO-Neg (Morante and Daelemans 2012), and use TextBlob[8] and Profanity[9] to calculate subjectivity and profanity scores. Finally, we use the lexicons by Mohammad and Turney (2013) to count positive and negative words, the lexicon by Kralj Novak et al. (2015) to get the sentiment polarity of emojis, and TextBlob to calculate sentiment polarity. These features are barely correlated with each other across all questions and labels: all inter-feature correlations are below 0.30 except a few involving the number of tokens in the replies. The supplementary materials detail the inter-feature correlations.

We check the predictive power of the above linguistic features using statistical tests (Table 4, t-test). We also indicate whether each features passes the Bonferroni correction. The p-values reveal several interesting insights:

- Longer replies, with more misspellings, or more subjective (a) are *counter hate* and include a *justification*, or (b) include *additional hate* if they are not counter hate.

[6]https://github.com/vivekanand1101/witter-sentiment-analysis
[7]https://www.cs.cmu.edu/ biglou/resources/bad-words.txt
[8]https://github.com/sloria/TextBlob
[9]https://github.com/ben174/profanity

| | Counter Hate? | | Justification? | | Attacks Author? | | Additional Hate? | |
|---|---|---|---|---|---|---|---|---|
| | p-value | Bonf. | p-value | Bonf. | p-value | Bonf. | p-value | Bonf. |
| Number of tokens | ↑↑↑ | ✓ | ↑↑↑ | ✓ | | | ↑↑↑ | ✓ |
| Count of misspelled words | ↑↑↑ | ✓ | ↑↑↑ | ✓ | | | ↑↑↑ | ✓ |
| Count of pronoun *you* | ↑↑↑ | ✓ | ↑↑↑ | ✓ | ↑↑↑ | ✓ | | |
| Count of negation cues | ↑↑↑ | ✓ | ↑↑↑ | ✓ | ↓↓↓ | ✓ | | |
| Subjectivity score | ↑↑↑ | ✓ | ↑↑↑ | ✓ | | | ↑↑↑ | ✓ |
| Profanity wrt. hateful tweet | ↓↓↓ | ✓ | | | ↓↓↓ | ✓ | ↓↓↓ | ✓ |
| Sentiment features | | | | | | | | |
| Count of positive words | ↑ | ✗ | ↑↑↑ | ✓ | ↓↓ | ✗ | ↑ | ✗ |
| Count of negative words | ↑↑↑ | ✓ | ↑↑↑ | ✓ | ↑ | ✗ | ↑↑↑ | ✓ |
| Emojis Polarity | ↓↓↓ | ✓ | | | | | | |
| Polarity wrt. hateful tweet | ↓↓↓ | ✓ | | | | | ↓↓↓ | ✓ |

Table 4: Linguistic analysis comparing the labels assigned to hateful tweets and replies. Number of arrows indicate the p-value (t-test; one: $p < 0.05$, two: $p < 0.01$, and three: $p < 0.001$), and arrow direction indicates whether higher values correlate with *yes* (up) or *no* (down). We also indicate whether the test passes the Bonferroni correction.

- The pronoun 'you' signals that the reply is *counter hate*, includes a *justification* and *attacks the author* of the tweet. This is unsurprising since a natural way to do so is to refer to the author with 'you' in the reply.
- Negation is used to express *counter hate* and provide *justifications*. On the other hand, people use negations when they do *not attack the author* of hateful tweets.
- Low profanity in the reply relative to the hateful tweet indicates that replies (a) are *not counter hate* and do *not* include *additional hate*, or (b) do *not attack the author* of the hateful tweet if they are *counter hate*.
- Regarding positive and negative words, we observe that (a) negative words are more indicative of *counter hate*, (b) both positive and negative words are used when the reply includes a *justification*, (c) unsurprisingly, positive words are rare when the reply *attacks the author* of the hateful tweet, and (d) negative words indicate the presence of *additional hate*.
- If emojis in the reply have positive sentiment, the reply tends to *not be counter hate*, indicating that sarcasm is common when replies agree with the hateful tweet.
- Replies to hateful tweets that are more negative than the hateful tweets tend to not be *counter hate* and *not* include *additional hate*.

## 5 Experiments and Results

We cast the problem of determining the relationship between a hateful tweet and a reply as a classification task. More specifically, we build a binary classifier for each question: whether the reply is counter hate, includes a justification, attacks the author of the hateful tweet, or includes additional hate. A (hateful tweet, reply) pair becomes an instance, and we split the dataset as follows: 70% for training, 10% for validation, and 20% for testing.

**Baselines** We experiment with two simple baselines: predicting *random* labels (*yes* or *no*) and the *majority* label. Recall that the majority label is *no* for all questions (Table 3).

**Neural Network Architecture and Training** We experiment with neural classifiers build on top of BERT-based transformers. First, we use the original BERT transformer (Devlin et al. 2019), which is pretrained with 800M words from the BooksCorpus (Zhu et al. 2015) and English Wikipedia (2,500M words). Second, we use BERTweet (Nguyen, Vu, and Tuan Nguyen 2020), a BERT model pretrained using the RoBERTa approach (Liu et al. 2019) with 850M English tweets and 5M tweets related to the COVID-19 pandemic.

The neural architecture consists of a pretrained transformer (BERT or BERTweet), a fully connected layer with 128 neurons and ReLU activation, and another fully connected layer with 2 neurons and softmax activation which outputs the prediction (*no* or *yes*). In order to explore the roles of the hateful tweet and reply in determining their relationship, we consider three textual inputs:

- the hateful tweet alone,
- the reply alone, and
- the hateful tweet and the reply.

In order to feed to the network the hateful tweet and the reply, we concatenate them with the '[SEP]' special token. As we shall see, models only taking as their input the hateful tweet outperform the baselines, meaning that some hateful tweets are more likely to elicit certain replies.

In order to implement the models, we use the Transformers library by HuggingFace (Wolf et al. 2020), TensorFlow (Abadi et al. 2016), and PyTorch (Paszke et al. 2019). We report hyperparameters and other implementation details in the supplementary materials.

### Quantitative Results

We present results (F1-measure for labels *no* and *yes*, and the weighted average) in Table 5. The supplementary materials present detailed results including Precision and Recall for all labels and models listed in Table 5.

The top block in Table 5 presents the results with the baselines. Since our dataset is biased (Table 3), the majority baselines obtain relatively high F1s (averages: 0.70, 0.66, 0.47,

| | Counter Hate? | | | Justification? | | | Attacks Author? | | | Additional Hate? | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | No | Yes | Avg. | No | Yes | Avg. | No | Yes | Avg. | No | Yes | Avg. |
| Baselines | | | | | | | | | | | | |
| Random | 0.63 | 0.29 | 0.55 | 0.60 | 0.30 | 0.53 | 0.54 | 0.38 | 0.48 | 0.61 | 0.31 | 0.54 |
| Majority | 0.88 | 0.00 | 0.70 | 0.87 | 0.00 | 0.66 | 0.76 | 0.00 | 0.47 | 0.87 | 0.00 | 0.67 |
| BERT trained with ... | | | | | | | | | | | | |
| hateful tweet | 0.87 | 0.25 | 0.74 | 0.81 | 0.39 | 0.71 | 0.70 | 0.50 | 0.63 | 0.87 | 0.05 | 0.68 |
| reply | 0.89 | 0.45 | 0.80∗ | 0.91 | 0.56 | 0.82∗ | 0.75 | 0.62 | 0.70 | 0.91 | 0.69 | 0.86∗ |
| hateful tweet + reply | 0.89 | 0.50 | 0.81∗ | 0.91 | 0.60 | 0.83∗ | 0.76 | 0.63 | 0.71∗ | 0.89 | 0.67 | 0.84∗ |
| pretrained with ... | | | | | | | | | | | | |
| best individual task | 0.90 | 0.54 | 0.83∗†‡ | 0.92 | 0.65 | 0.85∗ | 0.81 | 0.68 | 0.76∗ | 0.92 | 0.72 | 0.88∗‡ |
| all beneficial tasks | 0.88 | 0.57 | 0.82∗ | 0.90 | 0.61 | 0.83∗ | 0.83 | 0.66 | 0.76∗†‡ | 0.92 | 0.69 | 0.87∗‡ |
| BERTweet trained with ... | | | | | | | | | | | | |
| hateful tweet | 0.88 | 0.19 | 0.73 | 0.82 | 0.38 | 0.72 | 0.73 | 0.49 | 0.64 | 0.87 | 0.02 | 0.67 |
| reply | 0.88 | 0.57 | 0.82∗ | 0.91 | 0.68 | 0.86∗ | 0.80 | 0.66 | 0.75∗ | 0.93 | 0.79 | 0.90∗ |
| hateful tweet + reply | 0.90 | 0.58 | 0.83∗ | 0.90 | 0.68 | 0.85∗ | 0.83 | 0.67 | 0.77∗ | 0.93 | 0.78 | 0.89∗ |
| pretrained with ... | | | | | | | | | | | | |
| best individual task | 0.89 | 0.60 | 0.83∗ | 0.93 | 0.74 | 0.88∗‡ | 0.84 | 0.73 | 0.80∗ | 0.94 | 0.80 | 0.91∗ |
| all beneficial tasks | 0.90 | 0.57 | 0.83∗† | 0.92 | 0.72 | 0.87∗ | 0.83 | 0.69 | 0.78∗ | 0.93 | 0.78 | 0.89∗ |

Table 5: Results obtained with several systems (F1-measures; *Avg.* refers to the *weighted average*). The supplementary materials provide detailed results per label and subtask (P, R and F1). We indicate statistical significance (McNemar's test (McNemar 1947) with $p < 0.05$) as follows: $*$ indicates that a model obtains statistically significant results with respect to the *hateful tweet* model, † with respect to the *reply* model, and ‡ with respect to the *hateful tweet + reply* model.

and 0.67 for each question) despite they always predict *no*.

The results with models built upon BERT and BERTweet using different inputs (the hateful tweet, the reply, or both the hateful tweet and the reply) are more interesting:

- First, we note that BERTweet, as expected, yields better results than BERT as it is pretrained in the same domain we work with.
- Second, the hateful tweet alone outperforms the majority baseline by a significant margin (*counter hate*: 0.74 (BERT) and 0.73 (BERTweet) vs. 0.70, *justification*: 0.71 and 0.72 vs 0.66, *attacks author*: 0.63 and 0.64 vs 0.47, and *additional hate*: 0.68 and 0.67 vs 0.67). This empirical finding leads to the conclusion that some (hateful) tweets are more likely to elicit certain types of replies. For example, taking into account only the hateful tweet allows the neural network to improve results 36.2% with respect to the majority baseline predicting whether an (unknown) reply will attack the author (0.47 vs. 0.64).
- Third, feeding to the network both the hateful tweet and the reply obtains the best results by a small margin.

**Pretraining with Complementary Tasks** There are multiple tasks and corpora that are related to determining the relationship between a hateful tweet and a reply. Intuitively, replies that attack the author of a hateful tweet or contain additional hate are likely to have negative sentiment. Similarly, expressing certain emotions (e.g., anger, sadness, desperation) may signal counter hate.

In order to explore whether pretraining with related tasks is actually beneficial, we experiment with several corpora publicly available. Specifically, we pretrain our best models (*BERT* and *BERTweet* trained with the *hateful tweet + reply*) with existing corpora annotating:

- hateful comments: hateful or not hateful (Basile et al. 2019), and hate speech, offensive, or neither (Davidson et al. 2017);
- offensive language: offensive or not offensive (Zampieri et al. 2019);
- the function of vulgar words: aggression, emotion, emphasis, auxiliary, group identity, nonvulgar (Holgate et al. 2018);
- emotions: anger, joy, optimism or sadness (Mohammad et al. 2016);
- irony: tweets using and not using irony (Van Hee, Lefever, and Hoste 2018);
- sentiment: negative, neutral, or positive (Rosenthal, Farra, and Nakov 2017); and
- stance regarding abortion, atheism, climate change, feminism, and Hillary Clinton: in favor, against or none (Mohammad et al. 2016).

Table 5 also presents results pretraining with (a) the best *individual* tasks described above, and (b) all *tasks that are beneficial* individually. The best individual tasks are:

- *emotion* to determine whether the reply counters the hateful tweet,
- *stance regarding abortion* to determine whether the reply provides a justification,
- *emotion* to determine whether the reply attacks the author of the hateful tweet, and
- *offensive language* to determine whether the reply contains additional hate.

We observe that pretraining benefits the neural network overall with both BERT and BERTweet trained with the hateful tweet and reply (0.83 vs. 0.81, 0.85 vs. 0.83, 0.76 vs. 0.71, and 0.88 vs. 0.84 with BERT; and 0.83 vs. 0.83, 0.88

| | Counter Hate? | Justification? | Attacks Author? | Additional Hate? |
|---|---|---|---|---|
| Intricate text | | | | |
| Sarcasm, irony, implicit meaning | 32 | 17 | 19 | 22 |
| Mentions many users, unclear whom the author refers | 5 | 13 | 7 | 7 |
| Mentions many named entities | 1 | 0 | 6 | 5 |
| All | 38 | 30 | 33 | 34 |
| General knowledge | 16 | 29 | 9 | 15 |
| Lack of information | | | | |
| A picture or video is key, text alone insufficient | 12 | 13 | 3 | 4 |
| The text is very short, less than 5 tokens | 4 | 0 | 10 | 7 |
| All | 16 | 13 | 13 | 11 |
| Misspellings | 13 | 7 | 12 | 15 |

Table 6: Major error types made by the best performing model in each task (as shown in Table 5). An instance that a model misclassifies may fall under two error types. All the numbers are percentages.

vs. 0.85, 0.80 vs. 0.77, and 0.91 vs. 0.89 with BERTweet). In other words, the benefits of pretraining are not about adapting to Twitter data—BERTweet is already pretrained with millions of tweets—but about transferring knowledge from these related tasks. Finally, there is no benefit of pretraining with all tasks that individually are beneficial.

## Error Analysis

We manually analyzed a sample of the wrong predictions made by the best model in each task. The sample included 100 random (hateful tweet, reply) pairs for the *counter hate* task, 100 random (hateful tweet, reply) pairs for the *additional hate* task, 30 (hateful tweet, reply) pairs for the *justification* task, and 67 (hateful tweet, reply) pairs for the *attacks the author* task. The sample included all the wrong predictions in the test set for the *justification* and *attacks the author* tasks as there are less than 100. Table 6 shows the most common error types in each task. In our analysis, an instance may belong to several error types, e.g., a tweet may use sarcasm (intricate text) and have misspellings.

While there are some differences, the percentages show a similar trend across all four tasks. The most common errors are intricate text (38%, 30%, 33%, and 34%), including using sarcasm and irony, and mentioning several Twitter user handles or named entities. We include under *general knowledge* (16%, 29%, 9%, and 15%) commonsense and world knowledge such as (potentially offensive) comparisons with celebrities and references to height, weight, or gender. The model also faces challenges when the reply includes a picture or other media (12%, 13%, 3%, and 4%). Finally, misspellings and the often informal language of Twitter is also often present in the replies the model struggles with (13%, 7%, 12%, and 15%). Here are some examples:

- Intricate text. Hateful tweet: *Dear everyone using "natural" deodorant, it's not working, you stinky bi**h.*
  Reply: *@USER What you eat is what makes you stink.*
  - Counter Hate? Predicted: No, Gold: Yes
- General knowledge. Hateful tweet: *This b**ch still talking when I already told her to shut the f**k up..*
  Reply: *@USER Ivanka was raised with class so those raised in a barn are just background noise.*
  - Justification? Predicted: No, Gold: Yes
- Lack of information. Hateful tweet: *i'm on a boat, b**ch [picture of selfie riding a boat].*
  Reply: *@USER Took me a couple replays to realize he was looking at the camera well I think he is.*
  - Attacks Author? Predicted: No, Gold: Yes
- Misspellings. Hateful tweet: *Ni***s cheat on the most loyal beautiful women with fu**ing trash.*
  Reply: *@USER u beautiful f**k em.*
  - Additional Hate? Predicted: No, Gold: Yes

# 6 Conclusions

Hate speech in Twitter and other social platforms is common. Users often do not face consequences for their hateful messages. Social media platforms invest substantial resources to combat hate speech. A common strategy is to tag or block hateful content (and ban repeated offenders).

In this paper, we study the relationship between hateful tweets and replies. Crucially, we investigate (a) how *real* users react to hateful messages (as opposed to expert moderators or other synthetic interventions), and (b) how often they counter the hateful message. We have presented a corpus of 5,652 (hateful tweet, reply) pairs. Our analysis does not paint a rosy picture: only 20% of replies counter the hateful tweet, although only 22% of replies that agree with the hateful tweet contain additional hate. In other words, most replies simply agree with the hateful tweet. When people counter the hateful tweet, they rarely provide a justification (21%) and often attack the author of the hateful tweet (38%). Our experimental results show that identifying the four fine-grained aspects of the relationship between a hateful tweet and a reply can be automated. In particular, pretraining with related tasks is useful to identify justifications and whether the reply attacks the author of the hateful tweet.

We believe that the fine-grained relationships between hateful tweets and replies we work with could help in combating hate speech in social media. For example, a reply that counters a hateful message and provides a justification could be sufficient to stop spreading the hateful content. On the other hand, replies that include additional hate or counter the hateful tweet by attacking the author may need attention.

## References


Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Artstein, R.; and Poesio, M. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4): 555–596.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63.

Burnap, P.; and Williams, M. L. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, 7(2): 223–242.

Cécillon, N.; Labatut, V.; Dufour, R.; and Linarès, G. 2020. WAC: A Corpus of Wikipedia Conversations for Online Abuse Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ISBN 979-10-95546-34-4.

Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, S. S.; and Guerini, M. 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*, 2819–2829. Florence, Italy: Association for Computational Linguistics.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1).

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; and Bhamidipati, N. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, 29–30. ISBN 9781450334730.

European Commission. 2019. The EU Code of Conduct on Countering Illegal Hate Speech Online. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en. Accessed: 2021-5-12.

Fernandes, C. M.; Dolenec, M.; Boche, T.; and Silva, V. 2014. Final Report on Online Hate Speech. *ELSA International*.

Fiok, K.; Karwowski, W.; Gutierrez, E.; and Ahram, T. 2020. Predicting the Volume of Response to Tweets Posted by a Single Twitter Account. *Symmetry*, 12(6): 1054.

Fortuna, P.; and Nunes, S. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4): 1–30.

Fortuna, P.; Soler, J.; and Wanner, L. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 6786–6794. ISBN 979-10-95546-34-4.

Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Francis, W. N.; and Kucera, H. 1979. Brown Corpus Manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Gagliardone, I.; Gal, D.; Alves, T.; and Martinez, G. 2015. *Countering online hate speech*. Unesco Publishing.

Garland, J.; Ghazi-Zahedi, K.; Young, J.-G.; Hébert-Dufresne, L.; and Galesic, M. 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*.

Holgate, E.; Cachola, I.; Preoţiuc-Pietro, D.; and Li, J. J. 2018. Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4405–4414.

Howard, J. W. 2019. Free speech and hate speech. *Annual Review of Political Science*, 22: 93–109.

Hua, Y.; Danescu-Niculescu-Mizil, C.; Taraborelli, D.; Thain, N.; Sorensen, J.; and Dixon, L. 2018. WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Jha, A.; and Mamidi, R. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, 7–16.

Karan, M.; and Šnajder, J. 2019. Preemptive Toxic Language Detection in Wikipedia Comments Using Thread-Level Context. In *Proceedings of the Third Workshop on Abusive Language Online*, 129–134.

Kralj Novak, P.; Smailović, J.; Sluban, B.; and Mozetič, I. 2015. Sentiment of Emojis. *PLOS ONE*, 10(12): e0144296.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Madukwe, K.; Gao, X.; and Xue, B. 2020. In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 150–161.

Malmasi, S.; and Zampieri, M. 2018. Challenges in Discriminating Profanity from Hate Speech. *CoRR*, abs/1803.05495: 187–202.

Mathew, B.; Kumar, N.; Goyal, P.; and Mukherjee, A. 2020. Interaction Dynamics between Hate and Counter Users on Twitter. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, 116–124. ISBN 9781450377386.

Matsumoto, K.; Hada, Y.; Yoshida, M.; and Kita, K. 2019. Analysis of Reply-Tweets for Buzz Tweet Detection. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), Hakodate, Japan*, 13–15.

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

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

Mohammad, S. M.; and Turney, P. D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3): 436–465.

Morante, R.; and Daelemans, W. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 1563–1568.

Neumann, M.; King, D.; Beltagy, I.; and Ammar, W. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*.

Nguyen, D. Q.; Vu, T.; and Tuan Nguyen, A. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 145–153. ISBN 9781450341431.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8026–8037. Curran Associates, Inc.

Pavlopoulos, J.; Sorensen, J.; Dixon, L.; Thain, N.; and Androutsopoulos, I. 2020. Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Qian, J.; Bethke, A.; Liu, Y.; Belding, E.; and Wang, W. Y. 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)*.

Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Seetharaman, D. 2018. Facebook Throws More Money at Wiping Out Hate Speech and Bad Actors . *The Wall Street Journal*.

Tekiroğlu, S. S.; Chung, Y.-L.; and Guerini, M. 2020. Generating Counter Narratives against Online Hate Speech: Data and Strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Van Hee, C.; Lefever, E.; and Hoste, V. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 39–50.

van Rosendaal, J.; Caselli, T.; and Nissim, M. 2020. Lower Bias, Higher Density Abusive Language Datasets: A Recipe. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, 14–19. ISBN 979-10-95546-49-8.

Vidgen, B.; Harris, A.; Nguyen, D.; Tromble, R.; Hale, S.; and Margetts, H. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, 80–93.

Wang, A.; Chen, T.; and Kan, M.-Y. 2012. Re-tweeting from a linguistic perspective. In *Proceedings of the Second Workshop on Language in Social Media*, 46–55.

Warner, W.; and Hirschberg, J. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, 19–26.

Waseem, Z. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142.

Waseem, Z.; and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, 88–93.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1391–1399. ISBN 9781450349130.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Zampieri, M.; Nakov, P.; Rosenthal, S.; Atanasova, P.; Karadzhov, G.; Mubarak, H.; Derczynski, L.; Pitenis, Z.; and Çöltekin, Ç. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*.