Not All Counterhate Tweets Elicit the Same Replies: A Fine-Grained Analysis

Anonymous EMNLP submission

Abstract

Counterhate arguments can effectively fight and limit the spread of hate speech. However, they may worsen the situation, as some people may aggressively respond if they believe they are being threatened or targeted. In this paper, we investigate replies to counterhate tweets beyond whether the reply agrees or disagrees with the counterhate tweet. We present a corpus with 2,621 (hateful tweet, counterhate tweet, reply) triples annotated with fine-grained characteristics. Our corpus shows that almost half of the replies to the counterhate tweets do not agree with the counterhate tweet (51%), and it is more common for them to support the hateful tweet when they do not agree. We also analyze the language of the counterhate that leads to certain types of replies. Experimental results show that it is feasible to anticipate the kind of replies a counterhate tweet will elicit.

1 Introduction

Hate messages and offensive language are commonplace in social media platforms. Twitter reported that more than 1.1M accounts spread hateful content in the second half of 2020, a 77% increase with respect to the first half of the same year.[1] In a recent survey of 10,093 adults in the U.S., 41% of participants reported online harassment on a personal level, and almost two-thirds of adults under the age of 30 reported experiencing internet harassment (Vogels, 2021). These figures, alongside other surveys,[2,3] demonstrate the prevalence of hate speech on the internet. To address this problem, the European Commission partnered with popular social media platforms to announce a "Code of conduct on countering illegal hate speech online" (European Commission, 2019), which contains several commitments to prevent the spread of online hate speech across Europe.

The enormous amount of daily data makes these platforms rely on users who manually flag hateful content (Crawford and Gillespie, 2016). This results in spending millions of dollars yearly on manual hate speech verification and moderation (Seetharaman, 2018). An alternative is to automatically fight hate speech by using hate speech classifiers (Section 2). However, removing users' content—even if it is helpful—restricts free speech. According to the Pew Research Center (Duggan, 2017), "Despite this broad concern over online harassment, 45% of Americans say it is more important to let people speak their minds freely online, and 53% feel that it is more important for people to feel welcome and safe online."

A complementary strategy to address hateful content that does not interfere with free speech is to counter the hate with new content to divert the discourse away from the hate. Counterhate tweets can effectively fight and limit the spread of hate speech without removing or blocking the content (Gagliardone et al., 2015; Schieb and Preuss, 2016). These replies are positive arguments that are used to offer non-negative feedback to oppose hate speech with logic and facts. However, counterhate replies may worsen the situation, as some people may aggressively respond if they believe they are being threatened or targeted (Rains, 2013; Clayton et al., 2019).

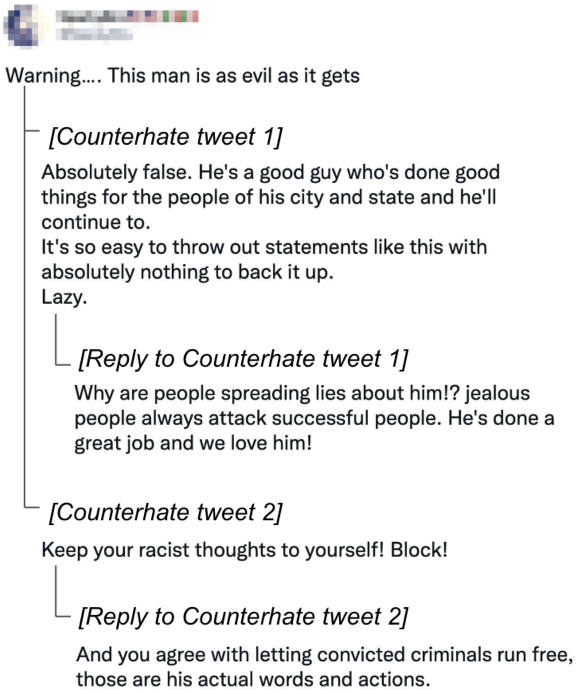
Upon these motivations, this study aims to identify the kind of replies counterhate tweets elicit. Specifically, we investigate replies to counterhate tweets beyond whether the reply agrees or disagrees with the counterhate tweet. We consider three-level Twitter threads: a hateful tweet, a counterhate tweet, and all the replies to the counterhate tweet. We define a hateful tweet as any tweet that contains abusive language directed to individuals or groups of people. On the other hand, a counterhate tweet is a response tweet that explicitly or implicitly disagrees with the hateful tweet. A reply

Use another term "three-level Twitter threads"

[1] https://time.com/6080324/twitter-hate-speech-penalties/

[2] https://legalresearch.elsa.org/library/ohs/

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



[redacted]

Warning.... This man is as evil as it gets

[Counterhate tweet 1]

Absolutely false. He's a good guy who's done good things for the people of his city and state and he'll continue to.
It's so easy to throw out statements like this with absolutely nothing to back it up.
Lazy.

[Reply to Counterhate tweet 1]

Why are people spreading lies about him!? jealous people always attack successful people. He's done a great job and we love him!

[Counterhate tweet 2]

Keep your racist thoughts to yourself! Block!

[Reply to Counterhate tweet 2]

And you agree with letting convicted criminals run free, those are his actual words and actions.

Figure 1: Twitter thread originating with a hateful tweet. This paper investigates the replies to counterhate tweets. In the first example, the reply not only agrees with the counterhate tweet, but also adds additional counterhate. On the other hand, the second reply not only disagrees with the counterhate tweet, but also shows support for the hateful tweet.

is any response to the counterhate tweet. Consider the example in Figure 1. The hateful tweet contains hateful content towards a man (shown in a picture in the real tweet). The reply to the first counterhate tweet not only agrees with the counterhate tweet, but also includes additional counterhate (i.e., *he's done a great job*). Conversely, the reply to the second counterhate tweet not only disagrees with the counterhate tweet, but also includes an opinion supporting the hateful tweet (i.e., *letting convicted criminals run free*). While the author of the second counterhate tweet may have had good intentions, the tweet elicited more hate and made the discourse undesirable. This paper presents a fine-grained characterization of replies to counterhate tweets and opens the door for forecasting which counterhate tweets may elicit more hate instead of stopping it.

In summary, the main contributions of this paper are:[4] (a) a corpus with 2,621 (hateful tweet, counterhate tweet, reply) triples annotated with fine-grained characteristics (whether the reply agrees with the counterhate tweet, supports the hateful tweet, attacks the author of the counterhate tweet, or adds additional counterhate); (b) linguistic analysis of the counterhate tweets depending on our fine-grained characterization of the replies they elicit; (c) experimental results showing it is feasible to anticipate the kind of replies a counterhate tweet will elicit, and modest improvements when applying data augmentation or blending with related datasets; and (d) qualitative analysis revealing when it is harder to perform any of the four classification tasks.

2 Previous Work

Recently, considerable literature has grown around identifying hateful content in user-generated content (Fortuna and Nunes, 2018). Existing research has created a variety of datasets to detect hate speech from several sources, including Twitter (Waseem and Hovy, 2016; Davidson et al., 2017), Reddit (Qian et al., 2019), Fox News (Gao and Huang, 2017), Yahoo! (Nobata et al., 2016; Djuric et al., 2015), and Gab (Mathew et al., 2021). Other studies have worked on identifying the target of hate, including whether the hateful content was directed toward a group, a person, or an object (Basile et al., 2019; Zampieri et al., 2019a; Ousidhoum et al., 2019).

Previous efforts have also attempted to detect and generate counterhate content. For counterhate detection, Garland et al. (2020) work with hateful and counterhate German tweets published from two well-known groups. Mathew et al. (2020) collect and analyze pairs of hateful tweets and replies using the hate speech template *I hate <group>*, and detect whether a reply to a hateful tweet is a counterhate reply or not. In addition to analyzing or detecting counterhate replies, Albanyan and Blanco (2022) identify four fine-grained aspects of the relationship between a hateful tweet and a reply (e.g., whether the reply counters the hateful tweet with justification). For counterhate generation, some studies have worked on collecting datasets with the help of crowd workers (Qian et al., 2019) or trained operators (Fanton et al., 2021; Chung et al., 2019).

There are several attempts to predict if a content will lead to additional hateful content. Zhang et al. (2018) identify whether a reply will result in a personal attack. Liu et al. (2018) predict the number of hateful comments that an instgram post would

[4] Corpus and implementation available at anonymous.link

receive. On the other hand, there are few efforts on investigating the impact of counterhate content, as stated in a recent survey by Alsagheer et al. (2022). Mathew et al. (2019) analyze YouTube comments and found that counterhate comments received more likes and interactions than non-counterhate comments. Other studies found that there is a positive association between counterhate efficiency and both its author's ethnicity (Munger, 2017) and how immediate the response to the hateful content is posted (Schieb and Preuss, 2018). Finally, Garland et al. (2022) analyze hateful and counterhate German tweets and find that organized counterhate tweets elicit more counterhate replies and decrease the severity of the hate speech. Unlike the previous studies, we (a) cover broader types of tweets by applying three approaches in the data collection process, (b) consider three-level Twitter threads (hateful tweet, counterhate tweet, and reply) which preserves the conversational nature of the discussion, and (c) go beyond whether the reply agrees and disagrees with the counterhate tweet.

The list here needs to be rewritten

3 Dataset Collection and Annotation

We start our study by collecting triples consisting of hateful tweets, counterhate tweets, and replies. Then, we annotate the triples with our fine-grained characterization of the replies. Unlike previous works (Section 2), our corpus enables us to (a) investigate whether counterhate tweets are successful at stopping the hate (Section 4), (b) analyze the language people use in counterhate tweets depending on the replies they elicit (Section 4), and (c) predict the kind of replies a counterhate tweet will elicit (Section 5).

Collecting Hateful Tweets, Counterhate Tweets, and Replies We use three strategies to collect a sufficient number of hateful tweets, counterhate tweets, and replies. First, we collect replies from corpora that provide (hateful tweet, counterhate tweet) pairs and include their tweet identifiers (Mathew et al., 2020; Albanyan and Blanco, 2022) using Twitter API. This approach resulted in only 260 triples because some tweets are no longer available and not all counterhate tweets have replies.

In the second strategy, we collect hateful tweets from corpora that only provide hateful tweets (Mathew et al., 2021; Chandra et al., 2021; He et al., 2021; Vidgen et al., 2020). Then, we implement the following steps:

1. Extract the replies (candidate counterhate tweets) for the collected hateful tweets using the Twitter API.
2. Use a counterhate classifier (Albanyan and Blanco, 2022) to discard non-counterhate replies.
3. Collect the replies to the counterhate tweets to construct the triples.

This approach resulted in 230 triples. Since the total number of the collected triples from the first and second approaches is still relatively low (490 triples), we apply a third strategy.

In the third strategy, we also utilize the three steps listed in the second strategy. However, we start with two additional steps. First, we collect candidate hateful tweets using the hate speech template *I <hateful_verb> <target_group>* defined by Silva et al. (2021). Second, we use a hate classifier, HateXPlain (Mathew et al., 2021), to discard non-hateful tweets. Then, we follow the three steps from the second approach to collect counterhate tweets and their replies. We retrieved 3,820 triples from this approach.

The total number of triples after combining the three approaches is 4,310. We finalized the collection process by validating the results. The final size of our corpus after the validation process is 2,621 triples (hateful tweet, counterhate tweet, and replies). The total number of hateful tweets is 1,147, while the number of counterhate tweets is 1,685. The number of counterhate tweets ranges between 1 and 20 per hateful tweet, while the number of replies ranges between 1 and 88 per counterhate tweet.

Annotation Guidelines Along with determining whether a reply agrees with the counterhate tweet, we identify finer-grained characterization of the replies. Accordingly, we define three steps to answer four questions in the annotation process (whether the reply *agrees* with the counterhate tweet, whether the not-agreed reply *supports* the hateful tweet and *attacks the author* of the counterhate tweet, and whether the agreed reply adds *additional counterhate*).

The initial step is determining if the reply **agrees** with the counterhate tweet. We define agrees as any reply that does not oppose the counterhate tweet either explicitly or implicitly. On the other hand, we consider not agreeing with the counterhate tweet when the reply uses sarcasm (e.g., you are missing something!) or casts doubt (e.g., are you kidding?)

<table>
<tr><td colspan="2"><i>Hateful Tweet 1</i>: I f**king hate <ethnicity> people. [...] I hope you all f**king die.
<i>Counterhate Tweet</i>: not all <ethnicity> part take in this. cant discriminate a whole race bc some f**k up; do sick things. White's abuse animals too
<i>Reply</i>: but down in <country> they are horrible f**king people</td></tr>
<tr><td>Agree? No
Attacks Author? No</td><td>Support? Yes
Addtl. Counterhate? n/a</td></tr>
<tr><td colspan="2"><i>Hateful Tweet 2</i>: I admit it, I don't like white people
<i>Counterhate Tweet</i>: Appreciate the honesty. You do realize that makes you racist, right?
<i>Reply</i>: thats not wt racism means. f**k off w that bullshit.</td></tr>
<tr><td>Agree? No
Attacks Author? Yes</td><td>Support? No
Addtl. Counterhate? n/a</td></tr>
<tr><td colspan="2"><i>Hateful Tweet 3</i>: If &lt;country> had only shown the true numbers and severity of this virus then maybe some countries would have taken it more seriously much earlier.
<i>Counterhate Tweet</i>: &lt;country> has shown you that ten of thousands people infected for about two months. Few of countries take serious action.
<i>Reply</i>: &lt;country> is doing such a good job[...] truthful government that cares about it citizens. A shining beacon on a hill for the world to emulate.</td></tr>
<tr><td>Agree? Yes
Attacks Author? n/a</td><td>Support? n/a
Addtl. Counterhate? Yes</td></tr>
</table>

Table 1: Three annotation examples of hateful tweets, counterhate tweets, and replies from our corpus. Annotations include four binary questions: whether the reply (a) *Agrees* with the counterhate tweet, (b) *Supports* the hate when disagrees with the counterhate tweet, (c) *Attacks the Author* of the counterhate tweet when disagrees with the counterhate tweet, and (d) adds *Additional Counterhate* when agrees with the counterhate tweet.

	Observed (%)	Cohen's κ
Agree?	91.1	0.82
Support?	89.1	0.77
Attacks Author?	92.3	0.79
Addtl. Counterhate?	91.7	0.81

Table 2: Inter-annotator agreements in our corpus. We provide the observed agreements (percentage of times annotators agreed) and Cohen's κ. κ coefficients between 0.6 and 0.8 are considered *substantial* agreement, and above 0.8 (nearly) perfect (Artstein and Poesio, 2008).

Subsequently, we ask two questions only when the reply does not agree with the counterhate tweet. First, we ask if the not-agreed reply **supports** the hateful tweet. We consider the reply to support the hateful tweet when trying to justify or increase the severity of the hate. For example, providing evidence to justify the content of the hateful tweet (e.g., the news says the opposite!) or introducing a new hateful argument (see the first example in Table 1). Second, we identify whether the reply **attacks the author** of the counterhate tweet. We define attack the author as any tweet that mocks or insults the author of the counterhate tweet (e.g., stupid people never understand!).

Finally, when the reply agrees with the counterhate tweet, we distinguish whether the reply includes **additional counterhate**. We consider that the reply contains additional counterhate when it provides a new opinion or factual arguments to support the counterhate tweet (e.g., he is also known for his charitable work and donations). Agreeing with the counterhate tweet does not necessarily contain additional arguments (e.g., you are correct!).

Examples Table 1 shows three triples examples of the hateful tweet, counterhate tweet, and reply extracted from our corpus. In the first example, the reply not only disagrees with the counterhate tweet but also *supports* the hateful tweet with new hate content against the mentioned people. Replies can show disagreement without including any justification to support the hateful tweet (e.g., do you have any evidence?!!).

In the second example, the reply *attacks the author* of the counterhate tweet without including any justification or support for the hateful tweet. This also indicates that the reply disagrees with the counterhate tweet. Replies can simply disagree without attacking the author (e.g., don't be their lawyer).

Finally, the reply in the third example does not only agree with the claim in the counterhate tweet but also includes *additional counterhate* (honest vs. successful governments). The reply can agree with the counterhate tweet without adding additional counterhate (e.g., convincing response!).

	%Yes	%No
Agree?	49	51
Support?	40	60
Attacks Author?	24	76
Addtl. Counterhate?	35	65

Table 3: Percentages for Yes and No labels per question.

Annotation Process and Inter-Annotator Agreements We used the Label Studio annotation tool.[5] The tool showed the hateful tweet, counterhate tweet, and reply. It displayed the screenshots of the tweets taken from the Twitter website to prevent readability issues when displaying the tweets (e.g., special characters). Additionally, annotators are provided with instructions for each question (i.e., definitions and examples).

The 2,621 triples (hateful tweet, counterhate tweet, reply) were independently annotated by two graduate students active on social media platforms. We are interested in how regular social media users interpret hateful tweets, counterhate tweets, and replies. Table 2 presents the inter-annotator agreements. For all questions, the observed agreements are almost 90%. Cohen's k indicates *substantial* agreement in whether the reply *supports* the hateful tweet and *attacks the author* of the counterhate tasks, and indicates *nearly* perfect agreements in whether the reply *agrees* with the counterhate tweet and includes *additional counterhate* tasks. k coefficients between 0.60 and 0.80 are considered *substantial* agreement, and above 0.80 are considered *nearly* perfect (Artstein and Poesio, 2008). We note that it is easier to determine whether a reply *agrees* and adds *additional counterhate* tasks than *supports* and *attacks the author* tasks. This is due to the use of sarcasm and nuanced language when the reply supports the hateful tweet and attacks the author of the counterhate tweet. After the two annotators finished all the annotations independently, they debated the points of disagreement they had in common and decided on the final label.

4 Corpus Analysis

Label Distribution Table 3 presents the percentages of *yes* and *no* labels per question. Almost half of the replies to the counterhate tweets do not agree with the counterhate tweet (51%), and it is more common for them to *support* the hateful tweet when they do not agree (40%). In addition, it is somewhat rare for these replies to *attack the author* of the counterhate tweet when they disagree (24%). On the other hand, it is less likely for the replies to include *additional counterhate* arguments when they agree (35%). This shows that most replies that agree with the counterhate tweet do not include any additional arguments to support the counterhate tweet (e.g., you are correct).

Linguistic Insights We analyze the language people use in the counterhate tweets that lead to certain types of replies. We count the number of tokens, pronouns, and proper nouns using spaCy (Neumann et al., 2019). We use the lexicons of offensive words [6] and lexicons by Mohammad and Turney (2013) to count offensive, positive, negative, and sadness words. Finally, we use Profanity-check[7] to calculate the profanity score and TextBlob[8] to calculate the subjectivity score. All inter-feature correlations are below 0.30, except for a few that involve the number of tokens feature (Appendix, Figures 2–5). We check the predictive power of the selected features using t-test. We also report if a feature passes the Bonferroni correction (Table 4). The p-values indicate several interesting insights:

- Counterhate tweets with more tokens or pronouns elicit replies that do *not attack the author* of the counterhate tweet or include *additional counterhate* if they agree.
- If the counterhate tweet contains more question marks, this leads to replies that (a) *agree* with the counterhate tweets and do not add *additional counterhate*, or (b) *support* the hateful tweet and do *not attack the author*.
- For different kinds of words (positive, negative, or offensive), we find that (a) positive words elicits replies that do *not attack the author* or adds *additional counterhate*, (b) negative words elicits replies that do not add *additional counterhate*, and (c) offensive words elicits replies that *agree* with the counterhate, or *attack the author* if they disagree.
- Profane counterhate tweets elicit replies that *agree* with the counterhate tweet or do *not support* the hateful tweet.
- Differences between the hateful tweet and counterhate tweet reveal that counterhate

[5] https://github.com/heartexlabs/label-studio

[6] https://github.com/zacanger/profane-words

[7] https://github.com/vzhou842/profanity-check

[8] https://github.com/sloria/TextBlob

	Agree?		Support?		Attacks Author?		Addtl. Counterhate?	
	p-value	Bonf.	p-value	Bonf.	p-value	Bonf.	p-value	Bonf.
Number of …								
tokens					↓↓↓	✓	↑↑↑	✓
pronouns					↓↓↓	✓	↑↑↑	✓
proper nouns	↑	✗			↓	✗		
question marks	↑	✗	↑↑↑	✓	↓↓↓	✓	↑	✗
positive words					↓↓↓	✓	↑↑↑	✓
negative words					↓	✗	↓↓	✓
offensive words	↑	✗			↑	✗		
Profanity score	↑	✗	↓	✗				
With respect to the hateful tweet								
offensive words	↑↑	✓	↓↓	✗				
sadness words	↑↑	✗			↓↓	✗		
subjectivity					↑↑	✗	↓	✗

Table 4: Linguistic analysis of the counterhate tweets depending on our fine-grained characterization of the replies they elicit. 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 *yes* (up) or *no* (down). We use a check mark to indicate tests that pass the Bonferroni correction. Counterhate tweets without offensive words tend to elicit replies that *agree* with the counterhate tweet and *do not support* the hate when they *disagree*.

tweets with (a) less offensive content lead to replies that *agree* with the counterhate tweet or do *not support* the hateful tweet, (b) less sadness words elicit replies that *agree* with the counterhate or do *not attack the author* of the counterhate tweet, and (c) less subjectivity end with replies that *attack the author* of the counterhate or do not add *additional counterhate*.

5 Experiments and Results

We create a binary classifier for each task; whether the reply: (a) agrees with the counterhate tweet, (b) supports the hateful tweet, (c) attacks the author of the counterhate tweet, or (d) includes additional counterhate arguments. We split the dataset into 70:10:20 ratios for training, validation, and testing. Each instance consists of a hateful tweet, a counterhate tweet, and a reply.

Baselines The baseline models we use in our experiments are the *majority* and *random* models. In the *majority* model, the majority label is predicted (*no* label for all tasks, Table 3). In the *random* model, a random label of *no* or *yes* is predicted.

Neural Network Architecture and Training In all experiments, we used the transformer-based BERTweet model (Nguyen et al., 2020). BERTweet is a BERT-based (Devlin et al., 2019) model but was pre-trained using the RoBERTa training strategy (Liu et al., 2019) on 850M English tweets. The neural architecture consists of the base architecture of BERTweet followed by a linear layer with 128 neurons and a ReLU activation. Then, we added a final linear layer with 2 neurons and a Softmax activation to do the binary classification between labels *yes:1* and *no:0*. We perform the experiments using different textual inputs:

- the hateful tweet alone,
- the counterhate tweet alone,
- the reply alone, and
- combination of different tweets.

We use the '</s>' special token to concatenate the inputs. Then, we apply three strategies to enhance the performance of neural models:

Data Augmentation We use the data augmentation technique presented by Marivate and Sefara (2020) called Easy Data Augmentation (EDA). We use the EDA methods: *Synonym Replacement* (randomly replacing a word), *Random Insertion* (inserting a synonym of a random word), and *Random Swap* (randomly swapping the positions of two words).

Concatenating Language Features Language features have been demonstrated to improve pre-trained models' performance in text classification tasks (Lim and Tayyar Madabushi, 2020). We experiment with concatenating language features with input embeddings. We calculate count-based lan-

	Agree?			Support?			Attacks Author?			Addtl. Counterhate?		
	No	Yes	Avg.	No	Yes	Avg.	No	Yes	Avg.	No	Yes	Avg.
Baselines												
Majority	0.67	0.00	0.34	0.75	0.00	0.45	0.87	0.00	0.66	0.79	0.00	0.51
Random	0.52	0.48	0.50	0.51	0.44	0.48	0.58	0.30	0.51	0.54	0.39	0.49
BERTweet trained with . . .												
reply	0.71	0.70	0.70	0.82	0.64	0.75	0.89	0.62	0.83	0.89	0.78	0.85
counterhate tweet	0.64	0.60	0.62	0.70	0.38	0.57	0.86	0.13	0.69	0.73	0.51	0.66
hateful tweet	0.61	0.59	0.60	0.72	0.30	0.55	0.86	0.00	0.66	0.76	0.42	0.64
reply + counterhate tweet	0.72	0.75	0.73	0.80	0.69	0.76	0.89	0.64	0.83	0.89	0.79	0.85
reply + hateful tweet	0.67	0.75	0.71	0.82	0.73	0.78	0.88	0.59	0.81	0.87	0.76	0.83
best pair + the other tweet	0.74	0.71	0.73	0.80	0.68	0.75	0.88	0.56	0.81	0.88	0.76	0.83
best input + EDA	0.75	0.74	0.75	0.84	0.74	0.80	0.89	0.64	0.83	0.89	0.77	0.85
best input + LF	0.74	0.74	0.74	0.84	0.67	0.78	0.90	0.64	0.84	0.88	0.77	0.84
best input + Blending	0.76	0.74	0.75	0.84	0.79	0.82	0.90	0.66	0.84	0.88	0.80	0.85

Table 5: Results obtained with several systems (F1-scores; *Avg.* refers to the *weighted average*). *Best pair*: the pair input that leads to the best pair result (reply+counterhate tweet or reply+hateful tweet). *The other tweet*: either counterhate tweet or hateful tweet. *Best input*: a textual input or combination of (reply, counterhate tweet, and hateful tweet) that leads to the best performance (undeline). *EDA*: easy data augmentation. *LF*: language features. Tables 8–11 provide detailed results per label and subtask.

guage features for the replies, such as the *number of tokens, pronouns, nouns with verbs, negative and positive words* using lexicons by Mohammad and Turney (2013), *question marks, proper nouns, and first-person pronouns.* Examples are shown in the appendix (Table 7). We then use the significance test (t-test) to keep the significant features ($p < 0.05$). The common significant features between the tasks are the *number of tokens, bad words, nouns with verbs, and positive words.* We concatenate these features with each other and with the input embeddings using the '</s>' special token.

Blending with Complementary Tasks We finally investigate pretraining with complementary tasks. We adopt the method by Shnarch et al. (2018), which integrates labeled data from related tasks with various ratios in each training epoch. This is done by blending the related task instances with our dataset for training, and decrease the ratio in each epoch to reach zero in the last one. The corpora we pretrain with are: (a) stance dataset (Mohammad et al., 2016) consists of 4,163 tweets about abortion, atheism, climate change, feminism, and Hillary Clinton annotated in favor, against, or none, (b) offensive dataset (Zampieri et al., 2019b) contains over 14K tweets labeled whether a tweet is offensive or not, and (c) hateful tweet-reply dataset (Albanyan and Blanco, 2022), annotated on whether the reply counters the hateful tweet (5,652 pairs), counters the hate with justification (1,145 pairs), attacks the author of the hateful tweet (1,145 pairs), and includes additional hate (4,507 pairs).

5.1 Quantitative Results

Table 5 shows the results using the F1-score for *no* and *yes* labels, and the weighted average. The appendix contains detailed results showing the precision, recall, and F1-score of each label. The average F1-scores for the majority baseline are 0.34, 0.45, 0.66, and 0.51.

The results of using the neural models with different inputs (the hateful tweet, the counterhate tweet, the reply, or a combination of different tweets) reveal several insights:

- Using the hateful tweet or counterhate tweet alone as an input outperforms the baseline models, showing that hateful tweets or counter-hate tweets are presumably elicit particular kinds of replies.
- Feeding the network with the reply alone yields the best results for single-tweet input.
- Combining the reply with the hateful tweet outperforms the models in *support* the hateful tweet task since, in this task, the reply is related to the hateful tweet. On the other hand, including the counterhate tweets improves the

	Agree?	Support?	Attacks Author?	Additl. Counterhate?
Intricate text				
Sarcasm and implicit meaning	18	20	15	18
Mentions many named entities	6	5	7	6
All	24	25	22	24
General knowledge	16	19	17	12
Short text, less than 5 tokens	20	12	21	14
Misspellings and abbreviations	11	9	11	12
Rhetorical question	8	14	9	9

Table 6: Error types made by the best performing model in each task (*best input* + *blending*, as shown in Table 5). All the numbers are percentages.

results in the other three tasks. We note that it barely affects the *attacks the author* task; this is because the attack can be detected from the reply alone.

- Including a third input (either the counterhate tweet or hateful tweet) to the best pairs (reply+counterhate tweet or reply+hateful tweet) worsens the results (0.73, 0.78, 0.83, and 0.85 vs. 0.73, 0.75, 0.81, and 0.83).

Additionally, the results show modest improvements when applying the three strategies (Data Augmentation, Language Features, and Blending with Complementary Tasks). We find that:

- Data Augmentation strategy benefits the neural network trained with the best input combination in two tasks: *agree* with the counterhate tweet and *support* the hateful tweet.
- Language Features strategy barely improves the results.
- Blending with complementary tasks strategy outperforms all models. More details about the related datasets that lead to the best results in all tasks can be found in the appendix.

We also tried combining the three strategies and found out the combining strategy did not improve the results.

When do the best models make errors? While our best models in each task produce high results (best input + blending, Table 5), we manually analyzed the wrong predictions made by each model. Table 6) shows the error types used in the analysis. We started the analysis by randomly selecting 100 samples from the model produced in the *agree* task. We considered all the wrong predictions for the other three tasks since they were less than 100 samples. They were 59 samples in the *support* task, 46 in the *attack the author* task, and 43 in the *additional counterhate* task. The error types are:

- Intricate text (24%, 25%, 22%, and 24%), which involves using sarcasm and implicit meaning, or mentioning many individuals or entities (e.g., Reply: don't block me I need you so bad. *Agree?* Gold: *No*, Predicted: *Yes*).
- General knowledge (16%, 19%, 17%, 12%), which requires world knowledge and commonsense to understand the meaning of the tweet (e.g., Reply: it's on sky news mate!. *Supports?* Gold: *Yes*, Predicted: *No*).
- Short text (20%, 12%, 21%, and 14%), tweets with less than 5 tokens (e.g., Reply: chill out. *Attack the Author?* Gold: *No*, Predicted: *Yes*).
- misspellings and abbreviations (11%, 9%, 11%, 12%), (e.g., Reply: @auscoups Why r they trending these things. *Addit. counterhate?* Gold: *Yes*, Predicted: *No*).
- Rhetorical question (8%, 14%, 9%, 9%), where a question in a tweet is asked to deliver a point (e.g., Reply: you think this is funny?. *Agree?* Gold: *Yes*, Predicted: *No*).

6 Conclusions

Countering hateful content is an effective way to fight hate speech (Gagliardone et al., 2015). However, counterhate replies may worsen the situation. In this work, we analyze how social media users reply to counterhate tweets (whether the reply agrees with the counterhate tweet, supports the hateful tweet, attacks the author of the counterhate tweet, or includes additional counterhate). We believe detecting these four tasks can help observe the undesirable threads to avoid increasing the hate associated with the discourse (i.e., replying to a counterhate tweet to support the hateful tweet or to attack the author of the counterhate tweet). We find that 51% of the replies do not agree with the counterhate tweet, and it is common for them to *support* the hateful tweet when they do not agree (40%). We analyze the language social media users use in the counterhate tweets that lead to certain types of replies and find several insights. For example, profane counterhate tweets elicit replies that agree with the counterhate tweet. Experimental results show that it is feasible to anticipate the kind of replies a counterhate tweet will elicit.

Limitations

In the data collection process (Section 3), we try to collect the triples (hateful tweet, counterhate tweet, and reply) from existing hateful tweet-reply and hateful tweet corpora (the first and second approaches). However, this ends with fewer triples since some tweets are no longer available and not all counterhate tweets have replies. In addition, we use hate speech and counterhate classifiers to discard non-hateful and non-counterhate tweets. However, this step might (a) discard actual hateful or counterhate tweets that are detected wrongly and (b) keep hateful or counterhate tweets that should be discarded. Another limitation is that we only consider the tweet text. However, some tweets contain text accompanied by images or sometimes images only. Including the tweets' images in the analysis may add more insights.

References

Abdullah Albanyan and Eduardo Blanco. 2022. Pinpointing fine-grained relationships between hateful tweets and replies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10418–10426.

Dana Alsagheer, Hadi Mansourifar, and Weidong Shi. 2022. Counter hate speech in social media: A survey. *arXiv preprint arXiv:2203.03584*.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mohit Chandra, Manvith Reddy, Shradha Sehgal, Saurabh Gupta, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2021. "a virus has no religion": Analyzing islamophobia on twitter during the covid-19 outbreak. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 67–77, New York, NY, USA. Association for Computing Machinery.

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.

Russell B Clayton, Annie Lang, Glenn Leshner, and Brian L Quick. 2019. Who fights, who flees? an integration of the lc4mp and psychological reactance theory. *Media Psychology*, 22(4):545–571.

Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 29–30, New York, NY, USA. Association for Computing Machinery.

Maeve Duggan. 2017. Online harassment 2017. *Pew Research Center*.

European Commission. 2019. The EU Code of Conduct on Countering Illegal Hate Speech Online. Accessed: 2021-5-12.

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.

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

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

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.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ data science*, 11(1):3.

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

Wah Meng Lim and Harish Tayyar Madabushi. 2020. UoB at SemEval-2020 task 12: Boosting BERT with corpus level information. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2216–2221, Barcelona (online). International Committee for Computational Linguistics.

Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 181–190. AAAI Press.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Vukosi Marivate and Tshephisho Sefara. 2020. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 385–399. Springer.

Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. Interaction dynamics between hate and counter users on twitter. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, page 116–124, New York, NY, USA. Association for Computing Machinery.

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. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):369–380.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

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

Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 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*, pages 9–14, Online. Association for Computational Linguistics.

Chikashi Nobata, Joel 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 ’16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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

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.

Stephen A Rains. 2013. The nature of psychological reactance revisited: A meta-analytic review. *Human Communication Research*, 39(1):47–73.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Carla Schieb and Mike Preuss. 2018. Considering the elaboration likelihood model for simulating hate and counter speech on facebook. *SCM Studies in Communication and Media*, 7(4):580–606.

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

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.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2021. Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):687–690.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Emily A Vogels. 2021. The state of online harassment. *Pew Research Center*, 13.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

A Inter-Feature Correlations

Figures 2–5 show the inter-feature correlations for the the linguistic features used in the linguistic analysis (Section 4, Table 4). Most correlation coefficients are less than 0.30 in all four tasks (whether the reply agrees with the counterhate tweet, supports the hateful tweet, attacks the author of the counterhate tweet, or includes additional counterhate). This shows that our analysis captures various kinds of counterhate tweets.

B Implementation Details

We used the transformer-based BERTweet model. The neural architecture consists of the base architecture of BERTweet followed by a linear layer with 128 neurons and a ReLU activation. Then, we added a final linear layer with 2 neurons and a Softmax activation. We prepared the dataset by removing URLs, symbols, additional spaces and then, normalized all text to lowercase. We used the pre-processed data as input to the BERTweet model architecture provided by HuggingFace (Wolf et al., 2020) with its own tokenizer. We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 1e-5, a batch size of 16, and a sparse categorical cross-entropy loss function. The number of tokens per input was 128 with automatic padding enabled for shorter inputs using the <pad>

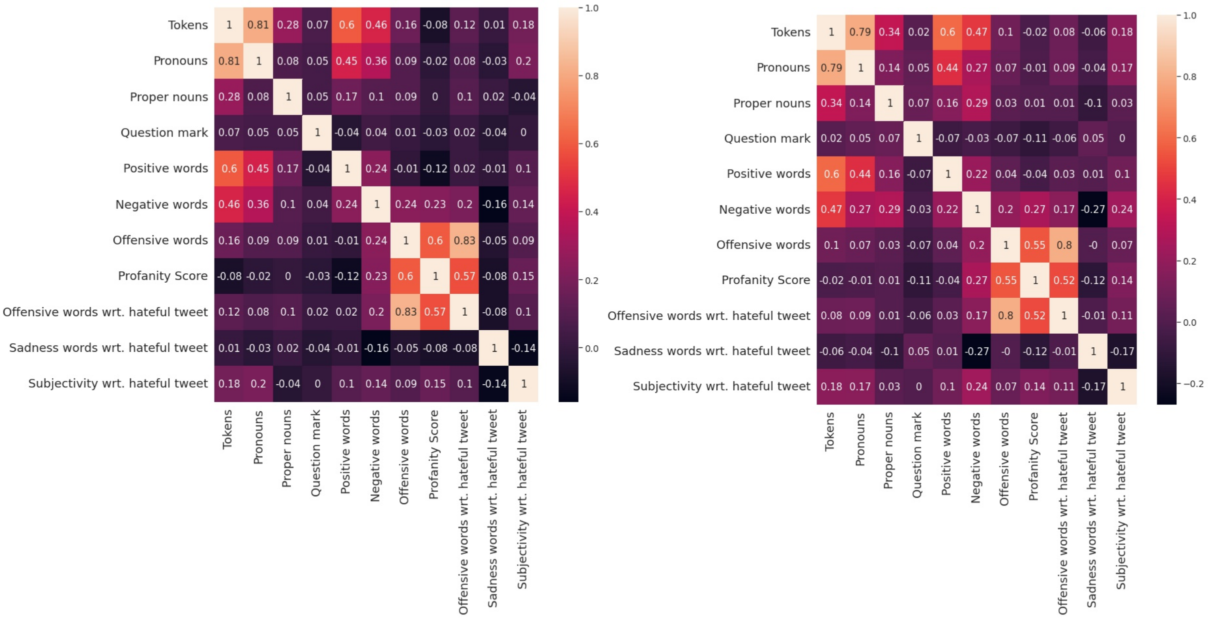


Figure 2: Correlation coefficients between features used in the linguistic analysis. The left and right heatmaps show the correlations with counterhate tweet for the replies that *agree* and do *not agree* with the counterhate tweet respectively.

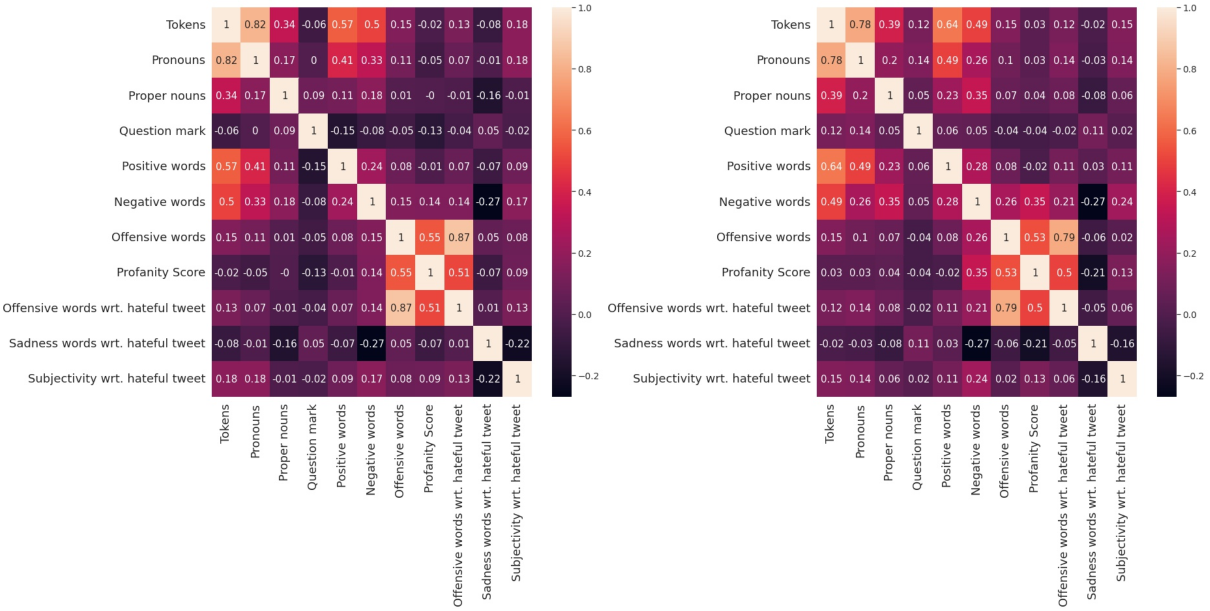


Figure 3: Correlation coefficients between features used in the linguistic analysis. The left and right heatmaps show the correlations with counterhate tweet for the replies that *support* and do *not support* the hateful tweet respectively.

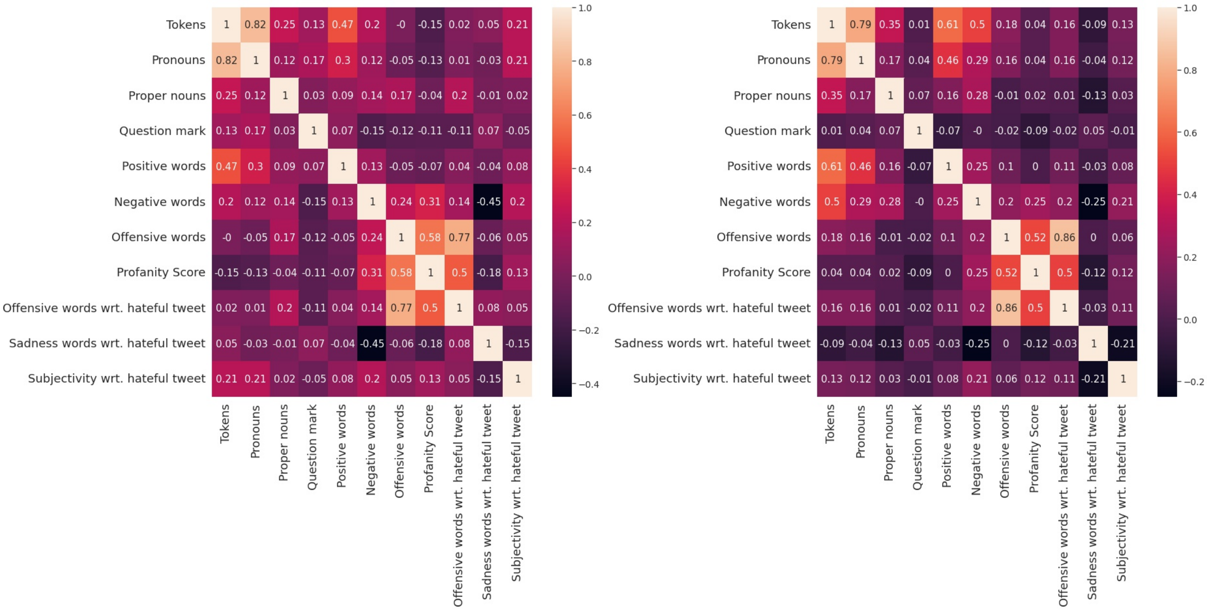


Figure 4: Correlation coefficients between features used in the linguistic analysis. The left and right heatmaps show the correlations with counterhate tweet for the replies that *attack* and do *not attack* the author of the counterhate tweet respectively.

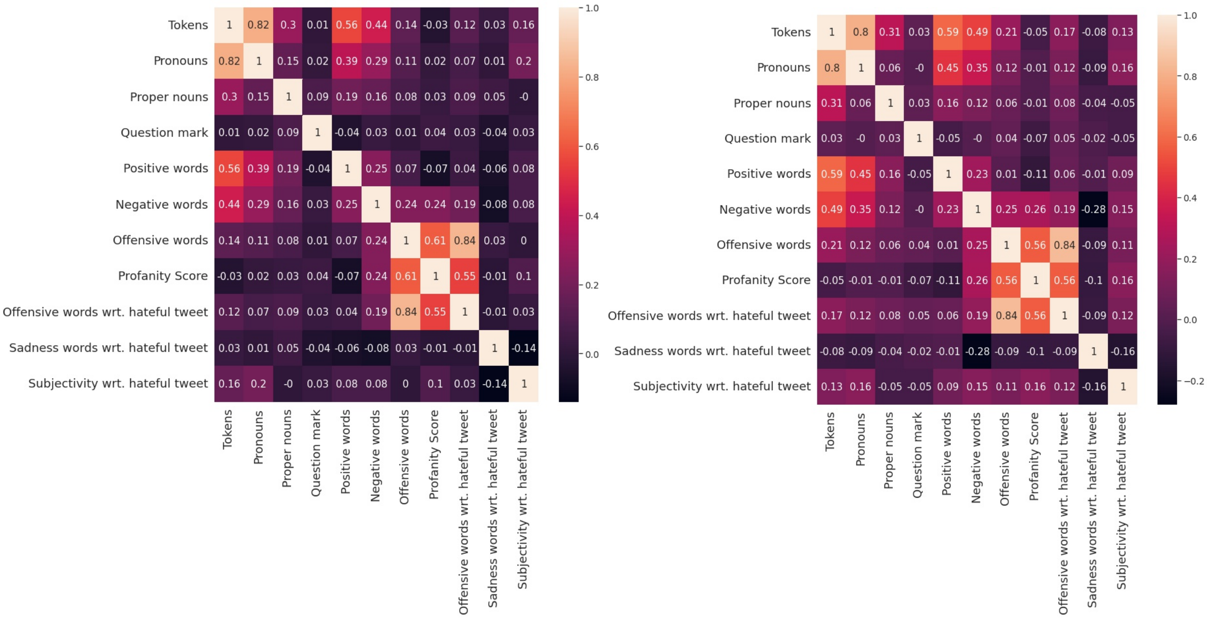


Figure 5: Correlation coefficients between features used in the linguistic analysis. The left and right heatmaps show the correlations with counterhate tweet for the replies that *include* and do *not include* additional counterhate respectively.

token. Models were fine-tuned for 6 epochs and the final fine-tuned model is loaded after the epoch in which it achieved the lowest validation loss.

C Language Features

Table 7 presents examples of applying the language feature strategy on the replies (Section 5). We experiment with concatenating language features presented in the table with input embeddings. The selected language features are number of tokens, pronouns, nouns with verbs, negative and positive words, question marks, proper nouns, and first-person pronouns.

D Detailed Results

Tables 8–11 show the detailed results of Table 5. We provide Precision, Recall and F1-score (a) using different tweet combinations and (b) applying the three strategies to enhance the results. In additiona, we show the results of each related dataset used in the *Blending with Complementary Tasks* strategy. The **related datasets** that lead to the best results in *each task* are:

- **stance dataset**, to determine whether the reply *agrees* with counterhate tweet task;
- **hateful tweet-reply dataset** regarding if a reply includes additional hate, to determine whether the reply *supports* the hateful tweet task;
- **hateful tweet-reply pair dataset** regarding if a reply attacks the author of the hateful tweet, to determine whether the reply *attacks the author* of the counterhate tweet; and
- **hateful tweet-reply pair dataset** regarding if a reply counters the hate with justification, to determine whether the reply adds *additional counterhate*.

	Language Features							
	tokens	pron.	N-and-V	pos.	neg.	QM	PR	FP-pron.
the least you can do is watch what u say, but ur too ignorant.	14	3	4	0	1	0	0	0
Also why poor Becky? She's with a great leading man. I get hating Franco but why the RoHo hate?	19	2	5	0	2	2	3	1
b**ch you lame as f**k hope you got that sh*t if you love gays	14	3	9	2	2	0	1	0
Right??? Like this dude is insane	6	0	1	0	1	3	0	0
Also, I never had the thought to bully someone because I found them weird, that's so toxic wth???	18	5	6	1	3	3	0	2
Who is this one? Are you dumb?	7	2	0	0	1	2	0	0
If there overprotective dosent mean they hate u you know??	10	3	5	0	1	2	0	0
Oh so we are doing that huh , Well Imo killing irl people is cool sounds dumb doesn't it ?	20	3	4	1	2	1	1	1

Table 7: Examples of the calculated language features for the replies. We explore pretraining with the language features as shown in Table 5. *pron.*: Pronouns. *N-and-V*: Nouns and Verbs. *pos.*: Positive words. *neg.*: Negative words. *QM*: Question Marks. *PR*: Proper Nouns. *FP-pron.*: First Person Pronouns.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.50	1.00	0.67	0.00	0.00	0.00	0.25	0.50	0.34
Random	0.51	0.54	0.52	0.50	0.47	0.48	0.50	0.50	0.50
BERTweet trained with ...									
reply	0.70	0.72	0.71	0.72	0.69	0.70	0.70	0.70	0.70
counterhate tweet	0.61	0.67	0.64	0.63	0.57	0.60	0.62	0.62	0.62
hateful tweet	0.60	0.62	0.61	0.58	0.59	0.59	0.60	0.60	0.60
reply + counterhate tweet	0.77	0.68	0.72	0.71	0.79	0.75	0.74	0.73	0.73
reply + hateful tweet	0.81	0.57	0.67	0.66	0.86	0.75	0.73	0.71	0.71
best pair + the other tweet	0.71	0.78	0.74	0.75	0.68	0.71	0.73	0.73	0.73
best input + EDA	0.75	0.74	0.75	0.74	0.75	0.74	0.74	0.74	0.74
best input + LF	0.75	0.73	0.74	0.73	0.75	0.74	0.74	0.74	0.74
best input + Blending with ...									
stance	0.73	0.78	0.76	0.77	0.71	0.74	0.74	0.75	0.75
offensive	0.65	0.83	0.76	0.87	0.49	0.62	0.76	0.71	0.69
counterhate	0.72	0.70	0.71	0.70	0.72	0.71	0.71	0.71	0.71
justification	0.71	0.78	0.74	0.75	0.68	0.71	0.73	0.73	0.73
attack	0.73	0.81	0.76	0.78	0.69	0.73	0.75	0.75	0.75
additional hate	0.69	0.71	0.70	0.70	0.68	0.69	0.70	0.70	0.70

Table 8: Detailed results (P, R, and F) predicting whether the reply *agrees* with the counterhate tweet. *Best pair*: the pair input that leads to the best pair result (reply+counterhate tweet or reply+hateful tweet). *The other tweet*: either counterhate tweet or hateful tweet. *Best input*: a textual input or a combination of (reply, counterhate tweet, and hateful tweet) that leads to the best performance (underline). *EDA*: easy data augmentation. *LF*: language features. This table complements Table 5.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.60	1.00	0.75	0.00	0.00	0.00	0.36	0.60	0.45
Random	0.58	0.45	0.51	0.39	0.51	0.44	0.50	0.48	0.48
BERTweet trained with . . .									
reply	0.74	0.91	0.82	0.79	0.53	0.64	0.76	0.65	0.74
counterhate tweet	0.63	0.80	0.70	0.51	0.30	0.38	0.58	0.60	0.57
hateful tweet	0.62	0.86	0.72	0.51	0.21	0.30	0.57	0.60	0.55
reply + counterhate tweet	0.78	0.83	0.80	0.72	0.66	0.69	0.76	0.76	0.76
reply + hateful tweet	0.81	0.83	0.82	0.74	0.72	0.73	0.78	0.78	0.78
best pair + the other tweet	0.77	0.83	0.80	0.71	0.64	0.68	0.75	0.75	0.75
best input + EDA	0.82	0.86	0.84	0.77	0.72	0.74	0.80	0.80	0.80
best input + LF	0.75	0.96	0.84	0.89	0.54	0.67	0.81	0.79	0.78
best input + Blending with . . .									
stance	0.84	0.73	0.78	0.66	0.80	0.72	0.77	0.76	0.77
offensive	0.78	0.72	0.75	0.63	0.70	0.66	0.72	0.71	0.71
counterhate	0.82	0.80	0.81	0.71	0.73	0.72	0.77	0.77	0.77
justification	0.83	0.83	0.83	0.75	0.75	0.75	0.80	0.80	0.80
attack	0.86	0.78	0.82	0.72	0.81	0.76	0.80	0.79	0.79
additional hate	0.89	0.79	0.84	0.73	0.86	0.79	0.83	0.82	0.82

Table 9: Detailed results (P, R, and F) predicting whether the reply contains *support* to the hateful tweet. *Best pair*: the pair input that leads to the best pair result (reply+counterhate tweet or reply+hateful tweet). *The other tweet*: either counterhate tweet or hateful tweet. *Best input*: a textual input or a combination of (reply, counterhate tweet, and hateful tweet) that leads to the best performance (underline). *EDA*: easy data augmentation. *LF*: language features. This table complements Table 5.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.76	1.00	0.87	0.00	0.00	0.00	0.58	0.76	0.66
Random	0.74	0.47	0.58	0.22	0.47	0.30	0.62	0.47	0.51
BERTweet trained with ...									
reply	0.88	0.90	0.89	0.66	0.59	0.62	0.82	0.83	0.83
counterhate tweet	0.77	0.97	0.86	0.45	0.08	0.13	0.70	0.76	0.69
hateful tweet	0.76	1.00	0.86	0.00	0.00	0.00	0.58	0.76	0.66
reply + counterhate tweet	0.88	0.91	0.89	0.67	0.61	0.64	0.83	0.84	0.83
reply + hateful tweet	0.87	0.90	0.88	0.64	0.55	0.59	0.81	0.82	0.81
best pair + the other tweet	0.85	0.91	0.88	0.64	0.50	0.56	0.80	0.81	0.81
best input + EDA	0.89	0.89	0.89	0.64	0.64	0.64	0.83	0.83	0.83
best input + LF	0.88	0.92	0.90	0.69	0.59	0.64	0.83	0.84	0.84
best input + Blending with ...									
stance	0.85	0.97	0.91	0.81	0.47	0.59	0.84	0.85	0.83
offensive	0.87	0.86	0.87	0.57	0.59	0.58	0.80	0.80	0.80
counterhate	0.91	0.85	0.88	0.61	0.73	0.67	0.84	0.83	0.83
justification	0.88	0.92	0.90	0.70	0.61	0.65	0.84	0.84	0.84
attack	0.89	0.92	0.90	0.71	0.62	0.67	0.85	0.85	0.85
additional hate	0.87	0.93	0.90	0.70	0.55	0.61	0.83	0.84	0.83

Table 10: Detailed results (P, R, and F) predicting whether the reply *attacks the author* of the counterhate tweet. *Best pair*: the pair input that leads to the best pair result (reply+counterhate tweet or reply+hateful tweet). *The other tweet*: either counterhate tweet or hateful tweet. *Best input*: a textual input or a combination of (reply, counterhate tweet, and hateful tweet) that leads to the best performance (underline). *EDA*: easy data augmentation. *LF*: language features. This table complements Table 5.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.65	1.00	0.79	0.00	0.00	0.00	0.42	0.65	0.51
Random	0.63	0.47	0.54	0.33	0.48	0.39	0.52	0.47	0.49
BERTweet trained with …									
reply	0.88	0.90	0.89	0.80	0.76	0.78	0.85	0.85	0.85
counterhate tweet	0.74	0.73	0.73	0.51	0.51	0.51	0.66	0.66	0.66
hateful tweet	0.70	0.82	0.76	0.52	0.36	0.42	0.64	0.66	0.64
reply + counterhate tweet	0.88	0.90	0.89	0.81	0.77	0.79	0.85	0.86	0.85
reply + hateful tweet	0.88	0.86	0.87	0.75	0.77	0.76	0.83	0.83	0.83
best pair + the other tweet	0.86	0.90	0.88	0.80	0.73	0.76	0.84	0.84	0.84
best input + EDA	0.85	0.94	0.89	0.85	0.70	0.77	0.85	0.85	0.85
best input + LF	0.87	0.89	0.88	0.78	0.75	0.77	0.84	0.84	0.84
best input + Blending with …									
stance	0.91	0.85	0.88	0.75	0.84	0.79	0.85	0.85	0.85
offensive	0.89	0.83	0.86	0.71	0.82	0.76	0.83	0.82	0.82
counterhate	0.90	0.83	0.86	0.72	0.84	0.77	0.84	0.83	0.83
justification	0.91	0.85	0.88	0.76	0.85	0.80	0.86	0.85	0.85
attack	0.88	0.84	0.86	0.72	0.78	0.75	0.82	0.82	0.82
additional hate	0.89	0.81	0.84	0.69	0.80	0.74	0.82	0.81	0.81

Table 11: Detailed results (P, R, and F) predicting whether the reply contains *additional counterhate*. *Best pair*: the pair input that leads to the best pair result (reply+counterhate tweet or reply+hateful tweet). *The other tweet*: either counterhate tweet or hateful tweet. *Best input*: a textual input or a combination of (reply, counterhate tweet, and hateful tweet) that leads to the best performance (underline). *EDA*: easy data augmentation. *LF*: language features. This table complements Table 5 .