Pinpointing Fine-Grained Relationships between Hateful Tweets and Replies Supplementary materials

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

A Implementation Details

The Python transformers library was used to load both the base BERT and BERTweet models. Our dataset was pre-processed by removing URLs, removing symbols and the "RT" prefix, removing any additional spaces, and at the end, converting all words to lower-case. The pre-processed data is then fed to BERT and BERTweet models where BERT and BERTweet tokenizers were used to tokenize tweets and to convert them into ids and masks. All models are trained using a learning rate of 1e-5 and a batch size of 16 using AdamW optimizer (Loshchilov and Hutter 2019) with a warm-up steps of 4, and sparse categorical cross entropy loss function. All models are trained for 6 epochs while saving a checkpoint of the model parameters after the epoch in which the model achieved the lowest validation loss.

B Inter-Feature Correlations

Figures 1–4 display the inter-feature correlations used in the linguistic analysis presented in Section 4 of the paper. All correlation are below 0.3 across all questions (whether the reply counters the hateful tweet, includes a justification, attacks author, or includes additional hate) and labels (*yes* and *no*) expect a few involving the number of tokens in the reply. Thus, the statistical analysis in Table 4 of the paper captures different kinds of replies.

C Detailed Results

Tables 1–8 present detailed results complementing Table 5 in the paper. We provide Precision, Recall and F1-measure using BERT (Tables 1–4) and BERTweet (Tables 5–8 using (a) all individual tasks selected for pretraining and (b) all tasks that are beneficial individual when pretraining.

References

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

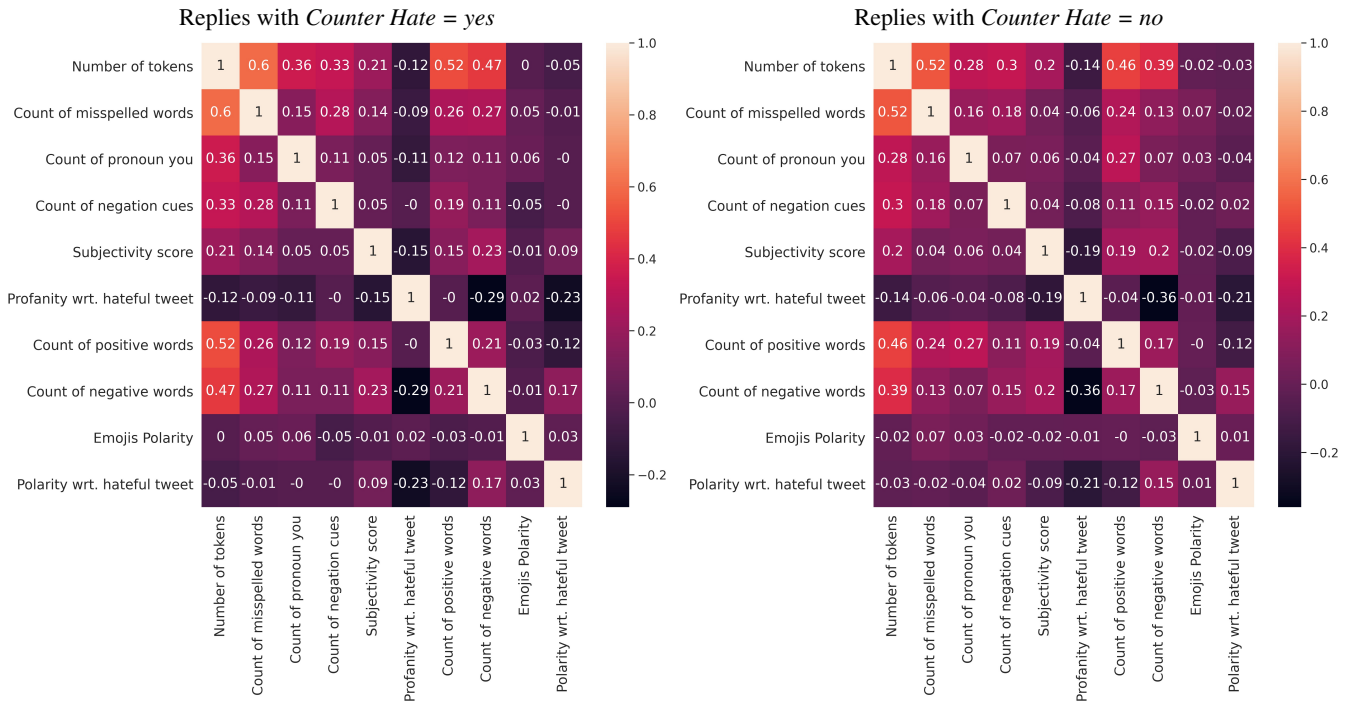

Figure 1: Correlation coefficients between features used in the linguistic analysis. Note that most coefficients are close to zero, meaning that our analysis looks into different characteristics of the replies to hateful tweets. The left and right heatmaps show the correlations with replies that are and are not counter hate respectively.

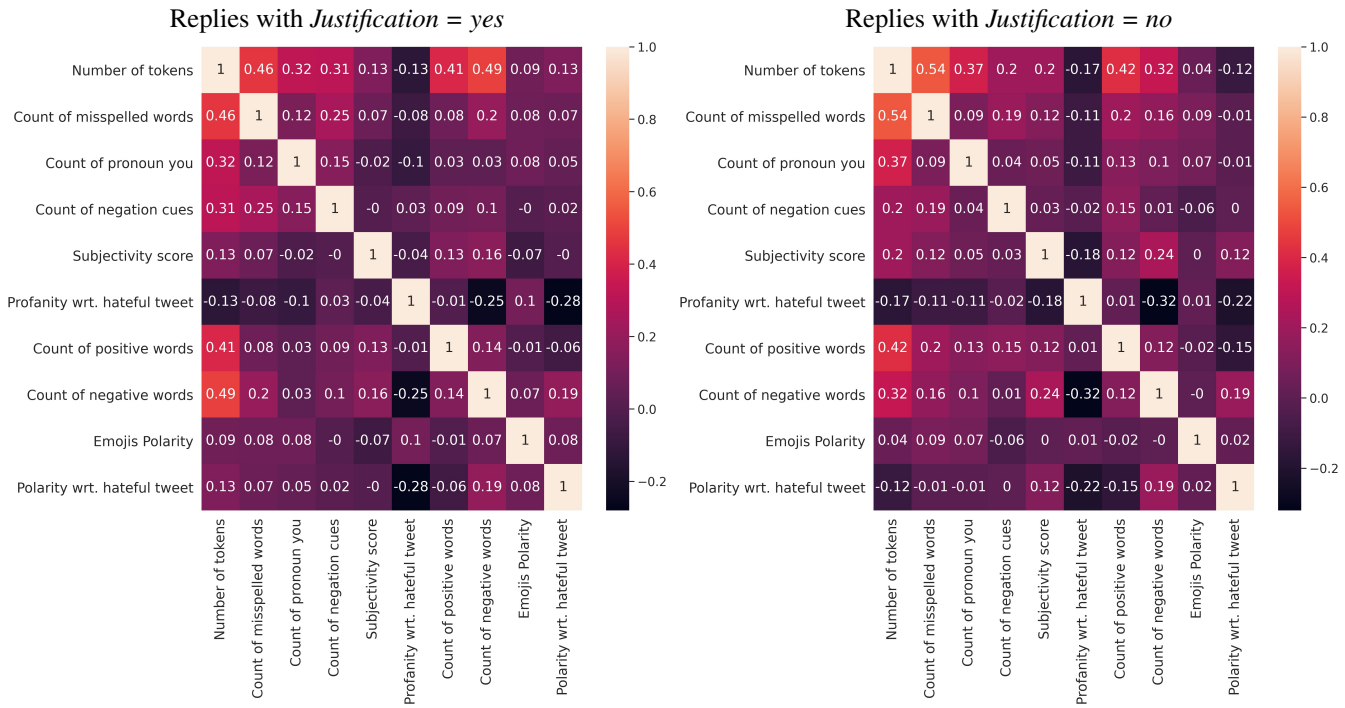

Figure 2: Correlation coefficients between features used in the linguistic analysis. Note that most coefficients are close to zero, meaning that our analysis looks into different characteristics of the replies to hateful tweets. The left and right heatmaps show the correlations with replies that include and do not include a justification respectively.

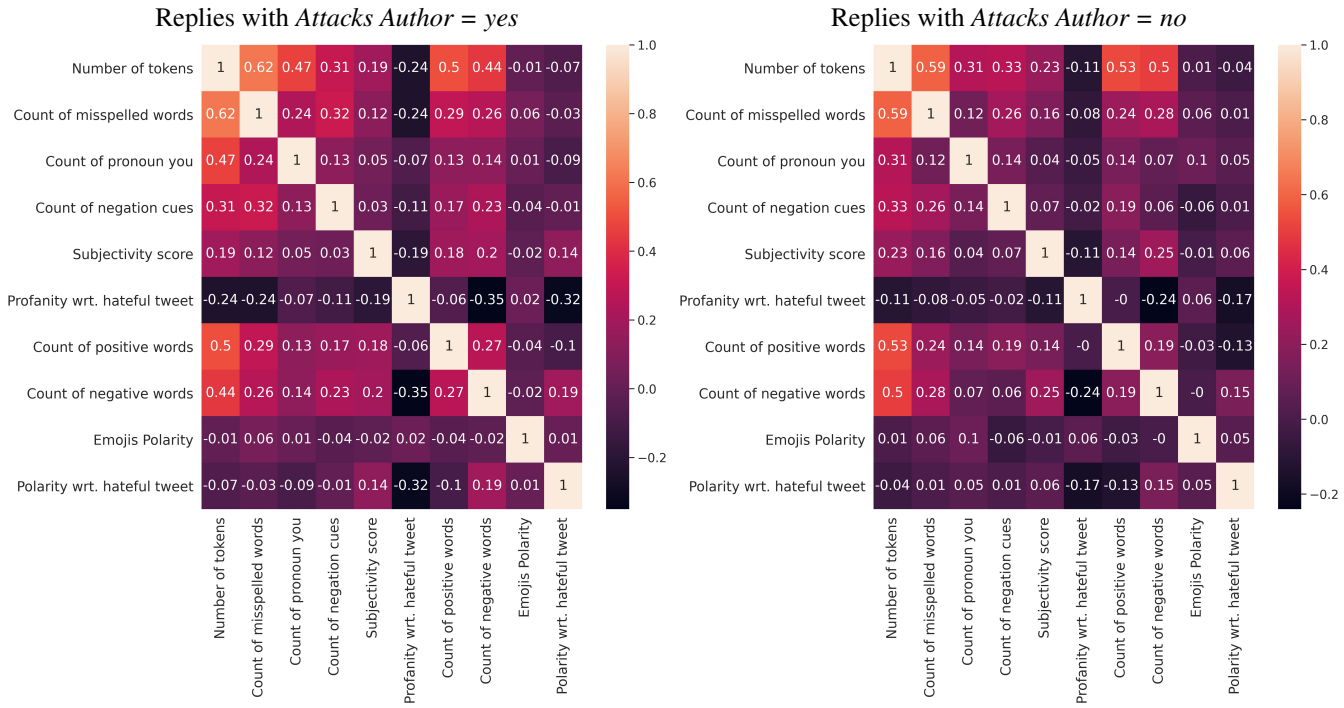

Figure 3: Correlation coefficients between features used in the linguistic analysis. Note that most coefficients are close to zero, meaning that our analysis looks into different characteristics of the replies to hateful tweets. The left and right heatmaps show the correlations with replies that attack and do not attack the author respectively.

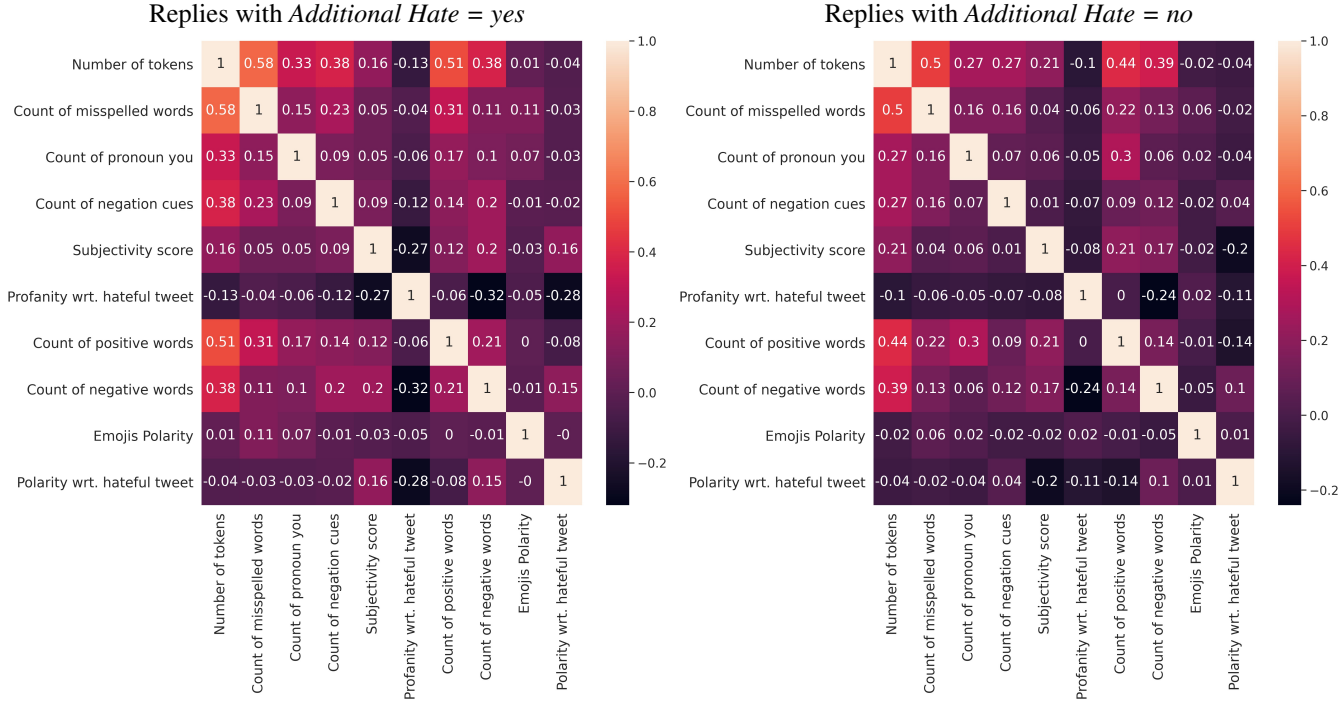

Figure 4: Correlation coefficients between features used in the linguistic analysis. Note that most coefficients are close to zero, meaning that our analysis looks into different characteristics of the replies to hateful tweets. The left and right heatmaps show the correlations with replies that add and do not add additional hate respectively.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.79	1.00	0.88	0.00	0.00	0.00	0.62	0.79	0.70
Random	0.79	0.52	0.63	0.21	0.47	0.29	0.69	0.51	0.55
BERT trained with ...									
hateful tweet	0.81	0.95	0.87	0.47	0.17	0.25	0.74	0.79	0.74
reply	0.85	0.94	0.89	0.60	0.36	0.45	0.79	0.82	0.80
hateful tweet + reply	0.86	0.92	0.89	0.59	0.44	0.50	0.80	0.82	0.81
pretrained with ...									
HateSpeech dataset	0.86	0.93	0.89	0.61	0.42	0.50	0.81	0.82	0.81
Vulgar Dataset	0.86	0.92	0.89	0.61	0.45	0.52	0.81	0.82	0.82
Hate	0.85	0.93	0.89	0.58	0.37	0.45	0.79	0.81	0.80
Offensive	0.85	0.95	0.90	0.66	0.39	0.49	0.81	0.83	0.81
Emotion	0.87	0.95	0.90	0.69	0.44	0.54	0.83	0.84	0.83
Irony	0.86	0.94	0.90	0.65	0.41	0.50	0.82	0.83	0.81
Sentiment	0.86	0.90	0.88	0.54	0.45	0.49	0.79	0.81	0.80
Stance-Abortion	0.88	0.88	0.88	0.55	0.55	0.55	0.81	0.81	0.81
Stance-Hillary	0.87	0.92	0.89	0.60	0.48	0.54	0.81	0.83	0.82
Stance-Climate	0.87	0.92	0.89	0.61	0.46	0.53	0.82	0.82	0.82
Stance-Atheism	0.86	0.91	0.89	0.57	0.46	0.51	0.80	0.82	0.81
Stance-Feminist	0.88	0.87	0.87	0.53	0.56	0.55	0.81	0.80	0.81
All Beneficial Tasks	0.89	0.88	0.88	0.56	0.57	0.57	0.82	0.82	0.82

Table 1: Detailed results (P, R, and F) predicting whether the reply to a hateful tweet is *counter hate*. These results are using BERT and pretraining with (a) all tasks individually and (b) all the tasks that are beneficial individually when pretraining. This table complements Table 5 in the paper.

	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	0.75	0.51	0.60	0.22	0.46	0.30	0.62	0.49	0.53
BERT trained with ...									
hateful tweet	0.81	0.81	0.81	0.39	0.39	0.39	0.71	0.71	0.71
reply	0.85	0.97	0.91	0.83	0.42	0.56	0.84	0.84	0.82
hateful tweet + reply	0.86	0.96	0.91	0.80	0.47	0.60	0.84	0.85	0.83
pretrained with ...									
HateSpeech dataset	0.82	0.94	0.88	0.63	0.32	0.43	0.77	0.78	0.77
Vulgar Dataset	0.86	0.92	0.89	0.65	0.51	0.57	0.81	0.82	0.81
Hate	0.88	0.92	0.90	0.69	0.58	0.63	0.84	0.84	0.84
Offensive	0.86	0.90	0.88	0.61	0.51	0.56	0.80	0.81	0.80
Emotion	0.88	0.92	0.90	0.71	0.61	0.65	0.84	0.85	0.84
Irony	0.88	0.88	0.88	0.60	0.61	0.61	0.81	0.81	0.81
Sentiment	0.85	0.93	0.89	0.68	0.46	0.55	0.81	0.82	0.81
Stance-Abortion	0.87	0.97	0.92	0.84	0.53	0.65	0.85	0.85	0.85
Stance-Hillary	0.85	0.97	0.90	0.81	0.42	0.56	0.84	0.84	0.82
Stance-Climate	0.86	0.95	0.90	0.74	0.49	0.59	0.83	0.84	0.83
Stance-Atheism	0.85	0.92	0.88	0.65	0.47	0.55	0.80	0.82	0.81
Stance-Feminist	0.89	0.86	0.88	0.60	0.66	0.63	0.82	0.81	0.82
All Beneficial Tasks	0.86	0.95	0.90	0.75	0.51	0.61	0.84	0.84	0.83

Table 2: Detailed results (P, R, and F) predicting whether the reply to a hateful tweet contains a *justification*. These results are using BERT and pretraining with (a) all tasks individually and (b) all the tasks that are beneficial individually when pretraining. This table complements Table 5 in the paper.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.61	1.00	0.76	0.00	0.00	0.00	0.37	0.61	0.47
Random	0.58	0.50	0.54	0.35	0.42	0.38	0.49	0.47	0.48
BERT trained with ...									
hateful tweet	0.69	0.72	0.70	0.52	0.48	0.50	0.62	0.63	0.63
reply	0.77	0.74	0.75	0.61	0.64	0.62	0.70	0.70	0.70
hateful tweet + reply	0.77	0.76	0.76	0.63	0.64	0.63	0.71	0.71	0.71
pretrained with ...									
HateSpeech dataset	0.75	0.86	0.80	0.71	0.56	0.62	0.73	0.74	0.73
Vulgar Dataset	0.72	0.89	0.79	0.72	0.44	0.55	0.72	0.71	0.71
Hate	0.72	0.88	0.79	0.70	0.45	0.55	0.71	0.71	0.70
Offensive	0.78	0.79	0.78	0.66	0.64	0.65	0.73	0.73	0.73
Emotion	0.80	0.83	0.81	0.71	0.66	0.68	0.76	0.76	0.76
Irony	0.72	0.80	0.76	0.62	0.52	0.56	0.68	0.69	0.68
Sentiment	0.75	0.73	0.74	0.59	0.61	0.60	0.69	0.68	0.69
Stance-Abortion	0.74	0.83	0.78	0.67	0.54	0.59	0.71	0.71	0.71
Stance-Hillary	0.72	0.77	0.74	0.58	0.52	0.55	0.67	0.67	0.67
Stance-Climate	0.74	0.84	0.78	0.67	0.53	0.59	0.71	0.71	0.71
Stance-Atheism	0.73	0.84	0.78	0.67	0.49	0.57	0.70	0.70	0.70
Stance-Feminist	0.70	0.86	0.77	0.65	0.40	0.50	0.67	0.68	0.67
All Beneficial Tasks	0.77	0.89	0.83	0.77	0.58	0.66	0.77	0.77	0.76

Table 3: Detailed results (P, R, and F) predicting whether the reply to a hateful tweet *attacks the author of the hateful tweet*. These results are using BERT and pretraining with (a) all tasks individually and (b) all the tasks that are beneficial individually when pretraining. This table complements Table 5 in the paper.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.77	1.00	0.87	0.00	0.00	0.00	0.59	0.77	0.67
Random	0.77	0.51	0.61	0.23	0.48	0.31	0.64	0.50	0.54
BERT trained with ...									
hateful tweet	0.77	0.99	0.87	0.43	0.03	0.05	0.69	0.77	0.68
reply	0.91	0.91	0.91	0.70	0.68	0.69	0.86	0.86	0.86
hateful tweet + reply	0.91	0.88	0.89	0.64	0.71	0.67	0.85	0.84	0.84
pretrained with ...									
HateSpeech dataset	0.92	0.91	0.91	0.71	0.72	0.72	0.87	0.87	0.87
Vulgar Dataset	0.89	0.95	0.92	0.79	0.61	0.69	0.87	0.87	0.87
Hate	0.90	0.91	0.91	0.69	0.68	0.68	0.85	0.85	0.85
Offensive	0.91	0.93	0.92	0.76	0.70	0.72	0.87	0.88	0.88
Emotion	0.90	0.93	0.91	0.73	0.65	0.69	0.86	0.87	0.86
Irony	0.87	0.92	0.89	0.66	0.52	0.58	0.82	0.83	0.82
Sentiment	0.89	0.94	0.92	0.77	0.63	0.69	0.86	0.87	0.87
Stance-Abortion	0.91	0.86	0.89	0.61	0.73	0.66	0.84	0.83	0.83
Stance-Hillary	0.90	0.93	0.91	0.73	0.65	0.69	0.86	0.87	0.86
Stance-Climate	0.89	0.93	0.91	0.72	0.63	0.67	0.85	0.86	0.85
Stance-Atheism	0.90	0.89	0.89	0.63	0.65	0.64	0.84	0.83	0.83
Stance-Feminist	0.89	0.91	0.90	0.67	0.64	0.66	0.84	0.84	0.84
All Beneficial Tasks	0.89	0.95	0.92	0.78	0.62	0.69	0.87	0.87	0.87

Table 4: Detailed results (P, R, and F) predicting whether the reply to a hateful tweet contains *additional hate*. These results are using BERT and pretraining with (a) all tasks individually and (b) all the tasks that are beneficial individually when pretraining. This table complements Table 5 in the paper.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.79	1.00	0.88	0.00	0.00	0.00	0.62	0.79	0.70
Random	0.79	0.52	0.63	0.21	0.47	0.29	0.69	0.51	0.55
BERTweet trained with ...									
hateful tweet	0.80	0.96	0.88	0.45	0.12	0.19	0.73	0.78	0.73
reply	0.89	0.88	0.88	0.57	0.57	0.57	0.82	0.82	0.82
hateful tweet + reply	0.88	0.92	0.90	0.60	0.56	0.58	0.82	0.83	0.83
pretrained with ...									
HateSpeech dataset	0.88	0.91	0.90	0.61	0.55	0.58	0.82	0.83	0.83
Vulgar Dataset	0.89	0.88	0.89	0.57	0.58	0.58	0.82	0.82	0.82
Hate	0.90	0.89	0.89	0.60	0.61	0.60	0.84	0.83	0.83
Offensive	0.88	0.89	0.89	0.58	0.56	0.57	0.82	0.82	0.82
Emotion	0.89	0.90	0.89	0.60	0.58	0.59	0.83	0.83	0.83
Irony	0.88	0.92	0.90	0.63	0.52	0.57	0.83	0.84	0.83
Sentiment	0.88	0.89	0.88	0.57	0.52	0.54	0.81	0.82	0.81
Stance-Abortion	0.87	0.90	0.89	0.57	0.52	0.54	0.81	0.82	0.81
Stance-Hillary	0.89	0.91	0.90	0.61	0.56	0.58	0.83	0.84	0.83
Stance-Climate	0.88	0.89	0.89	0.58	0.56	0.57	0.82	0.82	0.82
Stance-Atheism	0.88	0.89	0.88	0.57	0.54	0.56	0.81	0.82	0.82
Stance-Feminist	0.88	0.91	0.89	0.61	0.52	0.56	0.82	0.83	0.83
All Beneficial Tasks	0.87	0.94	0.90	0.67	0.49	0.57	0.83	0.84	0.83

Table 5: Detailed results (P, R, and F) predicting whether the reply to a hateful tweet is *counter hate*. These results are using BERTweet and pretraining with (a) all tasks individually and (b) all the tasks that are beneficial individually when pretraining. This table complements Table 5 in the paper.

	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	0.75	0.51	0.60	0.22	0.46	0.30	0.62	0.49	0.53
BERTweet trained with ...									
hateful tweet	0.81	0.83	0.82	0.40	0.36	0.38	0.71	0.72	0.72
reply	0.89	0.94	0.91	0.77	0.61	0.68	0.86	0.86	0.86
hateful tweet + reply	0.91	0.90	0.90	0.67	0.69	0.68	0.85	0.85	0.85
pretrained with ...									
HateSpeech dataset	0.88	0.96	0.92	0.83	0.58	0.68	0.86	0.87	0.86
Vulgar Dataset	0.89	0.94	0.92	0.77	0.68	0.72	0.87	0.88	0.87
Hate	0.89	0.95	0.92	0.79	0.63	0.70	0.87	0.87	0.87
Offensive	0.91	0.92	0.92	0.74	0.71	0.70	0.87	0.87	0.87
Emotion	0.87	0.94	0.90	0.72	0.53	0.61	0.83	0.84	0.83
Irony	0.84	0.95	0.89	0.71	0.42	0.53	0.81	0.82	0.81
Sentiment	0.91	0.94	0.93	0.78	0.71	0.74	0.88	0.88	0.88
Stance-Abortion	0.90	0.94	0.92	0.76	0.64	0.70	0.86	0.87	0.86
Stance-Hillary	0.91	0.94	0.93	0.78	0.71	0.74	0.88	0.88	0.88
Stance-Climate	0.90	0.95	0.93	0.81	0.66	0.73	0.88	0.88	0.88
Stance-Atheism	0.91	0.95	0.93	0.82	0.68	0.74	0.88	0.89	0.88
Stance-Feminist	0.91	0.93	0.92	0.76	0.71	0.74	0.88	0.88	0.88
All Beneficial Tasks	0.90	0.94	0.92	0.77	0.68	0.72	0.87	0.88	0.87

Table 6: Detailed results (P, R, and F) predicting whether the reply to a hateful tweet contains a *justification*. These results are using BERTweet and pretraining with (a) all tasks individually and (b) all the tasks that are beneficial individually when pretraining. This table complements Table 5 in the paper.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.61	1.00	0.76	0.00	0.00	0.00	0.37	0.61	0.47
Random	0.58	0.50	0.54	0.35	0.42	0.38	0.49	0.47	0.48
BERTweet trained with ...									
hateful tweet	0.69	0.77	0.73	0.55	0.44	0.49	0.64	0.65	0.64
reply	0.78	0.82	0.80	0.69	0.63	0.66	0.75	0.75	0.75
hateful tweet + reply	0.77	0.89	0.83	0.77	0.59	0.67	0.77	0.77	0.77
pretrained with ...									
HateSpeech dataset	0.80	0.77	0.78	0.65	0.70	0.68	0.75	0.74	0.74
Vulgar Dataset	0.80	0.86	0.83	0.75	0.67	0.71	0.78	0.79	0.78
Hate	0.82	0.85	0.84	0.75	0.71	0.73	0.80	0.80	0.80
Offensive	0.83	0.83	0.83	0.73	0.72	0.73	0.79	0.79	0.79
Emotion	0.80	0.82	0.81	0.70	0.68	0.69	0.76	0.76	0.76
Irony	0.79	0.79	0.79	0.67	0.67	0.67	0.75	0.75	0.75
Sentiment	0.79	0.80	0.79	0.67	0.66	0.67	0.74	0.75	0.74
Stance-Abortion	0.79	0.85	0.82	0.73	0.64	0.68	0.77	0.77	0.77
Stance-Hillary	0.79	0.82	0.81	0.70	0.66	0.68	0.76	0.76	0.76
Stance-Climate	0.80	0.84	0.82	0.72	0.66	0.69	0.77	0.77	0.77
Stance-Atheism	0.78	0.84	0.81	0.71	0.62	0.66	0.75	0.75	0.75
Stance-Feminist	0.73	0.82	0.77	0.64	0.52	0.57	0.69	0.70	0.69
All Beneficial Tasks	0.79	0.88	0.83	0.77	0.62	0.69	0.78	0.78	0.78

Table 7: Detailed results (P, R, and F) predicting whether the reply to a hateful tweet *attacks the author of the hateful tweet*. These results are using BERTweet and pretraining with (a) all tasks individually and (b) all the tasks that are beneficial individually when pretraining. This table complements Table 5 in the paper.

	No			Yes			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1
Baselines									
Majority	0.77	1.00	0.87	0.00	0.00	0.00	0.59	0.77	0.67
Random	0.77	0.51	0.61	0.23	0.48	0.31	0.64	0.50	0.54
BERTweet trained with ...									
hateful tweet	0.77	1.00	0.87	0.67	0.01	0.02	0.75	0.77	0.67
reply	0.95	0.91	0.93	0.74	0.85	0.79	0.91	0.90	0.90
hateful tweet + reply	0.95	0.91	0.93	0.74	0.83	0.78	0.90	0.89	0.89
pretrained with ...									
HateSpeech dataset	0.95	0.91	0.93	0.74	0.84	0.78	0.90	0.90	0.90
Vulgar Dataset	0.95	0.92	0.93	0.75	0.83	0.79	0.90	0.90	0.90
Hate	0.94	0.92	0.93	0.76	0.81	0.78	0.90	0.90	0.90
Offensive	0.95	0.93	0.94	0.78	0.82	0.80	0.91	0.91	0.91
Emotion	0.94	0.92	0.93	0.74	0.79	0.77	0.89	0.89	0.89
Irony	0.95	0.92	0.93	0.75	0.82	0.79	0.90	0.90	0.90
Sentiment	0.94	0.92	0.93	0.75	0.79	0.77	0.89	0.89	0.89
Stance-Abortion	0.95	0.91	0.93	0.74	0.82	0.78	0.90	0.89	0.90
Stance-Hillary	0.94	0.92	0.93	0.76	0.80	0.78	0.90	0.89	0.90
Stance-Climate	0.94	0.90	0.92	0.72	0.82	0.77	0.89	0.88	0.89
Stance-Atheism	0.96	0.90	0.93	0.71	0.86	0.78	0.90	0.89	0.89
Stance-Feminist	0.94	0.92	0.93	0.74	0.80	0.77	0.89	0.89	0.89
All Beneficial Tasks	0.94	0.91	0.93	0.74	0.82	0.78	0.90	0.89	0.89

Table 8: Detailed results (P, R, and F) predicting whether the reply to a hateful tweet contains *additional hate*. These results are using BERTweet and pretraining with (a) all tasks individually and (b) all the tasks that are beneficial individually when pretraining. This table complements Table 5 in the paper.