Improving Hate Speech Classification on Twitter

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Hate speech and offensive language have begun to perpetuate online communities that were originally designed to foster community and bring people together. Both anonymity and the ease of spreading content online have made it easier for hateful speech to infiltrate large communities like Twitter. Many instances of hate speech occur in contexts where no explicit hate terms are used. This problem could be helped by a Machine Learning classifier that identifies hate speech. However, until Davidson et al. (2017)'s research, all classifiers were binary, classifying speech as either offensive or not. Hate speech is a separate from category offensive speech because it targets individuals based on nationality, ethnicity, religion, gender, sexual discrimination, disability or class in an especially aggressive or demeaning manner (Tuckwood, 2017). Davidson et al. (2017) were the first to identify this field as needing at least three classes: Hate Speech, Offensive Language, or Neither.Davidson et al. (2017) identify classification of tweets without explicit hate speech as difficult to correctly classify. We attempted to take Davidson et al. (2017)'s research further by adding more robust features to the Logistic Regression model in order to better capture the context surrounding tweets that don't contain explicit hate terms and correctly classify them.

2 Data and Methodology

The Davidson et al. (2017) data was obtained using the Twitter API to obtain 85.4 million tweets from 33,548 users, of which 24,783 tweets were selected to make up the final dataset. Crowdflower, a crowd-sourcing website was used and annotators were provided with a formal definition of hate speech and asked to label each tweet as hate speech, offensive but not hate speech, or neither offensive nor hate speech. Every tweet was labeled by at least three annotators, and mean inter-annotator agreement was 92%.

2.1 Training

We trained all models using a set of 19K tweets from the dataset. Each model had a feature set concatenated with the baseline features. These models were then run through 5-fold cross-validation grid search on a Logisitic Regression model.

2.2 Features / Baseline Features

We implemented a number of hand-built features and utilized Flair embeddings Akbik et al. (2018) as well. These were used in conjunction with the baseline feature set. We utilized Davidson et al. (2017)'s feature set as our baseline. These features included uni/bi/trigrams weighted by TF-IDF, binary and count indicators for hashtags, mentions, retweets, and URLs. To capture syntactic structure information they used NLTK and Penn Part-of-Speech (POS) taggings

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

2.2.1 Hand Built Features

We built a lexicon of ethnic and group membership words. We used this lexicon to create a binary feature capturing if a tweet contains a statement targeting a specific group of people, i.e. "all you Asians" or "every Mexican." This could possibly capture the nuance of a statement that isn't explicitly hateful. Similarly, we tried to identify tweets where the name of a group was followed by a modal verb like "should" or "can." We also tried to capture self-reference when an allusion to specific group was made by implementing a feature that looked for first person pronouns followed by a word indicating group membership.

We implemented an indicator feature for offensive / hate speech geared towards women. We sourced gendered insults towards women to form a lexicon, where terms were scraped from a crowd-sourced post (sac, 2018). This context-based feature was performed in two-steps: first, identify if a tweet is aimed at a female (as indicated by pronouns). Second, check if the tweet has a gendered insult. This differs from a simple 'contains check' because many female-specific insults like "feisty", "bossy", etc. are offensive only in the context of being aimed at a woman.

Slang is an important characterization of tweets, so we wanted to capture the meaning behind slang words instead of ignoring them. To decode slang terms, we mapped common Twitter slang terms to their definitions and replaced any instance of slang with its definition. After replacing the slang, we extracted the sentiment of the tweet. We utilized the Marcus et al. (1993) PennTreebank to convert tweets to Wordnet tags (Miller, 1995) to get the sentiment of tweets with slang replaced with their definition.

We utilized the NRC Emotion Lexicon (Mohammad and Turney, 2013) to count the number of tokens in a tweet referring to a specific emotion. We used the count for each emotion as its own separate feature.

In addition to contextual features, we included some lexical features. The (Davidson et al., 2017) model utilizes a porter stemmer when processing the tweets; we included a feature that did not stem the words in tweets when creating tfidf weightings to see if that helped capture sentiment in tweets that may not be explicitly offensive. We also included indicator features such as if a tweet referenced immigrants directly and a feature that searched for a group membership word inside of quotes.

2.2.2 Flair

We wanted to include contextual string embeddings to better capture sentence-level context, since we were interested in capturing the nuanced context of a tweet that contains hateful speech without being explicit. Akbik et al. (2018) created an embedding library called **Flair** that provides word and sentence-level pre-trained embeddings. One of their word-level embeddings was trained using Twitter. We chose to use this set of embeddings converted to sentence level, which Akbik et al. (2018) call "Document Pool" embeddings. We also utilized Flair's contextual string embeddings, one of which was BERT embeddings (originally developed by Devlin et al. (2018)). The second set of embeddings was trained on a 1 billion word corpus from the news. We trained models using Twitter on its own as well as in combination with the news and BERT embeddings. Akbik et al. (2018) recommend "stacking" word and string embeddings for the best results.

3 Results

3.1 Model Performance

Final results were reported after all models were run on a held-out test set comprised of 5K tweets. We used 5-fold cross-validation grid search on a Logisitc Regression model to find the optimal parameters for each model. Our best performing model outperformed the baseline model by 2% in recall of the Hate Speech class and by 3% in macro averaged recall. Interestingly, hand built features did not seem to increase classification recall on the currently labeled data set. However, as will be explored in the error analysis, this does not necessarily mean that the hand-built features are in reality worse at correctly identifying hate speech.

References

2018. Everyday misogyny: 122 subtly sexist words about women (and what to do about them). *Sacraparental*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*.

Christopher Tuckwood. 2017. Hatebase: Online database of hate speech. *The Sentinal Project. Available at: https://www. hatebase. org*.

Improving Hate Speech Classification on Twitter

Anonymous Author(s)
Affiliation
Address
email

1 Combined Features

We wanted to compare the performance of our hand-built features with the embeddings both separately and combined. We concatenated the embedding features first with the baseline set of features and trained the model on those feature sets. We did the same for the baseline features plus our hand built features. After training separate models for these, we combined the hand-built features with the Twitter embeddings and the combined news, BERT, and Twitter embeddings.

Table 1: Davidson et al. (2017) Logistic Regression Baseline

	Precision	Recall	F1-score
Hate Speech	0.30	0.45	0.36
Offensive	0.94	0.86	0.90
Neither	0.66	0.81	0.73
Micro avg	0.83	0.83	0.83
Macro avg	0.64	0.71	0.66
Weighted avg	0.86	0.83	0.84

Table 2: Twitter Embeddings

	Precision	Recall	F1-score
Hate Speech	0.25	0.35	0.29
Offensive	0.92	0.87	0.90
Neither	0.68	0.75	0.71
Micro avg	0.83	0.83	0.83
Macro avg	0.62	0.66	0.63
Weighted avg	0.84	0.83	0.83

Table 3: News, Twitter, and BERT Embeddings

	Precision	Recall	F1-score
Hate Speech	0.31	0.47	0.38
Offensive	0.95	0.88	0.91
Neither	0.74	0.85	0.79
Micro avg	0.86	0.86	0.86
Macro avg	0.67	0.74	0.69
Weighted avg	0.88	0.86	0.87

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

Table 4: Hand-Built Features

	Precision	Recall	F1-score
Hate Speech	0.24	0.36	0.29
Offensive	0.91	0.86	0.89
Neither	0.65	0.71	0.68
Micro avg	0.81	0.81	0.81
Macro avg	0.60	0.64	0.62
Weighted avg	0.83	0.81	0.82

Table 5: Hand-Built Features with Twitter Embeddings

	Precision	Recall	F1-score
Hate Speech	0.27	0.38	0.32
Offensive	0.92	0.87	0.90
Neither	0.68	0.75	0.71
Micro avg	0.83	0.83	0.83
Macro avg	0.62	0.67	0.64
Weighted avg	0.85	0.83	0.84

Table 6: Hand-Built Features with News, BERT, and Twitter Embeddings

	Precision	Recall	F1-score
Hate Speech	0.31	0.47	0.38
Offensive	0.95	0.89	0.92
Neither	0.74	0.84	0.79
Micro avg	0.86	0.86	0.86
Macro avg	0.67	0.73	0.69
Weighted avg	0.88	0.86	0.87

References

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

Improving Hate Speech Classification on Twitter

Anonymous Author(s)
Affiliation
Address
email

0.1 Error Analysis

Error analysis with this dataset proved difficult because not all examples of tweets labeled as hate speech are truly hate speech. Some examples include :

"I'm not really a phone kinda guy.. I actually hate talking on the phone amp; texting kinda trash to me also."
"Whipped out some french in front of some babes at the post office. winning"

The Hand-Built Features with News, BERT, and Twitter Embedding (Table 6) model had 591 misclassified tweets. Of these, 88 were in the hate speech class, 99 in the neither class, 404 in the offensive class.

The News, Twitter, and BERT Embeddings ((without our hand built features)) model had 489 misclassified tweets. Of these, 32 were in the hate speech class, 82 neither, and 375 offensive. Out of the 489 misclassified tweets we do not agree with the labeling of 12% of the tweets with 36% of those both wrongly classified by our model and wrongly labeled and 64% correctly predicted by our model.

In the following sections we take a deep dive comparing the models Hand-Built Features with News, BERT, and Twitter Embedding (Table 6) and News, Twitter, and BERT Embeddings (Table 3).

0.2 Hand-Built Features with News, BERT, and Twitter Embedding model Class: Hate Speech

In the Hand-Built Features with News, BERT, and Twitter Embedding model we found 29 instances of tweets that we believe were incorrectly labeled. That's 35% of all missed tweets in the hate speech category. Of those, 3% of the tweets were both incorrectly labeled and incorrectly predicted by our model, leaving 32% of tweets in the hate speech class that our model correctly predicted.

Amongst the hate speech labels we agree with, the targeted groups were: 25% Female with one instance threatening violence and another suggesting the target commit suicide. 73% of these were predicted as offensive showing a bias towards the offensive class when concerning women and variations of the terms 'hoe' and 'bitches'.

20% Gay Community. The diverse variations and spellings of the term 'fag' make it difficult to weight it towards hate speech. One possible feature could be a regex for the term 'fag' or gay in conjunction with a swear word. Targets of the term are as follows: 6/12 males as a means to emasculate; 4/12 women; 1/12 males; 1/12 the gay community in general.

$<$ 14% Males with six instances including insults meant to emasculate with three of those also including threats of violence. A feature looking at males as a target and the usage of terms 'bitch' and 'pussy' could weigh it from offensive to hate speech.

$<$ 12% African American; with most of the tweets having hard to discern context identifying them as hate speech such as a link to an article, usage of the n word in different spellings, and one tweet

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

where the hate speech was in quotes making it difficult to know if it was commentary condemning it or agreeing with it. Targets of the term are as follows: 1/7 African American females; 6/7 African Americans in general.

$< 12\%$ White people with one threat of violence and two including politics. The term 'white trash' was in almost every instance. When running our features, we ran them over each token so a possible feature is to do a regex for the term and format it as one token.

8% Asian; with every instance targeting Chinese people including some variation of the term 'chink', making it hard to understand why our model failed to flag it as hate speech especially since half of these instances were predicted to the class 'neither' by the model. One possibility is the low number of hate speech in the training class making use of the term and the alternate definition of chink meaning "a narrow opening or crack, typically one that admits light." Targets of the term are as follows: 4/5 Chinese; 1/5 Asians and the gay community as a means to emasculate.

Remainder: 2/59 Politics, 1/59 general racist, 1/59 Jewish, 1/59 Latinos

Many of the tweets included masked swear or hate words surrounded by hash tags, or html tags. Making a regex for these terms could help identify the intent.

0.3 News, BERT, and Twitter Embedding model Class: Hate Speech

For the News, Twitter, and BERT Embeddings model, we found 4 instances of tweets that we disagree are hate speech. Three of these were correctly predicted by our model and one tweet was both incorrectly labeled and wrongly predicted by our model.

For the News, Twitter, and BERT Embeddings model (without our hand built features), out of the 12% of the tweets that were incorrectly labeled, 36% of those were both wrongly classified by our model and wrongly labeled and 64% correctly predicted by our model.

In the hate speech class, our model correctly predicted three instances labeled as hate speech as offensive and neither.

It also correctly predicted 17 instances as hate speech: 76% were incorrectly labeled as offensive speech; 18% were incorrectly labeled as neither; While some of the following target groups are represented in the same tweet, instances include: 11% African Americans; 28% Female; 28% male with three instances aiming to emasculate men; 22% gay community in general; 5% Asian; and 5 $\%$ White people.

We note that this model improved our recall for hate speech targeting African Americans and White people compared to the combined usage of the hand built features. It would be useful to comment out certain features to see what is reducing performance.

Our model missed 28 instances of hate speech our model incorrectly predicted 39% into the neither class, and 61% of tweets into the offensive class.

Of the 17 instances incorrectly predicted as offensive: 35% female with two of those encouraging suicide; 24% gay community; 24% male with three including threats of sexual violence / general violence; 12% targeted African Americans; and the remaining targeting politics and using the term 'retarded'.

We notice a reduction of overall missed classifications over all groups and a removal of missed hate speech tweets targeting Latinos and Asians. This provides us with a starting point of inspecting the ethnic group feature that focuses on Latinos and Asians (although also African Americans).

Of the 11 instances that were classified in the neither class: 36% African American with one including threat of violence; 27% gay community with one including encouragement of suicide and the rest emasculation of males; the remaining targeting Chinese, and White people.

We notice that the number of missed instances targeting African Americans, White people, and Asians rises showing a bias of this model to classify as neither. It should be noted that there are no missed instances targeting females.

0.4 Hand-Built Features with News, BERT, and Twitter Embedding model Class: Offensive

Out of the 403 Offensive class using the Hand-Built Features with News, BERT, and Twitter Embedding model, we disagree with 45% of the tweets ad believe they are incorrectly labeled, 7% of which our model also incorrectly predicted the label. The remaining 28% of the tweets were correctly predicted by our model.

Of the 28% tweets where we believe our model correctly predicted the label, 68% of the tweets should be labeled as hate speech. Within these, the targeted communities are: 26% gay community; 11 instances where it was used as a means to emasculate men and the remaining targeting the gay community in general

22% African American; this continues to be difficult as variations of 'nigga', 'nigguh', 'nig', 'niglet' are used both as an in-group and by other parties as part of an insult. 15% female; three instances including violence / sexual violence. 14% male; 3 emasculating. 8% used the term 'retarded'; a possible feature capturing instances of [What / He's][a][retard/ed] to weigh them towards the offensive class could reduce the errors. 8% White people. 4% Latinos; an interesting insight as that all instance in the missed tweets mentioning Latinos were either offensive or hate speech. A feature to capture more of these instances would be to decode masked hate words, e.g. 'buck all the beaners', in fact almost all tweets targeting Latinos included a variation of the term beaner so a feature checking for the term along with swear words would properly flag it as hate. The remaining 3% were generally offensive

32% should be labeled as class neither. Our model appropriately captured self-referential statements and usage of terms that were used in a self-affirmative manner, such as:

'RT @kaitlyn_lardi: "@17Seniors: so i basically become a fearless bitch when i'm mad"'
'RT @G0ldenG0ddess: Turn up about to be real , marriott with my bitches for the weekend , mansion tonight , adult swim tomorrow 128131;'

Of the 7% of tweets that were both incorrectly labeled and incorrectly predicted by our model and whose class should be hate speech, there were the following instances targeting: 3 female, 3 African American; 1 Latino; 1 using the term 'retarded'.

Of the missed predictions to true class offensive, the tweets were not targeted and were said as a statement as opposed to an attack. The terms 'hoe', 'pussy', and variations of 'nigga' were common. Making use of our targeted features could help to capture these tweets by toggling the targeted to off.

0.5 News, BERT, and Twitter Embedding model Class: Offensive

Of the 331 missed tweets in the offensive class, we only disagree with the labeling of 8 tweets. Five targeted males, showing a higher bias towards labeling male offensive speech as hate speech. 3 targeted the gay community and 1 was generally racist.

48% of the tweets were incorrectly predicted as neither; about 90% of the missed tweets referenced offensive speech towards women. This provides a case for our hand built feature that checks if a tweet is offensive to women and we look forward to doing additional manipulation of combining our features to improve performance.

52% predicted as hate speech; the overwhelming majority were offensive to women and then African Americans showing a bias our model has towards labeling offensive speech targeting these groups as hate speech. The tweets also included heavy usage of slang, hinting that our slang decoder does help in contextualizing the tweet to capture more information.

0.6 Hand-Built Features with News, BERT, and Twitter Embedding model Class: Neither

26% of the 99 missed tweets in the class neither should have alternate labels. Our model correctly predicted the label for 21% of the tweets with the remaining 5% both wrongly labeled and incorrectly predicted by our model.

Our model incorrectly classified 41% of tweets as offensive. The upside is that not many included slang outside as terms 'nig' and variations thereof, meaning that our slang decoder seemed to help minimizing previous biases that labeled tweets with slang as offensive. The majority of the tweets included pronouns, which may have triggered our pronouns checker feature to label these as offensive. We could fine tune that feature by noting if the tweet is a question, that it may not be offensive. Checking to see if ther tweet is commentary that flips the negative sentiment could be helpful in correctly labeling tweets as neither, e.g. 'RT @MobJoe: Word. And it don't make u a hoe RT @100granHman: It's okay to have sex on first date long as the feeling is mutual'.

More concerning is that our model labeled 32% of the missed tweets in the neither class as hate speech. Many of the tweets combined the term 'trash' with a noun. Capturing the term 'white trash' could down weigh and other instances of the term trash just by itself. The use of the term 'Jihadis' created a strong bias towards labeling the tweet offensive, even when in the context of reporting news. We could create a feature where it searched for the term 'Jihad' along with profanity to differentiate it from the term 'Jihad' just by itself.

0.7 News, BERT, and Twitter Embedding model Class: Neither

We disagree with the labeling of 47% of the missed tweets in the neither class. And of these we believe our model correctly predicts 65% of these tweets and 35% of these tweets were both incorrectly labeled and incorrectly predicted by our model.

Our model correctly identified 19 instances of hate speech targeting: 26% female; 26% male; 21% gay community; 11% African Americans; and the remaining 16% evenly distributed targeting Asians, general racist, and White people.

Of the instances where both the labeling and predictions were incorrect, 7 were hate speech targeting: 36% African Americans; 36% female; 28% gay community

Of our model's incorrect predictions of neither into the offensive class, the model failed to pick up on self-referential and in-group tweets highlighting the need for these hand built features. For the tweets that were incorrectly predicted as hate speech, the NRC emotions feature may help capture more nuanced information about the tone of the tweet.

0.8 Summary

Overall we hope to have highlighted how difficult it is to gauge model performance when dealing with a dataset that is almost 2/3 offensive and with which we feel a great disagreement in the labeling process. We look forward to continuing our work in this space as we have just received academic research API access from Twitter and plan to work on creating new labeled multi-class datasets available to everyone and to test our current and future models. (Davidson et al., 2017) used a Logistic Regression model, as did we. This task would benefit from other models (LSTMs, NNs) being run on larger, more accurately labeled datasets, as Wang (2018) began to explore.

References

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

Cindy Wang. 2018. Lexicon integrated deep neural networks for fine-grained hate speech detection on twitter.