Hate Speech Mining from Human Moderated Online Communities

Kailin Lu
Columbia University
kl2978@columbia.edu

Abstract

Hate speech and toxic speech detection are important for maintaining civil spaces in online forums and platforms. Detection of toxic speech is difficult because it can be highly context dependent, making basic methods such as keyword detection not accurate enough for production purposes. This paper explores the use of extracting moderator removed comments from Reddit as an existing source of labeled data. Using removed comments as a proxy for toxic comments as well as a submission titles as context, a maximum of 78% accuracy in toxic comment detection was achieved.

1 Introduction

Social media platforms today are being responsible for moderating abusive language detection in user comments on a real-time level (Gambäck and Sikdar, 2017). While researchers in natural language processing have been working on this task over the last twenty years, completely satisfactory and automated solutions have yet to be found (Schmidt and Wiegand, 2017).

Over-moderation is a concern for social media platforms, whose users will not be pleased if comments they deem to be not hate speech are preemptively removed. The distinction between hate speech and more common but less serious offensive language has been a rising new area of research (Saleem et al., 2017). While there is no standard definition of what comprises hate speech, most define hateful speech as language that expresses hatred of a person or group on the basis of characteristics such as religion, sexual orientation, race, or other demographic traits (Schmidt and Wiegand, 2017).

Hate speech classification is challenging for a variety of reasons including subjective annotation of datasets, the lack of sufficiently large data, and unreliable keyword starting point approaches. This paper builds on the idea of mining naturally segmented toxic speech from Reddit first presented by Saleem et. al. (2017).

2 Related Work

Offensive speech is often misclassified as hate speech. This is partly because comments that contain offensive words is a strong feature in classifying hate speech which leads to a high false positive rate (Saleem et al., 2017). Part of this is due to subjectivity in human annotators as to what constitutes hate speech. Human coders tend to annotate racist and homophobic language as hate speech, whereas sexist language was more often annotated as offensive (Davidson et al., 2017).

Often, the presence of offensive words leads classifiers to conflate hate speech with generally offensive language. In a Twitter corpus collected using a set of hate speech keywords, Crowdflower annotators only marked 5% of the tweets as hate speech, but only 1.3% of the hate speech tweets were unanimously coded as hate speech (Davidson et al., 2017).

2.1 Supervised Machine Learning

Supervised machine learning algorithms such as naive Bayes, decision trees, random forests, and support vector machines are common classifiers used to classify hate speech, as well as identifying offensive language at large (Burnap and Williams, 2015).

Davidson et. al (2017) achieved a F1 score 0.90 using logistic regression with L2 regularization on a human annotated Twitter corpus. Features that were used in this model included unigrams, bigrams, trigrams, as well as counts of URLs, hashtags, and sentiment score. However, in the error analysis they noted that the classifier tended to miscategorize offensive words as hate speech, but overall categorized a smaller of proportion of tweets as hate speech in comparison to human annotators.

2.2 Deep Learning

Recurrent neural network (RNN) variations including long short term memory (LSTM) networks have also been used to identify abusive language (Pavlopoulos et al., 2017; Sharma et al., 2017). Comparing against vanilla convolutional neural network (CNN) operating on word embeddings, (Pavlopoulos et al., 2017) found RNNs with attention mechanisms to better classify English Wikipedia abusive comments. However, the authors did not test against multiple embeddings such as character level embeddings. Given that classification tasks are known to typically not be competitive across multiple corpuses (Saleem et. al, 2017), it cannot be said whether RNN or CNN methods are superior.

Lack of robustness of the classifier has been a common problem across hate speech and cyberbullying classification tasks. Hosseini et. al (2017) demonstrated how the Google Perspective API, which assigns a toxicity score to comments, could be fooled by simple misspellings or added punctuation in the middle of words to fool the classifier to not recognize offensive language. Keyword based methods have the problem that the identified hate speech keywords need to exist in both the training and test datasets, which Hosseini et. al showed to be easily not the case.

2.3 Community-Defined Speech

(Saleem et al., 2017) proposed a method of data collection that does not involve human coders. They introduce and show the existence of ”community-defined speech” where self-selected communities exhibit distinct speech patterns. Rather than data mine a corpus and have human annotators mark hate speech, Saleem et. al (2017) introduce the concept of community-defined speech. On Reddit there exist subreddit communities that are both in support of a group as well as specifically hateful towards a target group. By extracting comments from subreddits in support and openly hating against a target group, Saleem et. al. show that human annotated corpuses are not necessarily when there are examples of self-identified groups exhibiting unique language patterns.

This approach also addresses the dataset size challenge, since (Saleem et al., 2017) chose popular subreddits. The authors use three machine learning classification techniques: naive Bayes, support vector machine, and logistic regression. Additionally, they show that the classifier performs well on non-Reddit sites such as the forum site Voat, achieving accuracies of over 60 %.

However, they do note that Voat is similar in format compared to Reddit. One downside of this approach is that these hate-identified subreddits may not be around for training perpetually. Since publication, Reddit has removed two of the three hate identified subreddits used in this study. This project will be an extension on the ideas of this work by first showing a proof of concept using removed comments rather than opposing subreddits and then extending the classification techniques to deep learning methods.

3 Project Goals

The goal of the project will be to:

1. Create a dataset of hate speech by retrieving comments removed for abuse on Reddit

2. Develop a model of hate speech and compare patterns across different subreddit communities.

This project will examine hate language targeted towards women and the LGBT community through using reported comments made in two prominent support subreddits: r/TwoXChromosomes and r/lgbt. Each of these subreddits has a list of community guidelines. Users can report comments that violate the subreddit rules. Moderators review reported comments and make the decision on whether or not to remove posts and comments. Additionally, moderators can themselves remove comments without a report.

It is not expected that all remove comments will fall under the category of hate speech. Therefore, the first step of this project will be to do a topic analysis using nonnegative matrix factorization (NMF) (Berry et al., 2007) to explore whether hateful comments end up within their own topic. Other clustering methods such as k-means will be tried as well. Once a suitable cluster of comments containing hate speech is manually identified, these comments will serve as the positive portion of the training set with a random set of non-reported comments serving as the negative set.

The following step will be to develop a classifier, trying both machine learning and deep learning methods. Since this method of training data generation does not naturally create a test set, labeled Reddit corpora of hate speech will be used as a test set.

4 Data Collection

The Reddit API `https://api.reddit.com` can be used to query specific subreddits, post content, and comment content. Comments on Reddit can be removed in one of two ways: the user can delete their own comment or a subreddit moderator can remove a comment as inappropriate. Data on the comment and submission body will reflect the current status of the post. For example, if a post has been deleted by the user, the comment body will read `[deleted]` whereas if a post has been removed by a moderator then the comment body will read `[removed]`. In other scenarios where the comment still exists on the Reddit platform, the comment body will read as the text stored at the time of querying, which prevents scrapers from tracking comment edits in a straightforward manner.

As the Reddit API reflects the status of comments and posts at the time of query, removed and deleted comment text are not available. However, Reddit user `u/Stuck_In_The_Matrix` has lead development of Pushshift (Baumgartner, 2017) which queries the Reddit API at frequent, regular intervals. The content of some removed comments can then be retrieved by first querying the Reddit API for removed comment IDs and then comparing the comment body of those IDs from Pushshift.

Subreddit	**lgbt**	**TwoX**
Submissions	1730	6750
All Comments	35,499	360,480
Removed Comments	1,603	15,624
Archived Removed Comments	1,603	1,603

Table 1: Count of submissions and top level comments retrieved from 2014 - 2017.

4.1 Comment Moderation on Reddit

Subreddits are each governed by a set of community rules and posting guidelines. The role of subreddit moderators is to monitor submissions for posts or comments which do not follow the guidelines and remove them. For the majority of subreddits, the first rule is to follow the overarching Reddit community guidelines which is to not post toxic or hateful language. However, other examples of rules include staying on the topic of the subreddit or requiring that users follow a particular style of posting. For this reason, we do not expect all removed comments to be examples of toxic language.

4.2 Dataset

In this paper we examine comments from two subreddits which might attract hate speech r/lgbt and r/TwoXChromosomes. r/lgbt is a subreddit meant for discussion and support of the LGBT community. r/TwoXChromosomes is a similar subreddit targeted at women's rights and equality issues. Submissions and top level comments from the period January 1, 2014 to January 1, 2017 were scraped from both subreddits. The total number of items scraped can be found in Table 1.

The proportion of removed comments whose text was able to be retrieved from Pushshift was less than 1% in r/lgbt and less than 2% in r/TwoXChromosomes. These proportions are consistent with how frequently hate speech is found in a corpus in previous works.

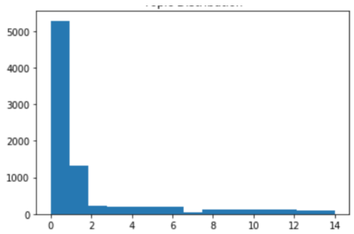
4.3 Topic Modeling for Toxic Language Identification

In order to develop a set of examples of toxic language from each subreddit, the removed comments need to be separated into categories of truly toxic versus comments which merely do not follow the subreddit's other rules. Topics were assigned to comments using nonnegative matrix factorization. A term-document matrix is factorized into a topic-document matrix multiplied with a topic-word matrix. The number of topics to categorize the comments into is predetermined.

TwoXChromosomes has four community rules: 1) Respect, 2) Grace, 3) Equanimity, and 4) Relevance. As such, the first attempt at a topic classification was to use only four topics. Qualitative inspection of the resulting topics showed that comments which fell under each of the four topics were not easily distinguishable. After several iterations of using different numbers of topics, fifteen was chosen as the number of topics to segment the comments.

The distribution of number of comments which fell under each topic is shown in Figure 1.

Figure 1: Topic distribution of removed r/TwoXChromosomes comments

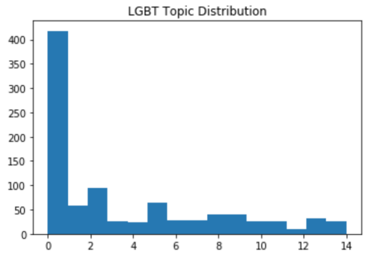


The majority of comments were categorized into Topic 0. Manual inspection of this topic showed no clear theme so comments within this topic were labeled as non-toxic and non-relevant. Examples of other topic within removed r/TwoXChromosomes comments include lack of sensitivity around abortion, judgemental sentiments of sexual behavior or partners, and comments around comparisons of men and women. All comments in topics besides topic 0 were labeled as toxic for women's issues.

In r/lgbt, there are only two rules for comments and submissions: 1) No homo or trans phobia, and 2) Must be open and have willingness to learn. The removed comments were again categorized into fifteen topics.

There is a similar pattern where the majority of removed comments fell into the first topic. Unlike the most common topic in r/TwoXChromosomes, this first topic was not a 'not relevant' cluster but rather a cluster that contained general expressions of anger. Other topics which emerged include support for Donald Trump, statements relating homosexuality to mental illness, and disrespectful sentiments towards the LGBT community. All of the removed comments from r/lgbt were kept as examples of toxic comments.

Figure 2: Topic distribution of removed r/lgbt comments



5 Experiments

Comments from the toxic class were labeled as part of the topic model. An equal number of comments were sampled from the non-toxic comments in each subreddit to form the negative example clas

The training data is split into a 80% train set and a 10% validation set. For this project, classification of toxic language in each of the subreddits is kept as separate tasks. In future work, combining toxic language data collected by topic into one classifier is worth exploring.

5.1 Linear Models

The first classification models which were tested were support vector machine (SVM) and logistic regression using L1 regularization. The words in the comments were transformed into feature vectors using term frequency-inverse document frequency (Tf-Idf) weighting. Common English stop words provided by Python scikit-learn library were removed from the corpus.

Examining incorrectly classified comments, we find that false positives were generated from the use of offensive language. While outrage sentiment

Subreddit	lgbt	TwoX
SVM	.57	.77
Logistic	.59	.75

Table 2: Test Set Accuracy of SVM and Logistic Regression

Subreddit	TwoX
No Context	.75
With Context	.78

Table 3: Test Set Accuracy of LSTM Models

might have been expressed in these comments, it was the outrage in support of the post rather than a toxic reaction against the post. Shorter comments which did not contain generally offensive language also tended to be misclassified as nontoxic. Both of these problems can be explored in future work where more training examples can be mined over more subjects and a longer time period.

5.2 Deep Models and Comment Context

In addition to linear models, a long short term memory network (LSTM) was used to classify comments within r/TwoXChromosomes. Two versions of the network were used. One which simply read the words from the comments themselves, and a second which read both in both the comments and the submission post which is used as context information. The context-LSTM is based off a model for sarcasm detection (Ghosh et al., 2017).

The no-context LSTM architecture included a one layer LSTM with 128 hidden units followed by a softmax classification layer. The Adam optimizer was used with a 1e-5 learning rate. Word embeddings were trained as part of the model. In the LSTM with context, separate LSTM layers with 128 hidden units each were used to summarize the post title and the comments. The results were then concatenated into a single tensor which was then fed into the softmax classification. This model also used the Adam optimizer with a 1e-5 learning rate.

The LSTM without context performed at roughly the same accuracy as the linear models whereas the LSTM with the context provided performed slightly better. It is possible that the LSTM models did not significantly improve classification accuracy over linear models due to insufficient sample size.

Both LSTM models did not provide a significant improvement over linear models. However since the LSTM which contained context information performed the best over all of the models, there is an indication that understanding the context of a conversation is necessary to better identifying toxic or hateful speech within a conversation.

6 Conclusions

Identifying and removing hateful speech in online forums is an ongoing challenge for social media companies. The presence of toxic language leads to a harmful environment that these companies do not wish to foster. However, over moderation of comments people find to not actually be toxic can lead to user frustration with the platform. Classifying toxic language has been a challenge in the past due to limited amounts of labeled training data and the inefficient and likely expensive methods of traditional methods of labeling through human annotation.

In this paper, the use of moderator removed comments in two Reddit subreddit communities is studied as a possible data source for toxic comment classification. What the results have identified is that comments which have been removed and seem qualitatively toxic can be identified through automated methods. Furthermore, including information about the context of the comment such as a post title can improve classification accuracy. However, the models are still confusing generally offensive words with toxic intent and also exhibit difficulty in classifying shorter comments.

7 Future Work

A natural continuation of the study of moderator removed comments is to use classification models trained on moderator removed comments on existing datasets of toxic language. The first step would be to do a direct comparison with forums that discuss similar topics such as gender or sexuality issues. Combining comments from focused topics into a more general classifier would also be an interesting area to pursue. Further work could be done on studying how using user moderated comments may or may not lead to benefits in time-relevant toxic comment classification.

Acknowledgments

I would like to thank Professor Smaranda Muresan for the opportunity to do this project and the continued feedback throughout the process.

References

Jason Baumgartner. 2017. Pushshift.

Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. *arXiv preprint arXiv:1707.06226*.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993*.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Hitesh Kumar Sharma, TP Singh, K Kshitiz, Harsimran Singh, and Prince Kukreja. 2017. Detecting hate speech and insults on social commentary using nlp and machine learning.