Udacity

Machine Learning Nanodegree Capstone Project

Identifying and classifying toxic and abusive text online

Neil Mistry
4-8-2018

Table of Contents

Problem Definition

Overview

This project aims to deliver a machine learning algorithm that can effectively identify and classify abusive or toxic language used online. It is inspired from a competition hosted on Kaggle by The Conversation AI team, a research initiative founded by Jigsaw and Google.

Many platforms online struggle to effectively facilitate public conversation. It is apparent if you are to use Facebook, Wikipedia, online forums, or other social media tools. The internet allows for anonymity and distance if one desires, which can encourage some people to use these mediums to express hate, negativity, or hurl abuses. There are numerous publically available models out (available through Perspective API) which are designed to identify and classify negative comments. But these current models still make errors.

In this competition, participants are challenged to build a multi-headed model that's capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate better than Perspective's current models.

Data Inputs

For this competition, datasets are being provided from Wikipedia's database. These comments are from users who submit comments to edits on Wikipedia's publically managed content. There is a training set and a testing set provided (with no testing labels). The training set contains 159,571 comments which have been classified to contain or not contain comments that are: toxic, severely toxic, obscenity, threats, insults, or identity hate. Each comment can have multiple positive labels (positive is represented by 1, negative by 0). Some comments are blank. The first 10 entries are shown below (**vulgar language warning**):

id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27	0	0	0	0	0	0
000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)	0	0	0	0	0	0
000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info.	0	0	0	0	0	0
0001b41b1c6bb37e	More I can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of accidents"" -I think the references may need tidying so that they are all in the exact same format ie date format etc. I can do that later on, if no-one else does first - if you have any preferences for formatting style on references or want to	0	0	0	0	0	0
0001d958c54c6e35	You, sir, are my hero. Any chance you remember what page that's on?	0	0	0	0	0	0
00025465d4725e87	Congratulations from me as well, use the tools well. Â Â· talk "	0	0	0	0	0	0
0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
00031b1e95af7921	Your vandalism to the Matt Shirvington article has been reverted. Please don't do it again, or you will be banned.	0	0	0	0	0	0
00037261f536c51d	Sorry if the word 'nonsense' was offensive to you. Anyway, I'm not intending to write anything in the article(wow they would jump on me for vandalism), I'm merely requesting that it be more encyclopedic so one can use it for school as a reference. I have been to the selective breeding page but it's almost a stub. It points to 'animal breeding' which is a short messy article	0	0	0	0	0	0
00040093b2687caa	alignment on this subject and which are contrary to those of DuLithgow	0	0	0	0	0	0

Data provided by Conversation AI team from Wikipedia user comments

The testing set contains 153,164 comments without any labels. Participants of the competition may submit a set of testing labels which will be immediately scored based by mean column-wise ROC AUC (Area under receiver operator curve). ROC curves are made by plotting a true-positive vs false positive for each type of classification (toxic, obscene, etc).

Problem Statement

When a user submits a comment or post in an online medium (such as Facebook, Wikipedia, or a blog) the medium owners would like to know if this comment contains any negative connotation such as a threat, obscenity, insult and identity-based hate. If it does, the medium owner may elect to prohibit the content being submitted, flag it for manual review before making the content visible, or give the user a warning before accepting the submission.

It is important that we can capture as much negative content at the source, so that it does not cause a toxic effect on these mediums. Therefore, getting close to 100% true positives for negative comments is ideal. However, those who are submitting content to these mediums would be less inclined to keep using these mediums if there are too many false positives. That is, if those who are submitting their content and not submitting negative comments, would have their content rejected, they will not be satisfied with using these mediums.

The proposed solution is to build a model that will classify the comments based on its content (words) if it represents any toxic, severely toxic, obscene, threatening, insulting or identity hate based

content. The model will be able to clearly identify each of these classifications and return a positive or negative flag indicating the presence of these types of content in each comment. The model will be able to flag for multiple content types for each comment. We will build the model using the training data provided, and set aside part of the training data (20%) for our own validation before submitting our test set (using predictions) to the competition.

Metrics

Since the competition uses mean-wise ROC AUC scoring, we will use that for the purpose of scoring our model. Our initial benchmark using a Naïve-Bayes implementation on a bag of words feature set resulted in a mean-wise ROC AUC score of 0.6981. I aim to achieve a score in the 0.98 region as obtained by the leaders on the competition leaderboard.

For the purpose of testing our models before submission, to ensure changes we make result in an improvement, we will use accuracy, precision, recall and the F1 score for each classification label. The Naïve-Bayes implementation presented in 'Benchmark Models' will serve as the benchmark statistics.

ROC AUC is calculated as follows:

$$AUC = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} 1\, p_i > p_j$$

Where m is all data points with a true label of 1 and n are all data points with a true label of 0, for each classification. Pi and Pj denote the probability score assigned by the classifier to data point 'i' and 'j' respectively. 1 is the function, the function outputs a 1 if the condition is met (pi > pj) [1]. AUC will be calculate for each classification (toxic, obscene, etc) and all the scores will be added for the final score.

Problem Analysis

Data Exploration and Visualization

Let us take a look our training dataset below where we have the first 10 entries.

id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27	0	0	0	0	0	0

000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)	0	0	0	0	0	0
000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info.	0	0	0	0	0	0
0001b41b1c6bb37e	" More I can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of accidents"" -I think the references may need tidying so that they are all in the exact same format ie date format etc. I can do that later on, if no-one else does first - if you have any preferences for formatting style on references or want to do it yourself please let me know. There appears to be a backlog on articles for review so I guess there may be a delay until a reviewer turns up. It's listed in the relevant form eg Wikipedia:Good_article_nominations#Transport "	0	0	0	0	0	0
0001d958c54c6e35	You, sir, are my hero. Any chance you remember what page that's on?	0	0	0	0	0	0
00025465d4725e87	" Congratulations from me as well, use the tools well. Â Â· talk "	0	0	0	0	0	0
0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
00031b1e95af7921	Your vandalism to the Matt Shirvington article has been reverted. Please don't do it again, or you will be banned.	0	0	0	0	0	0
00037261f536c51d	Sorry if the word 'nonsense' was offensive to you. Anyway, I'm not intending to write anything in the article(wow they would jump on me for vandalism), I'm merely requesting that it be more encyclopedic so one can use it for school as a reference. I have been to the selective breeding page but it's almost a stub. It points to 'animal breeding' which is a short messy article that gives you no info. There must be someone around with expertise in eugenics? 93.161.107.169	0	0	0	0	0	0
00040093b2687caa	alignment on this subject and which are contrary to those of DuLithgow	0	0	0	0	0	0

Each data point has an ID (as an index), a user comment, and 6 classification labels. Each classification label has a 0 (if the comment does not fall within the class) or 1 (if the comment does fall within the classification.

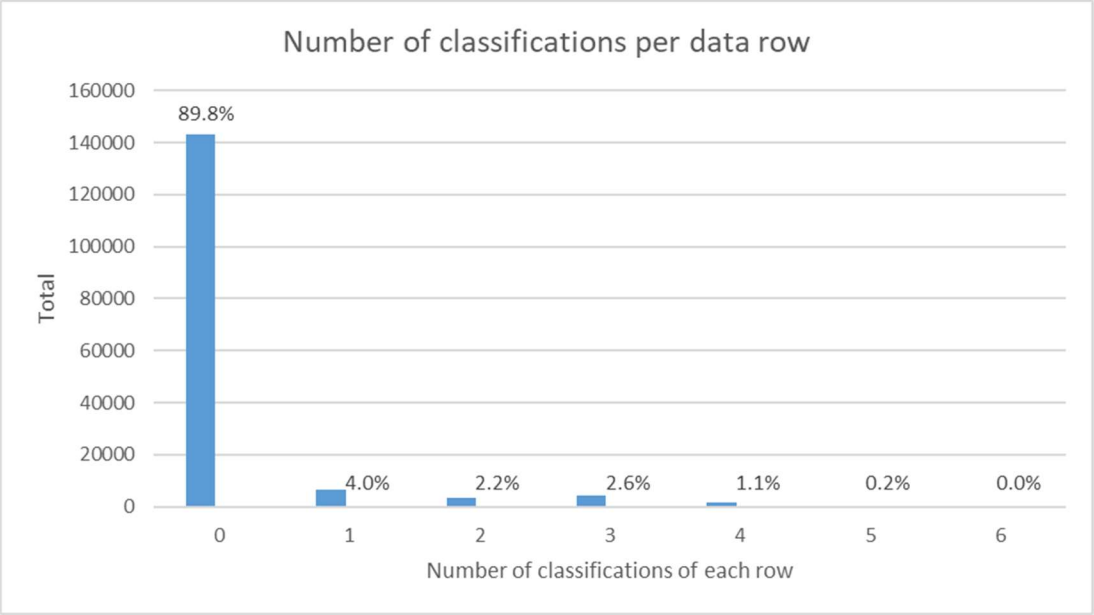
The first comment is:

Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27

We can see there is not toxic/vulgar/insulting content in this comment. Now let us look at comment 7,

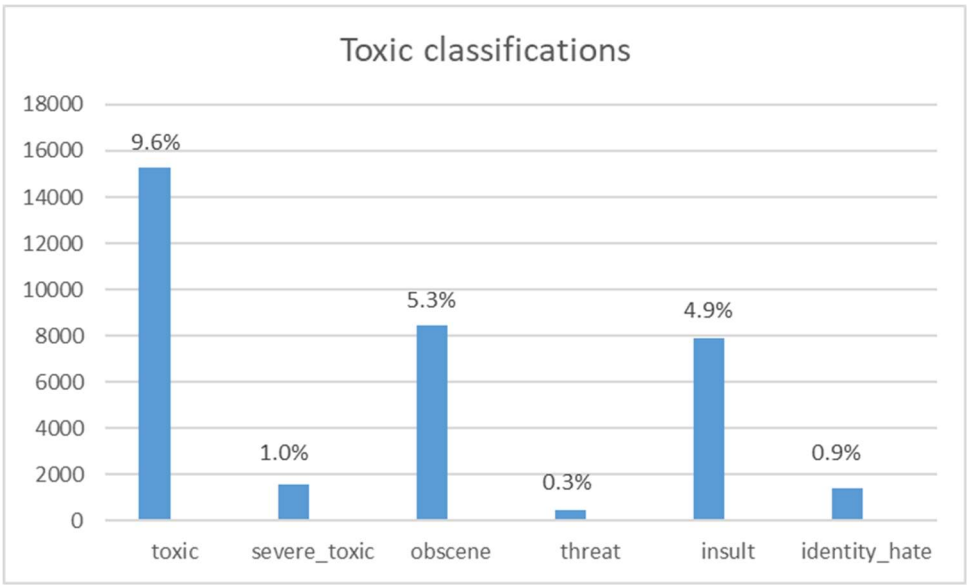
COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK

This comment is classified as being toxic, severe toxic, obscene, and an insult. We can see this comment clearly contains toxic/obscene/insulting content. We have summarized the total classifications below:

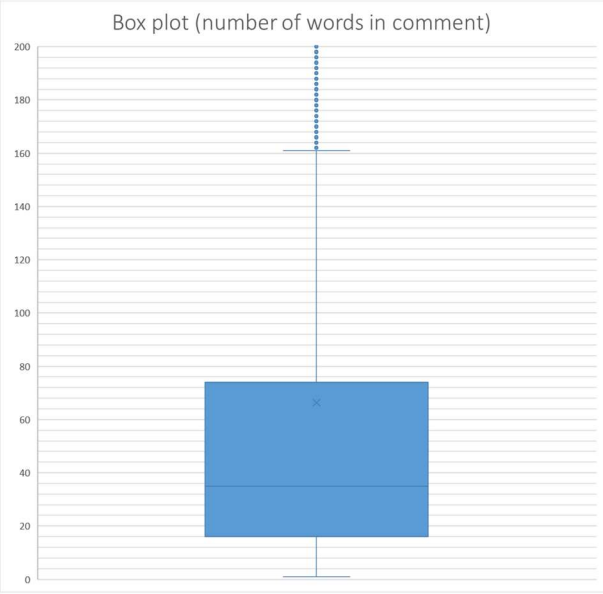


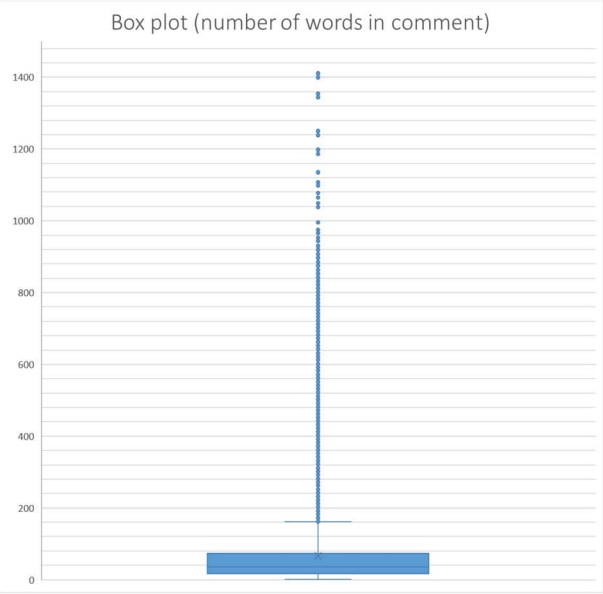
We can see that almost 90% of the comments do not fall into any toxic classification (and we consider these comments as clean comments). 4% of the comments are positively classified under 1 of the toxic categories, and 1 – 3% of the data are each classified as having between 2 and 4 toxic classifications. There could be a relationship between some of the classifications that we could explore further if needed.

Let us also consider which classifications are occurring in our comments. Below is the total of each classification in our training data and as a percentage of the total. We can see that the majority of our classifications are captured by the categories of 'toxic', 'obscene' and 'insult'. The other three categories altogether represent less than 3% of the comments. We have a risk of having an unbalanced model that will be highly adept to accurately detecting 'toxic', 'obscene' and 'insult' comments and not detecting the other three very well. We will want to pay close attention to the performance of each type of classification.



Next we will consider the length of each comment. Below we have a box plot of the number of words (we consider a word as any set of characters with a space separation) in each comment. We can see there is significant variation in comment length, from as low as 1 comment to over 1400 words. 50% of our comments are between 16 and 74 words. Our median comment length is 35 words.





Algorithms and Techniques and Benchmarks

Current models created by Conversation AI team, use CNN (convolutional neural networks) models (Keras) in python. An example is provided. Based on the results in the Kaggle competition, we

can see benchmark models with over 0.98 score using the mean-wise ROC AUC scoring method. These models use CNN, often with a pre-trained word classifier.

Prior to starting this project, I have created a simple Naïve-Bayes model using the dataset provided on Kaggle. This model takes the 80% of the training data, applies a bag of words algorithm prior to training on the Naïve-Bayes algorithm in python. I used the remaining 20% to score the classifier which resulted in these metrics:

```
toxic
('Accuracy score: ', '0.948394172019')
('Precision score: ', '0.80967032967')
('Recall score: ', '0.602748691099')
('F1 score: ', '0.691052335397')
severe_toxic
('Accuracy score: ', '0.988688704371')
('Precision score: ', '0.437888198758')
('Recall score: ', '0.439252336449')
('F1 score: ', '0.438569206843')
obscene
('Accuracy score: ', '0.969230769231')
('Precision score: ', '0.782140107775')
('Recall score: ', '0.592419825073')
('F1 score: ', '0.674187126742')
threat
('Accuracy score: ', '0.997211342629')
('Precision score: ', '0.222222222222')
('Recall score: ', '0.0810810810811')
('F1 score: ', '0.118811881188')
insult
('Accuracy score: ', '0.963684787717')
('Precision score: ', '0.691983122363')
('Recall score: ', '0.508054522924')
('F1 score: ', '0.585923544123')
identity_hate
('Accuracy score: ', '0.989409368635')
('Precision score: ', '0.319672131148')
('Recall score: ', '0.132653061224')
('F1 score: ', '0.1875')
```

This model also returned a 0.6981 mean-wise ROC AUC score on the kaggle leaderboard. This will set the benchmark for our model. We aim to improve upon the 0.6981 score and come to the 0.98 region using CNN.

Methodology

Data Preprocessing

In text based problems we often use each word or a group of words as a feature. This can create a problem because we can have many unique words (features) and could slow down our computation. We will often want to remove words that occur to frequently, or very infrequently. As words that occur too frequently could very well be irrelevant in our classification problem (consider that only about 10% of our comments are toxic, if we have a word that occurs a lot more than 10% it may not

be an important feature for classification). Similarly words that occur infrequently may also be irrelevant when creating a model as there may be no relation between that feature and a classification if a classification occurs more frequently than that feature.

We also want to move irrelevant information, such as punctuation and possibly numbers. As well as remove any common words such as “the” and “and”. We refer to these types of words as “stop-words”. Additionally, it would be ideal to correct misspelled word and group short form words and acronyms with their fully spelled words. However, this can be quite difficult to do.

We will try to address the first four issues by using the CountVectorizer method in sklearn. CountVectorizer will return us a sparse array with each word having a feature representation. The feature will have a number associated to the number of times that feature (word) occurs in a sentence. Additionally, we will remove words that occur 3 times or less (consider that a comment classified as a ‘threat’ occurred 478 times) and remove words that occur in more than 30% of the comments (consider that only about 10% of comments had a toxic classification).

Implementation

We have begun by using the CountVectorizer method on our training data. This will create a feature set we can use to train our naïve-bayes models. Since the Sklearn method for Naïve-Bayes can only fit input data to classify 1 label, we will use 6 independent models to predict each classification. We will train with all our training data provided by Kaggle and submit a set of predictions to Kaggle using the test data provided.

Our results on fitting the training set is as follows (we have reserved 20% of training data for testing):

toxic Accuracy score: 0.94 Precision score: 0.72 Recall score: 0.73 F1 score: 0.72	**obscene** Accuracy score: 0.97 Precision score: 0.66 Recall score: 0.76 F1 score: 0.71	**threat** Accuracy score: 0.99 Precision score: 0.11 Recall score: 0.18 F1 score: 0.14
insult Accuracy score: 0.96 Precision score: 0.60 Recall score: 0.69 F1 score: 0.63	**identity_hate** Accuracy score: 0.98 Precision score: 0.26 Recall score: 0.37 F1 score: 0.31	**severe_toxic** Accuracy score: 0.98 Precision score: 0.35 Recall score: 0.66 F1 score: 0.46

This resulted in an ROC AUC score of 0.6981.

Refinement

We can see that although our scores for toxic, obscene and insult were good for a simple model, our scores for threat, identity_date and severe_toxic were considerably lower. We can attribute this to the training data having much fewer examples of these classifications. Recall from our analysis earlier the training data only have these three classifications occur in less than 3% of the comments.

One thing we can do to improve our results is expand our training set. Since we are primarily concerned about the ROC AUC score (as that is how the competition is scored), we will use our entire training data to train our model. We can then submit our predictions to Kaggle using the testing set. Doing this resulted in a drastic improvement. Our score increased from 0.6981 ROC AUC to 0.7808.

Results

Model Evaluation and Validation

Do to the nature of the competition, we do not have access to the testing labels in order to fully evaluate our final model. From the Kaggle competition we can retrieve a ROC AUC score which was 0.7808. If we reduce our training size to allocate some of our data to testing (20%) we can clearly see the model performs very well at classifying 'toxic', 'obscene' and 'insult' comments. We lack data to improve our model performance for the other classifications. But we have shown the effectiveness of the model for this type of problem and am confident that with more data the model will be effective with all categories.

Justification

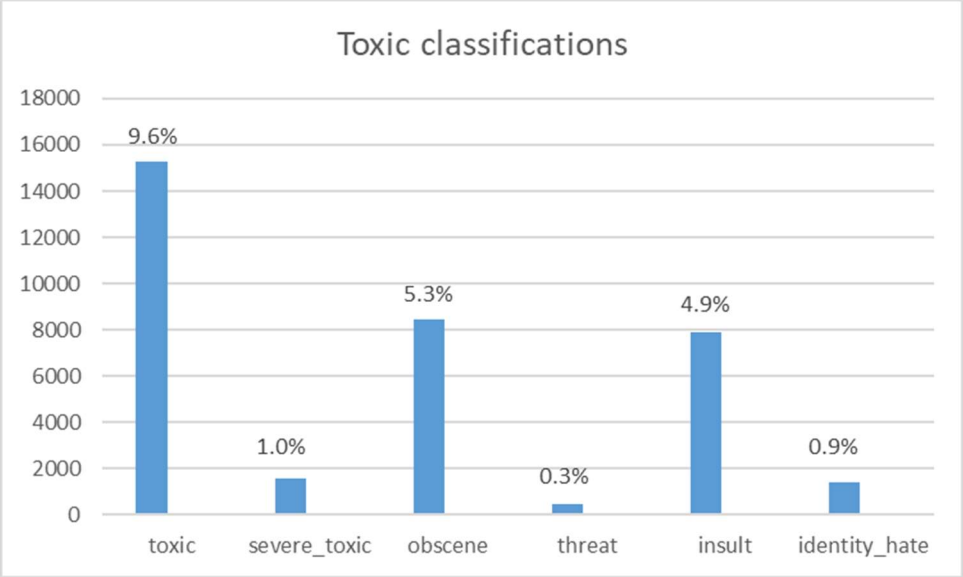
Compared to our original benchmarks we have improved almost 10%. Infact, we are on par with CNN models in performance used in this competition. Models that can outperform the naïve-bayes solution are using pre-trained CNN models.

Conclusion

There exists many mediums that allow public users to comment and provide feedback on online content. As the public nature of these mediums allow anonymous feedback, we can expect a high level of negative feedback on these mediums. There exists a real problem in filtering the negative and unnessary content from these mediums. We have shown that a simple naïve-bayes supervised model

can effectively classify content to determine if that content contains toxicity. Our model can effectively be deployed to flag these comments. Additionally a simple feedback loop, for example a moderator can add additional data into the model for retraining. If a moderator is to flag a comment as being toxic, and specifically belonging to one of the categories of toxic classified in this project, that information can further be used to retrain the model.

Currently, we are restricted in model performance for three categories of toxic classification: severe_toxic, threat, and identity_hate, as we have less than 4000 of these samples to train our model. Whereas we have over 15,000 samples of toxic comments to train our toxic classifier.



References

[1] "Calculating ROC curves and AUC scores" by 'Institute for Computing and Information Sciences' at Radboud University (Nijmegen, Netherlands), http://www.cs.ru.nl/~tomh/onderwijs/dm/dm_files/roc_auc.pdf