Udacity

Machine Learning Nanodegree Capstone Project Proposal

Identifying and classifying toxic and abusive text online

Neil Mistry

3-1-2018

Project Domain

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.

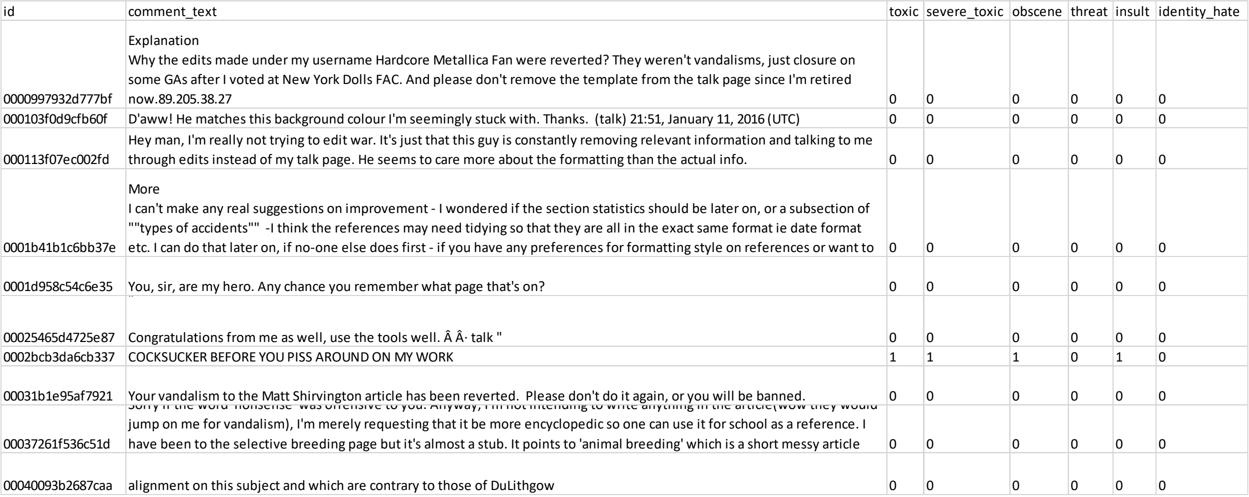
Problem Statement

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.

Datasets and 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).

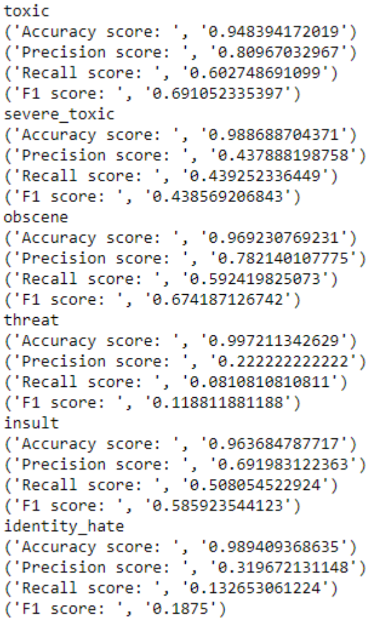
Solution Statement

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 testing before submitting our test set (using predictions) to the competition.

Benchmark Models

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.

Evaluation 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.

Project Design

I will begin by exploring the training data to understand how many of each type of classification exist. This will serve to understand if there would exist any model in-balance in detecting each type of toxic content. I will then remove any punctuation and stop-words (if, and, or, etc). The remaining words will then be converted into a count-array using the CountVectorizer method from Scikit-learn. We will then remove any words that occur extremely in-frequently, or occur too frequently (we can set the thresholds based on the number of toxic classified words we discover in our data exploration). The remaining table will serve as our feature set to submit to any model.

80% of the training data will be used to train our model. I will score each model using 20% of the training set using the predictions from the classification model. I will compare results from each model and model tweak using accuracy, precision, recall and the f1-score. If we see poor performance on some of the classification types due to model in-balance, I can then train each classification label separately.

The highest performing model will be then trained on 100% of the training set. This model will then make predictions using the testing set provided. The provided testing set will be submitted to the Kaggle competition to obtain the mean-wise ROC AUC score.