Machine Learning Engineer Nanodegree

Capstone Proposal

Kuan - Han Chen
July 15th, 2018

Proposal

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

Domain Background

The Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet) are working on tools to help improve online conversation. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far they've built a range of publicly available models served through the Perspective API, including toxicity. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content).

Problem Statement

In this competition, you're 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. You'll be using a dataset of comments from Wikipedia's talk page edits. Improvements to the current model will hopefully help online discussion become more productive and respectful.

Datasets and Inputs

The datasets are provided by Jigsaw and Google for competeion in Kaggle. Them contain training set and testing set.

- train.csv(160k rows x 8 columns) - the training set, contains 8 columns, namely id, comment_text, toxic, severe_toxic, obscene, threat, insult and identity_hate, and last six columns comments with their binary labels
- test.csv(153k rows x columns) - the test set, contains 2 columns, namely id and comment_text .You must predict the toxicity probabilities for these comments. To deter hand labeling, the test set contains some comments which are not included in scoring.
- test_labels.csv(153k rows x 7 columns) - labels for the test data; value of -1 indicates it was not used for scoring. It contains 7 columns, namely id, toxic, severe_toxic, obscene, threat, insult and identity_hate.
-

Note: I am not sure if the training set is imbalanced dataset. If it is, I will take following approach, copy the few materials to make it equivalent to the majority of the data.

Solution Statement

I will use Text-CNN as the main solution algorithm. Text-CNN has a good performance in natural language processing. The Text-CNN algorithm was published in this paper, Convolutional Neural Networks for Sentence Classification

Benchmark Model

I will use naive bayes as my benchmark model. And get a baseline score through naive bayes model for my dataset, and compare with Text-CNN, which algorithm has higher AUC, ROC and accuracy.

Evaluation Metrics

- AUC: In signal detection theory, a receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied.
- ROC: The AUC value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example.
- Accuracy: How many number of correct prediction in total number of correct predictions.

Project Design

The general sequence step are as follows:

1. Loading the Data
2. Pre-processing the Input
    - tokenization
    - remove stopwords
    - Stemming and lemmatization
    - Transforming Comments to Sequences
3. Training Model
    - Naive bayes
    - Text-CNN
4. Make the predections
5. Result Analysis

Reference

1. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge#description
2. http://www.aclweb.org/anthology/D14-1181
3. http://www.jeyzhang.com/cnn-apply-on-modelling-sentence.html