IST Data Mining II

TERM PROJECT PROPOSAL

Toxic Comment Classification
Identify and Classify Toxic Online Comments

Prepared by:

QING GONG

CHU WANG

February 11, 2018

1. Introduction

Nowadays, it has been unprecedentedly convenient for humans to express themselves and broadcast ideas online benefitting from the tremendous development of internet technology. People therefore can publish toxic comments and even assault people online immorally because it is fairly difficult to identify such people and online accountability system is still not mature. However, toxic comments have quite bad influence on the online communities. People might stop expressing themselves and give up on seeking different opinions when they receive abuse and harassment. It has been a huge problem for many platform, such electric merchandise website, academic forum website, and political opinion community. Thus, it is of importance to effectively facilitate conversations by classify toxic comments and identify abuse and harassment.

In order to effective solve the issue or improve the online conversation environment, many research institutes are working on tools to help improve online conversation. For example, there is a competition on Kaggle regarding this issue: "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). 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.”

The point of departure for this project is the competition on Kaggle. This project is aimed to compete on this topic with learning relative methods and equip us on natural language process.

2. Data Description

A large number of Wikipedia comments is provided which have been labeled by human raters for toxic behavior. The types of toxicity are:

toxic

severe_toxic

obscene

threat

insult

identity_hate.

The data has two sets, training set and test set. The training set contains comments with their binary label, and the test set is to predict the toxicity probabilities for these comments. The deter hand labeling, the test set contains come comments which are not included in scoring.

For example, there is one comment from the sample, “COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK”. It is identified as toxic, severe_toxic, obscene, and insult. “Don’t mean to bother you”, however, is not toxic comment at all.

3. Key Algorithm/Technology

From the public kernels shared on Kaggle, there are two strong baselines using NBSVM method and LSTM method separately. We plan to try them and improve them.

Idea1: how to detect misspellings in data, for example:

"::Yeah, that sounds more like WP:-) than ""f#*$ing c*^$."" I got a good laugh out of it. "

Idea2: classes are not independent, eg: class ‘severe_toxic’ must belongs to ‘toxic’

4. Expectation

A model is created to predict a probability of each type of toxicity for each comment. For each id in the test set, the model is able to predict a probability for each of the six possible types of comment toxicity(toxic, severe_toxic, obscene, threat, insult, identity_hate). The columns will be in the same orfer as shown below. The file will contain a header and have the following format:

id,toxic,severe_toxic,obscene,threat,insult,identity_hate

00001cee341fdb12,0.5,0.5,0.5,0.5,0.5,0.5

0000247867823ef7,0.5,0.5,0.5,0.5,0.5,0.5

etc.

The score of submissions are now evaluated on the mean column-wise ROC AUC. In other words, the score is the average of the individual AUCs of each predicted column.