MACHINE LEARNING CAPSTONE PROPOSAL

Title: Toxic Comment Classification

Domain Background

This project is about online comments classification, which is based on a public Kaggle competition "Toxic Comment Classification Challenge". Online discussion is everywhere, however, it can be difficult. According to 2016 Global Report on Online Commenting, there is a trend that high-profile news organizations (including NPR, CNN, The Verge, Toronto Star, Reuters and Popular Science) shutting down comment sections due to the abusive tone and poor quality of comments.

To combat such problem, the Conversation AI team, a research initiative founded by Jigsaw and Google, are working on tools to help improve online conversation. One of the studies focuses on the toxicity of online comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion).

Problem Statement

The current models by the Conversation AI 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). The goal of this project is to build a multi-headed model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate. Such improvements to the current model will hopefully help online discussion become more productive and respectful.

Datasets and Inputs

The dataset contains about 160k human labeled comments from Wikipedia Talk pages. The labeled annotations are obtained by asking 5000 crowd-workers to rate Wikipedia comments according to their toxicity (likely to make others leave the conversation).

The dataset has 159571 text instances with 6 attributes (1 for toxic, 0 for non-toxic). The types of toxicity are:

- toxic
- severe_toxic
- obscene
- threat
- insult
- identity_hate

Solution Statement

To classify the types of toxic comments is in the domain of Natural Language Processing (NLP). Deep Learning methods are starting to out-compete the classical and statistical methods on some challenging NLP problems with singular and simpler models. In this project, deep learning along with other NLP techniques will be used to train a convolutional neural network to classify the toxic comments to their respective classes.

Benchmark Model

A Logistic Regression model has been trained on the dataset and tihs yields 0.9722 in the submission leaderboard.

Evaluation Metrics

For each id in the test set, a probability for each of the six possible types of comment toxicity (toxic, severe_toxic, obscene, threat, insult, identity_hate) will be predicted. The models are now evaluated on the mean column-wise ROC AUC. In other words, the score is the average of the individual area under the ROC curves of each predicted column.

Project Design

Programming Llanguage: Python 3.5
Libraries : Keras, TensorFlow, Gensim, pandas, NumPy

Workflow:

- Download and inspect the dataset
- Clean and tokenize the text and save the results to a new file
- Establish logistic regression ad benchmark model
- Train a bag-of-words model as comparison
- Develop a word embedding layer (Word2Vec) + CNN model
- Develop a pre-trained embedding layer (GloVe) + CNN model
- Evaluate the models and find out the best algorithm for this project
- Fine tune the selected model to increase performance