Springboard--DSC Program

Capstone Project 1 - Proposal

Toxicity Classification

By: Ellen A. Savoye

December 16, 2019

Business problem:
Openly discussing things that you feel strongly about, or care about can be difficult; even more so online or in discussion boards where the threat of abuse and harassment can be prominent. Many messaging and discussion-based platforms struggle to filter out toxic and outwardly offensive comments including comments that are rude, disrespectful or otherwise likely to make someone feel vulnerable and leave a discussion. To combat this, ultimately, most platforms limit or completely remove offensive comment sections as a result. With the advancement of computing technology, including AI, assessment and learning, can a code be developed to recognize toxic comments in online conversations with respect to mentions of identities?

Client:
Conversation AI, a joint venture between *Jigsaw* and *Google* (both subsidiaries of *Alphabet*), have been the primary business prompting this problem. Even though this is one of the focuses of the *Conversation AI* team, the question of recognizing these potentially hazardous comments has a far-reaching impact on more online platforms and businesses than just this joint-venture.

Data:
The data used for this project is publicly available at the following address: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data. These data are originally from *Civil Comments*, a collection of about 2 million public comments. The dataset used is 'train.csv' which is comprised of the following features:

- id - unique comment id
- target - toxicity label between 0 and 1; target >= 0.5 is considered positive (toxic)
- comment_text: text of the individual comments
- severe_toxicity - toxicity subtype attribute
- obscene - toxicity subtype attribute
- threat - toxicity subtype attribute

- insult - toxicity subtype attribute
- identity_attack - toxicity subtype attribute
- sexual_explicit - toxicity subtype attribute
- 24 identity features - white, hispanic, male, female, other_disability, etc.

Problem Approach:
The raw comments together create my raw corpus (collection of documents). A document, in this instance, can also be called a comment. Once my raw corpus has been compiled, cleaning the data can commence. There are many different methods that can be applied: removing conjugations, stop words, and superfluous tokens to name a few. The clean corpus is then vectorized. The vector will have a value to indicate if a word is in the document as either a flag (0 or 1) or frequency count (a value greater or equal to 0), respectively, and be the same size as the corpus. In the end, a design matrix representing a stack of vectors representing documents is created. Given that target is known, the design matrix will be used in classification algorithms to build models, analyze their performance, and suggest how to use these models in the context of the business problem.

Deliverables:
The deliverables will include all code developed with each step contained in it's own Jupyter Notebook, a written final report, and a written presentation slide deck.