Machine Learning Engineer Nanodegree

Capstone Proposal

André Azevedo Muta
October 18st, 2018

Proposal

Domain Background

This project is based on the Kaggle's **Toxic Comment Classification Challenge** (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge) competition.

The internet is usually not the best place where healthy discussions happen. The constant clash of divergent views and the perceived distance of an physical confrontation by acting behind a screen make way for the prevalence of abuse and harassment and platforms that struggle to effectively facilitate conversations end up limiting or completely shutting down user comments.

The **Toxic Comment Classification Challenge** focused on tackling this problem by detecting different types of toxic comments using as data human-labeled comments from Wikipedia's talk page edits. The competition's goal was to improve current models used by the Conversation AI team (a research initiative founded by Jigsaw and Google) by accurrately predicting toxicity and its types.

Problem Statement

The goal is to build a **multilabel classification** algorithm that is capable of distinguish different types of comment's toxicity, like threats, obscenity, insults, and identity-based hate, that have a better performance than Perspective's current models (https://github.com/conversationai/unintended-ml-bias-analysis), the only data provided are the comment's text, making it an overall **supervised learning problem** that will mainly use **Natural Language Processing** methods for extracting information from the comments.

Datasets and Inputs

The competition's datasets are compromised of a larged number of comments from Wikipedia's talk page edits which have been labeled by human raters for toxic behavior.

The labels are:

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

The datasets are:

- **train.csv** the training set with 160 thounsand rows and 7 columns - contains comments with their binary labels

- **test.csv** the test set with 153 thousand rows and 2 columns - the goal of the competition was to predict the toxicity probabilities for these comments. To deter hand labeling, the test set contains some comments which are not included in scoring.

Solution Statement

The dataset was labeled in 7 different types of toxicity and they are not mutually-exclusive, but some are more problematic than others so it is important to be able to tell them apart.

To create a model that is capable of distinguish them, in this **multi classification problem**, using only the text from the comments, the solution will be compromised of mainly NLP transformations on the comment text that will be used to train a machine learning classification model, which will predict the probability of true for each one of the seven columns.

Benchmark Model

The benchmark model will composed of a TF-IDF (Term Frequency - Inverse Document Frequency) followed by a simple Logistic Regression.

TF-IDF is a simple numerical statistic that calculates the frequency of each term t in a document d and weight it by the inverse document frequency, i.e., the rarity of such term across all documents [1]. It can be expressed as:

> Let $C(t, d)$ be the count of ocurrences of t in d, the TD-IDF statistic can be expressed as:
> $tdidf(d, t) = tf(d, t) * idf(d, t)$
> $tf(t, d) = log(1 + C(t, d))$
> $idf(d, t) = log\frac{n}{df(d,t)}$

Altough the competition is already finished it is possible to compare the score against the leadeboard, so it will also be used as a form of final benchmark.

Evaluation Metrics

The evaluation metric used by the competition was the mean column-wise AUC-ROC. In other words, the score is the average of the individual AUCs of each predicted column.

The area under the curve (AUC) of a receiver operating characteristic (ROC) of a classifier can be interpreted as the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example [2], i.e.

$$P(score(x^+) > score(x^-))$$

Project Design

1. Template: build the template of the project focusing on reproducibility.
2. Exploratory Data Analysis (EDA): raw data.
3. NLP: extracting the data.
    - Tokenization - TD-IDF
    - Try Stemming/Lemmatization
4. EDA of features created by the TD-IDF.
    - Distributions by type of toxicity
    - Train and Test dataset

5. Benchmark model: TD-IDF Logistic Regression Classifier
6. Improvements: Try machine learning algorithms and NLP techniques combinations and save the scores.
    - NLP: TF-IDF, GloVe, FastText
    - ML: Logistic Regression, LightGBM, Naive Bayes
7. Combining: blend the predictions of the best scoring models

---

[1] https://en.wikipedia.org/wiki/Tf%E2%80%93idf (https://en.wikipedia.org/wiki/Tf%E2%80%93idf)

[2] https://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it (https://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it)