# Machine Learning Engineer Nanodegree

# Jigsaw恶毒评论分类

王金栋 优达学城

2019.3.12

## 问题的定义

### 项目背景

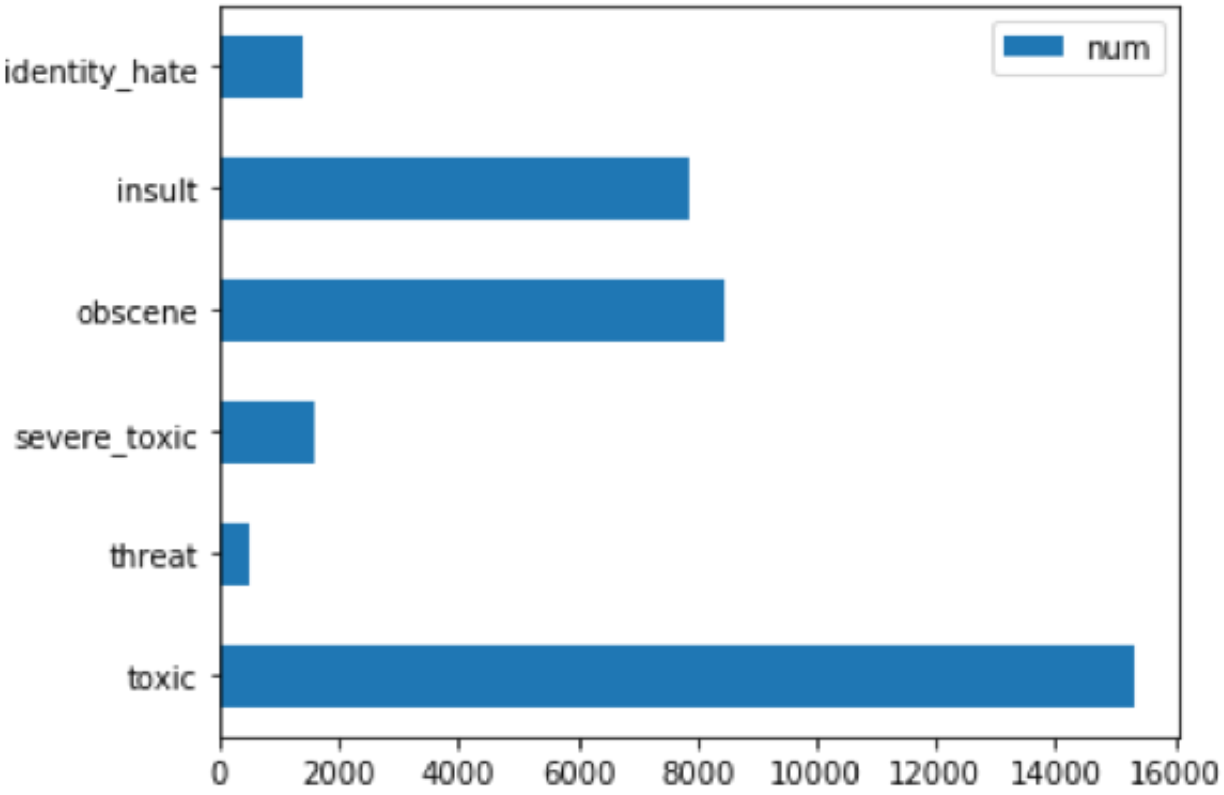
- *通过项目介绍中的信息大概判断这是一个文本分类的项目，目前主流的文本分类模型有LSTM,GRU,LSTM-CNN,GRU-CNN 等等*
- Jigsaw(前身为Google ideas）在kaggle平台上举办了一场文本分类比赛，旨在对于网络社区部分恶毒评论进行区分鉴别。在该赛题中，你需要建立一个可以区分不同类型的言语攻击行为的模型，该赛题一共提供了toxic,severe_toxic,obscene,threat,insult,identity_hate这六种分类标签，你需要根据提供的训练数据进行模型训练学习。

### 问题陈述

- *该项目简而言之就是文本多分类问题，找出涉及言语辱骂，威胁性质等评论*
- *根据以往**NPL**的处理经验优先选择**LSTM**及其升级版**GRU**模型还有其他融合该两种模型的变种模型*
- *分别列出分属于不同种类的概率，这样即可清晰的表示出分类的结果*

### 数据或输入

通过train.csv中可以看到数据集共有8个纬度，其中第一个ID 第二个comment_text也就是评论内容，剩下的8个纬度为分类的种类，在训练过程中X为经过分词量化处理后的comment_text，而Y就是6个种类，在该集合中，样本分布不均且同一个样本可能有多种类别，分布结果如下图:




该数据集大概有160K行样本，因为我们只能获取训练集的答案所以我们讲训练集分成训练、验证两个部分，随机划分10%的比例作为验证集，剩余的作为训练集

## 评估标准

- *ROC、AUC、K-S 曲线 而**K-S**是通过计算**max(TPR-FPR)**求得**KS**值 KS值越大表明模型的区分能力越强*
- AUC（area under curve）即ROC曲线下的面积。（随机给定一个正样本和一个负样本，分类器输出该正样本为正的那个概率值 比 分类器输出该负样本为正的那个概率值 要大的可能性）
- $$AUC = \frac{\sum_{ie\ positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N}$$
- *一般评判模型我们都是使用**TP**、**FP**、**TN**、**FN**来求得**ROC**、**AUC**曲线，这两个评判指标可以很清晰的描述出模型当前的表现能力，若加上**K-S**曲线可以使结果的表达能力升华*

## 基准模型

- *基准模型我首先选择**LSTM+Attention,**从理论层面讲**LSTM**当序列很长时难以学到合理的向量表示，且输入序列不论长短都会编码成一个固定长度的向量表示，打破了传统编码器-解码器结构在编解码时都依赖于内部一个固定长度向量的限制。*
- 且开篇中提到该项目是一个文本多分类项目，首先参考的就是LSTM模型进行文本分类，而独立的LSTM不足以完成我们的目标所以考虑融合Attention机制
- 其次考虑使用textCNN模型，对于文本分类问题，常规方法就是抽取文本的特征，使用doc2evc或者LDA模型将文本转换成一个固定维度的特征向量，然后基于抽取的特征训练一个分类器。而TextCNN 是利用卷积神经网络对文本进行分类的算法，并且有着卓越的表现

- 使用来自外部embedding的知识可以提高RNN的精度，因为它集成了有关单词的新信息（词汇和语义），这些信息是在大量数据集上训练和提炼出来的。

  预先训练的embedding我将使用GloVe

- GloVe是一种无监督学习算法，用于获取单词的矢量表示。 对来自语料库的汇总的全球单词共现统计进行训练，结果展示了词向量空间的有趣线性子结构。

  我将使用的Glove embedding是在一个非常大的公共互联网爬行训练得到的，其中包括：

  - **840亿标记**
  - **220万词汇**

- 首先从数据方面进行优化，对数据进行合理分析选择最优的优化方式，其次对模型进行参数的调整，不断实验找出最合适的参数
- 通过观察ROC和AUC曲线，以及模型收敛的速度最后的准确率

## 项目设计

- 首先我们对数据进行可视化分析，分析样本分布以及有无异常、空值，根据特征类型判断如何对异常进行处理以及对空值的补齐操作
- *这是一个典型的文本分类的项目，所以第一考虑使用**LSTM**进行分类操作*
- *因为**LSTM**为长短时记忆的机器学习模型，可以通过对每一个单词的分析来判断该句子是否属于哪一类。*
- 但是LSTM在过长序列中无法进行有效的语意分析，且所有的向量都被固定了长度，所以考虑加入Attention机制来解决上述问题
- 最后输出ROC及AUC曲线，观测模型性能

## Project Design

http://www.jeyzhang.com/understand-attention-in-rnn.html

https://www.cnblogs.com/wkslearner/p/8748141.html

https://www.jianshu.com/p/9570dd34ffb2

https://zhuanlan.zhihu.com/p/31547842

https://zhuanlan.zhihu.com/p/47976826

http://www.gzhshoulu.wang/article/2290189

---

**Before submitting your proposal, ask yourself. . .**

- Does the proposal you have written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your proposal?

- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?