Machine Learning Engineer Nanodegree

Capstone Proposal

Yuanchao Ma

March 20, 2019

Proposal

Domain Background

互联网的本质是为了降低人们获取信息的成本，更便捷的进行沟通和分享。因此，从互联网创建时，就允许世界各地的人们通过互联网进行自由的交流、讨论以及合作。而像国内的贴吧，微博，微信，国外的Twitter，Facebook，Wikipedia等社区平台的建立，形成了这些互动可以发生的基础。为了人们在社区中可以更有序的交流以及促进对话，许多的社区都制定了自己的标准和规则，并防止这些社区被有毒行为劫持或摧毁。然而，随着有毒评论的黑色产业化，利益驱使人们通过各种手段来规避规范和标准，使得通过人为来执行这些规范和标准变得越来越困难。事实上Facebook正在招聘越来越多的版主来筛选可疑的内容[1]。同时，许多新闻网站现在也已经开始禁用评论功能[2]。而这些人工的审核监控机制，是非常低效的做法。

综上，我们需要一种工具来自动化地对用户评论进行监视，分类和标记。此外，不同的网站可能需要监控不同类型的内容。因此需要建立一个能够区分不同类型的言语攻击行为的模型。

我们可以看到在论文[3]中，研究人员对情感分析进行了大量研究。他们的工作重点是情绪分析，这与我们正在研究的领域非常相似。论文中定义了一种使用词袋技术预处理文本的合理方法。他们接着使用SVM和朴素贝叶斯分类器来确定推文的情绪是积极的，中性的还是负面的，并且发现朴素贝叶斯分类器更准确。此外，当他们对推文进行矢量化时，他们通过使用bigrams来提高分类器的准确性。他们的工作可以为我的benchmark model参考。

Problem Statement

Toxic Comment Classification Challenge是kaggle上由Jigsaw提出的一个比赛，旨在找到更好的对恶毒评论的多分类模型。我通过参加这个比赛，使用比赛提供的人工标注的Wikipedia评论数据，训练一个能够在任意文本数据上判断多种恶意（威胁，色情，侮辱和种族歧视言论等）分类概率的多分类模型。

Datasets and Inputs

训练数据由Toxic Comment Classification Challenge比赛提供。数据为对恶性行为人工标注的Wikipedia评论数据。标注的类型为：

- toxic
- severe_toxic
- obscene
- threat
- insult
- indentity_hate

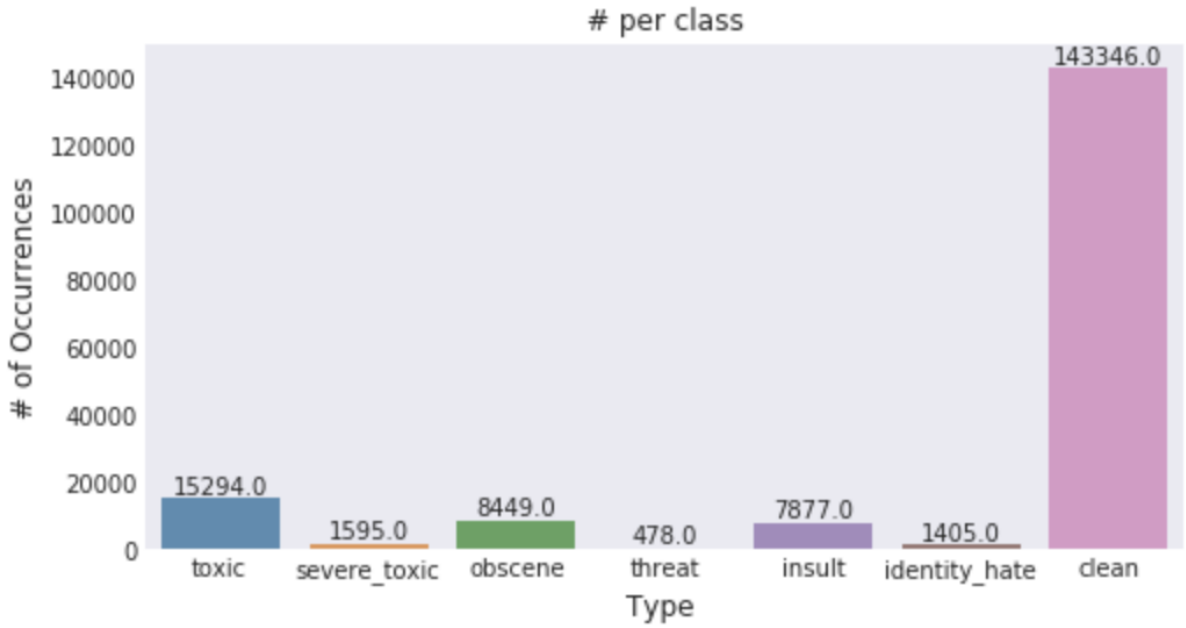
比赛提供的数据由如下四个文件构成：

- train.csv - 训练集，包括159571条已进行标注的评论数据
- test.csv - 测试集，包括153164条待检测数据

csv文件的数据格式为：

- id
- comment_text
- toxic
- severe_toxic
- obscene
- threat
- insult
- indentity_hate

其中，comment_text是模型的输入，模型的输出是输入被判断为每个分类（toxic，insult等）的概率。在训练集中，评论分类的个数分布如下图[4]，可见该数据集是一个非平衡的数据集。



Solution Statement

我的解决方案为：训练一个多分类的分类器，分类器的输入为评论数据c，输出为这条评论在每个分类上的概率r。c为一个文本字符串，r为取值范围(0,1)的数值。

一个在文本分类里效果不错的SVMNB算法作为我的Benchmark model，长期短期记忆网络（LSTM）[5]是一种回归神经网络（RNN）算法，也是一种专为自然语言处理而设计的算法。经过实践验证可以很好地运行，并且可以作为解决方案的基础。我将使用word embedding对数据进行预处理，将文本转换为可以馈送到神经网络的数字向量表示。作为解决方案的一部分，我将评估几种单词嵌入方法，如Word2Vec，Glove，FastText。

Benchmark Model

SVM是最常用的文本分类算法之一，可用作基准模型。基于SVM和朴素贝叶斯算法的SVMNB[6]，它提供了比传统SVM更好的性能，是在kaggle比赛中的推荐的benchmark，我将使用SVMNB作为我的benchmark

model。

Evaluation Metircs

我使用列平均的ROC AUC作为我的评估指标，它是单个类别预测结果ROC AUC的平均值。ROC曲线是在不同分类阈值下使用TPR和FRP绘制的图，而AUC则是ROC曲线下面积，当AUC值越大，当前的分类算法越有可能将正样本排在负样本前面，即能够更好的分类[7]。同时，偏差，方差，精度，召回和F1分数也将用作评估指标，以检查过度拟合和欠拟合。

Project Design

解决这个问题可拆分为如下的步骤[8]：

1. 数据探索
2. 数据预处理
3. 模型设计
4. 模型评估

第一步是数据探索。我会对训练数据集中的平均分类进行分布统计，并创建可视化图形。此外，还可以创建词云图，以了解每个类别中的常用词。同时，了解数据集中的独特单词，常出现单词，填充单词等，对于数据集的理解也很重要。

第二步是预处理数据，例如空值处理，异常处理处理等。在预处理期间，需要删除所有不需要的数据。这包括可能包含随机字母或单词的垃圾数据，非文本数据，用户名等。预处理措施一般包括：

- 大写变小写
- 去掉停顿词，标点，空白文本，英文之外的其他文本
- 分词
- 词性标注 - 帮助我们更好的理解单词/句子的含义
- 词干提取 - 减少输入的语料库
- 生成文档矩阵后计算tf-idf，去除频率较低的单词（例如去掉频率小于5的，或去掉在60%文档中出现的单词）

在将数据输入神经网络模型之前，需要将整个数据解析并标记为单独的单词，并且每个单词将使用其索引进行编码。然后，每个评论文本数据将使用索引值表示，其中每个单词都用其索引替换。每个索引值都是网络的特征，这被称为词嵌入。目前有几种效果不错的词嵌入方法，如word2vec，GloVe，FastText等。

我将使用循环神经网络（RNN）作为我的解决方案，它是神经网络的一种，网络会对前面的信息进行记忆并应用于当前输出的计算中，因此它可以处理顺序数据。RNN在NLP中取得了巨大的成功和广泛的应用。但是，传统的RNN使用BPTT，存在梯度消失的问题，它无法记住长期的信息。为了解决这个问题，创建了长短期记忆网络，它是一种特殊形式的RNN，具有4个神经网络层和3个门（输入，输出和遗忘）的LSTM单元。这些层和门有助于网络记住相关信息并忘记无关信息。GRU（Gated Recurrent Unit）是2014年提出来的新的RNN架构，它是简化版的LSTM，在超参数（hyper-parameters）均调优的前提下，这两种RNN架构的性能相当，但是GRU架构的参数少，所以需要的训练样本更少，易于训练。

当模型训练完成后，我将使用交叉验证集对数据进行交叉验证，以调整各种参数，然后使用测试数据集进行测试，并使用所描述的评估指标（AUC）进行评估。

Reference

1. http://fortune.com/2018/03/22/human-moderators-facebook-youtube-twitter/
2. https://www.theguardian.com/science/brain-flapping/2014/sep/12/comment-sections-toxic-moderation
3. http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf
4. https://github.com/udacity/cn-machine-learning/blob/master/toxic-comment-classification/pics/hist.png
5. https://www.researchgate.net/profile/Sepp_Hochreiter/publication/13853244_Long_Short-term_Memory/links/5700e75608aea6b7746a0624/Long-Short-term-Memory.pdf
6. https://nlp.stanford.edu/pubs/sidaw12_simple_sentiment.pdf
7. http://alexkong.net/2013/06/introduction-to-auc-and-roc/
8. https://github.com/Kirupakaran/Toxic-comments-classification/blob/master/proposal.pdf