Web 信息处理与应用项目报告

——百度贴吧帖子情感分析

刘同禹　　2019000171

1. 项目背景

百度贴吧是现在最大的中文交流论坛之一，用户可以在讨论某个相关主题的贴吧中发表帖子，表达自己对某人某事的看法。揣摩发帖用户的对帖子中讨论的事物的情感是帖子浏览者所必须要做的事情，因此希望能够借助神经网络来对帖子进行情感分类，对于发布的帖子，我们希望能够通过帖子的标题和内容摘要来判断该用户持积极情感还是消极情感。

为了缩小问题的讨论范围，我们选取了与“知名电竞选手 UZI 退役”相关的帖子进行情感分析，一方面是因为该事件讨论度高，相关帖子的数量较多，二是对该选手本身的讨论即存在较大争议，可以避免数据集的标签过于倾斜。

2. 数据收集

2.1 帖子爬取

我们选择在 LPL 吧、UZI 吧和抗压背锅吧这讨论主题与电竞高度相关的贴吧进行数据采集，并且只选取 UZI 退役三天内的帖子，以保证帖子的相关事件与我们所关心的事件有关。初步的爬取后，共收集得到 45023 篇帖子。

为了进一步筛选得到与“UZI 退役”这一事件或与 UZI 本人直接

相关的帖子，我们采取关键词过滤的方式进行筛选，通过筛查帖子的标题和内容摘要中是否出现了某些关键词来判断该帖子是否是目标帖子，最终我们筛选得到 6142 篇帖子。

2.2 数据标注

最初我们想使用基于规则的标注工具 Snorkel 来进行数据标注，但考虑到数据本身的复杂性，难以单纯地仅通过规则来判断发帖人所蕴含的情感，具体来说，有一些发帖人喜欢“阴阳怪气”，即表面上用一些积极情感的词语和描述，但句子的总体含义是对描述人物的贬低，即消极情感，因此，单纯地利用规则来进行标注会使得大量的负例无法被发现。

因此使用 CrowdGame[1]系统进行标注，该系统采用众包+规则的方式来进行数据的标注，一方面使用规则降低一部分人力成本，一方面采用 human-in-loop 的模式对规则进行监督，保证数据的质量。

最终得到 4236 条正例样本，1817 条负例样本。

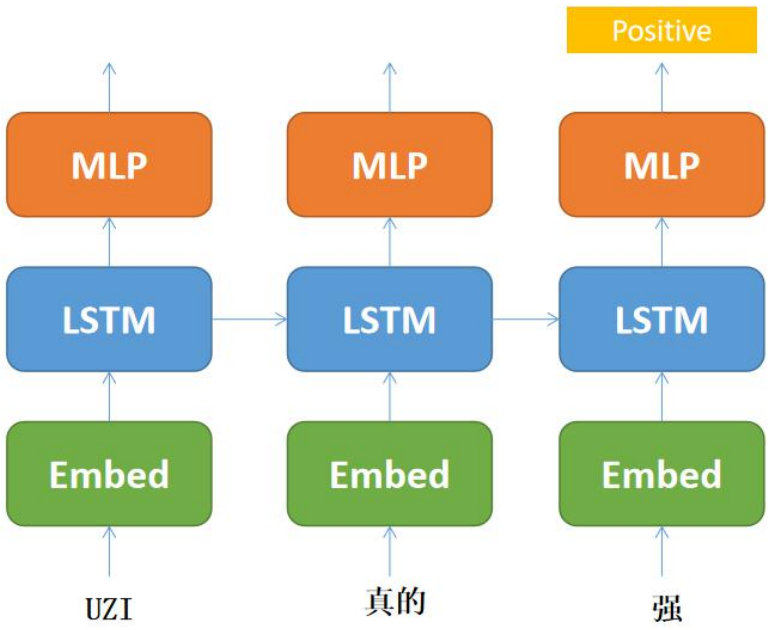
2.3 数据预处理

对于数据中的中文句子，我们采用 jieba 分词工具进行分词，并建立字典，将每一个词语映射为一个数字，以此将一个中文句子映射称一个向量，来作为模型的输入，考虑到句子含有的平均单词数量为 20，我们将长度不足 20 的向量以 0 进行填充，长度大于 20 的向量进行截断，以保证输入的序列长度都为 20。

3. 模型选择

3.1 基础结构

考虑到 RNN 模型在序列处理上的优异表现，我们使用 LSTM 模型作为我们的情感分析模型，基本模型设计如下图所示：



在先经过 Embedding 层进行语义的提取，再通过 LSTM cell 抽取序列特征，最终通过一个 MLP 层得到最终分类结果。

3.2 模型设置

LSTM 结构在单向输入序列的同时，还可以逆向进行序列输入，来同时得到正向序列和逆向序列的两个特征，在某些情况下，这种双向 LSTM 结构能够更准确地得到序列的特征，以提升模型表现，因此，我们尝试单向和双向两种 LSTM 结构，来观察在贴吧帖子的数据集上，双向 LSTM 的表现。

除此之外，LSTM 结构还可以引入 Attention 机制，对 hidden 变量赋予权重来使得模型能够集中关注在序列的关键部分，我们同样考

虑引入 Attention 和不引入 Attention 这两种情况，来观察模型的表现。

4. 实验结果

4.1 模型表现

每种设置的 model 在测试集上的 accuracy 表现如下：

	Without Attention	With Attention
Unidirectional	0.734(0.012)	0.755(0.015)
Bidirectional	**0.771(0.021)**	0.764(0.012)

4.2 训练时间

每种设置的 model 平均训练时间如下：

	Without Attention	With Attention
Unidirectional	22	24
Bidirectional	33	34

5. 结论

首先，未引入 Attention 机制的单向 LSTM 结构的表现最差，这一点是符合预期的，毕竟模型最为简单，考虑的因素更少。而横向比较来看，改变 LSTM 结构为双向结构所带来的提升效果会高于引入 Attention 机制带来的提升效果，这可能是由于该数据集句子长度普

遍较短，很少出现因为句子过长而带来的序列重要信息的遗忘，因此Attention机制的收益相对较小。

而总体来看，模型的分类效果并不算太好，一点原因可能是因为数据的标注还是存在一定的问题，毕竟我们采用弱监督的标注方式，不可避免地引入了噪音。

[1]T.Liu, J.Yang, J.Fan. CrowdGame: A Game-Based Crowdsourcing System for Cost-Effective Data Labeling. SIGMOD Conference 2019: 1957-1960