实验报告

题目：微博情感分析

姓名：周海、高涵

系别：工程系

指导老师：陈湘涛

完成时间：2020.12.15

一、引言

文本情感分析，又称意见挖掘（Opinion Minning），是指对带有情感色彩的主观性文本进行采集、收集、分析、归纳和推理的过程，涉及人工智能、机器学习、数据挖掘、自然语言处理等多个研究领域。

二、问题定义

本实验问题定义为：根据给定的10万左右数据集，运用监督/非监督机器学习算法，判断新来的测试数据集每条review的情感，是消极还是积极，以此实现分类功能。

三、研究思路及算法过程

整体框架流程如图所示：



图1 整体流框架

1.观察数据集，了解数据集的构成和基本特点。

微博文本一般具有短小随意的特点，一般均带有较强烈的情感色彩。利用给定的数据集训练实现分类。其中数据集包括两列，一列微博评价（review），另一列为消极/积极评价（label），label分为**两类**，分别为0（positive）和1（negative）。

若是将每条review进行jieba分词进行切分，画图（频数分布图）观察，近似满足正态分布，绝大多数的review长度集中在20左右，分布图如图X所示。必要时可以考虑对长文本的截断。

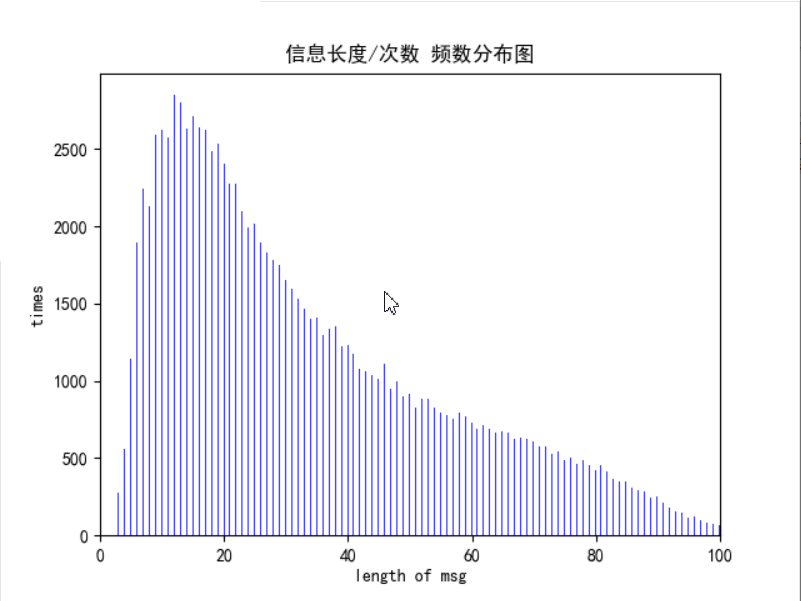


图2 review长度频数分布图

2.数据预处理

仔细观察数据集，我们要对review中对于判别情感没有用的文字信息进行剔除。

(1)删除类似于@某人的文字，如“//@曾在泸沽湖盛开的啥啥啊:”；

(2)删除类似于回复某人的文字，如“回复@中紫灰:”；

(3)删除类似于hashtag的文字，如“#开心早安#”；

(4)删除【】内（包含【】的）文字，如“【上学时最令你吐血那句话】”；

(5)删除无用的链接，如“<http://t.cn/zjt9bgC>”；

(6)删除类似于hashtag加空格的文字，如“@色色168”。

3.jiaba分词

4.加载停用词表cn\_stop\_list（中文停用词表），删除停用词

5.word2vec训练得到自己的词向量

6.将每条词向量用最大长度（填充0）的方法补充得到相同长度，设计、训练神经网络，得出神经网络模型model。

神经网络在情感分析中，不仅能够提取文本特征，还能自动学习并自动修正模型，与建立情感词典的方法相比，其优越性显而易见。设计的神经网络如下图3。

7.借助模型预测情感。

评价指标可以为：

整体7步综合起来步骤如下图：



图3 神经网络设计 图4 研究步骤

四、实验结果与讨论分析

根据二八原则，我们将测试集和训练集2-8分，并且是平衡数据集，positive和negative非常近似满足于1:1的比例。

以下是神经网络的训练过程截图：显示ETA、loss和acc等特征。

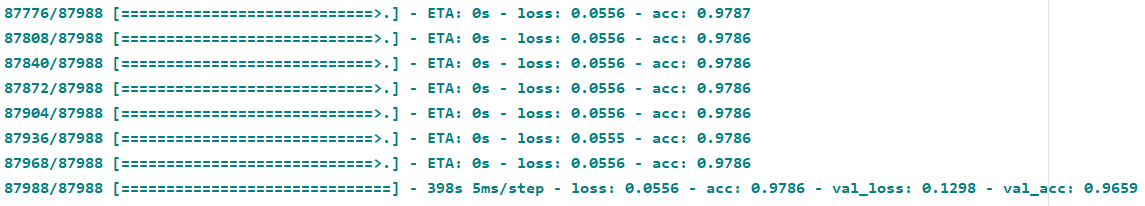


图5

初始时，神经网络迭代次数epoch设置为5，以下是每次迭代时acc和loss的变化图，如下图所示。从图中可以看出，train acc在不断升高，train loss在不断降低，但是当epoch=2之后增长幅度不再很明显。从val acc和val loss来看，好像不存在过拟合现象，原因可能是因为数据集小，然后使用的神经网络过于复杂。

最终，我们可以根据测试集的情况对epoch进行略微调整，已实现最大的预测准确率。

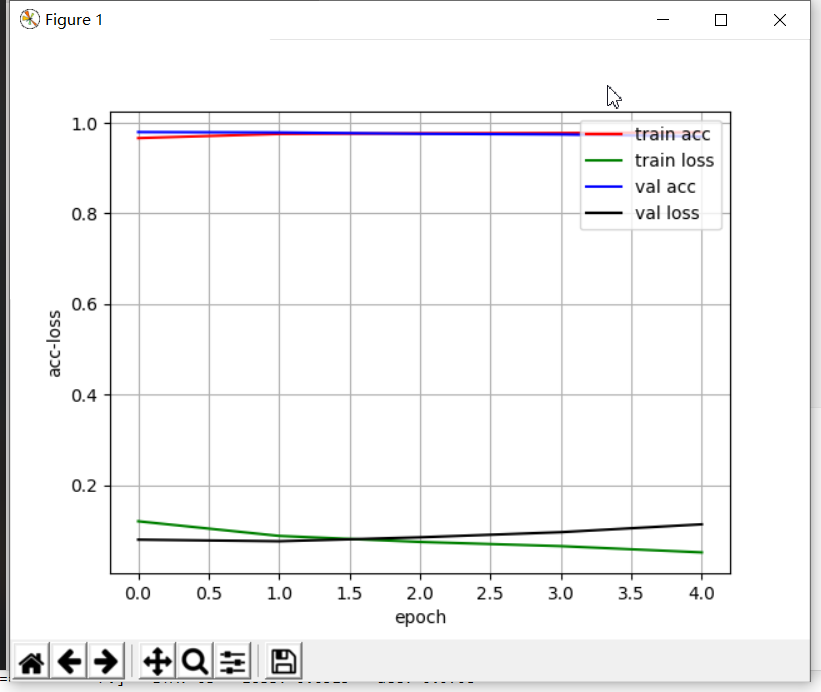


图6 acc-loss图

五、主要结论

因此，神经网络是一种性能高效，分类效率比较好的分类方法，[6]等提到，和SVM、Naïve Bayes相比，神经网络分类效果往往不会令人失望。

接下来的工作：可以通过与SVM和Naïve Bayes进行比较，然后得出神经网络更能够捕捉到文本之间的信息。

六、创新点

1.对数据集进行截断分析。

2.使用Google 很强大的Word2vec工具[7-8]，创建属于自己的词向量。

七、参考文献

[1]余容,李光强,尹健.微博文本挖掘研究综述[J].情报探索,2017(05):97-103.

[2]洪巍,李敏.文本情感分析方法研究综述[J].计算机工程与科学,2019,41(04):750-757.

[3]金志刚,胡博宏,张瑞.融合情感特征的深度学习微博情感分析[J].南开大学学报(自然科学版),2020,53(05):77-81+86.

[4]杨瑞丽. 基于深度特征和加权word2vec融合模型的情感分析算法研究[D].西安科技大学,2020.

[5]姜丹. 基于卷积神经网络和BiLSTM的网络舆情情感分类研究[D].杭州电子科技大学,2020.

[6]王立志,慕晓冬,刘宏岚.采用改进粒子群优化的SVM方法实现中文文本情感分类[J].计算机科学,2020,47(01):231-236.

[7]M. Al-Amin, M. S. Islam and S. Das Uzzal, "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words," 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, 2017, pp. 186-190, doi: 10.1109/ECACE.2017.7912903.

[8] J. Gao, Y. He, X. Zhang and Y. Xia, "Duplicate short text detection based on Word2vec," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, 2017, pp. 33-37, doi: 10.1109/ICSESS.2017.8342858.