上下文情感分析调研

目前对上下文情感分析方法主要分为基于语法规则，基于机器学习算法SVM及变种，深度学习算法LSTM,CNN，混合算法。 大部分论文主要在对特征处理上采用不同方法，再基于得到的特征采用适合的算法处理，目前调研的方法如下

A context-based model for Sentiment Analysis in Twitter

还运用一个马尔科夫链式表示方法。

对文本特征提取采用了不同的核心Bag of Word Kernel（word-level），Lexical Semantic Kernel (LSK)(Sentence Level)，User Sentiment Profile Context (USPK).再将通过这些核心生成的特征线性组合起来，用multi-class SVM进行分类，同时将上下文转换成一个序列标记问题再用基于马尔科夫链方法进行预测。

Modeling Rich Contexts for Sentiment Classification with LSTM

用了一个分层的LSTM模型，主要分为两层，一层为tweet-level基于上下文，并将上下文本同样分为三类，转换为二元特征在进入LSTM处理。同时在tweet-level上将当前tweet信息构建一个基于word-level的LSTM生成一种表示在进入tweet-level LSTM单元处理。

测试集为跨语种包含中文和英文。

Opinion Target Extraction in Chinese News Comments

采用语义规则分类器将目标分为显性和隐性，对评论进行分析的时候，当对隐性目标分析的时候加入了对主体内容分析，同时主要用互信息进行计算排序。在主体内容用类似提取中心概念方法进行处理，用信息熵进行对主题分别计算。

对于显性目标分析则基于sentence-level：Subjective主观》objective客观》others

再用中心定理（Centering theory）提取出最佳目标。

基于字典和监督学习

基于特征情感分类

找最常出现名词

论文测试中的准确率不高

A Context-based Neural Network Model for Twitter Sentiment Classification

该文章基于word-level进行情感分析

将上下文分为三种，

基于回复

基于作者

基于主题hashtag,

算法主要包括两种子神经网络，一种是原文（评论）的卷积网络主要有五层神经网络

输入层 将词用转换为低维实数词向量

卷积层 该层使用N-gram 来对进行卷积，N=3.

池化层 该层用来生成特征，同时减少过拟合，目前流行的要最大池，最小池和平均池，在这里这个算法将三种方法都采用了。

隐藏层 采用双正曲函数tanh对特征值进行非线性化修正（添加激活函数）。

输出层 将特征值用线性函数计算在各个标签下的可能性。

第二个是上下文的神经网络

将上下文的词转换为tf-idf值，取排序前100的词转换为相关关键词向量后不进行卷积直接池化层再与原文的词向量合并进入隐藏层处理。

除此之外，它还将characters(表情符号或者词根词缀)也考虑进去，采用与原文卷积神经网络的神经网络，也池化层处理后与其他特征合并。

测试集分为平衡集和非平衡集，主要为纯英文，采用基于上下文的神经网络准确率可以达到百分之90以上，相比于基线方法（SVM,NEC,SSWE）高百分之10左右的准确率，不采用上下文相比则略优。

Weakly-Supervised Deep Learning for Customer Review Sentiment Classification

[A Context-based Neural Network Model for TwitterSentiment Classification](http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/11922/11586)

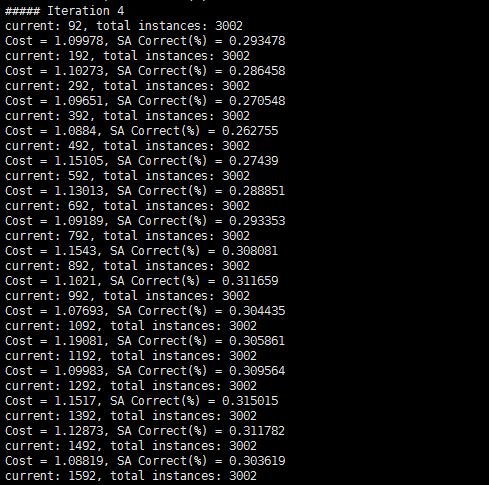
初步测试结果如下：

目前该算法程序仅能用于模型训练

数据量：约3000条评论（Positive 1000,negative 1000,neutural 1000 ）

模型训练耗时：超过4个小时

迭代次数：5次

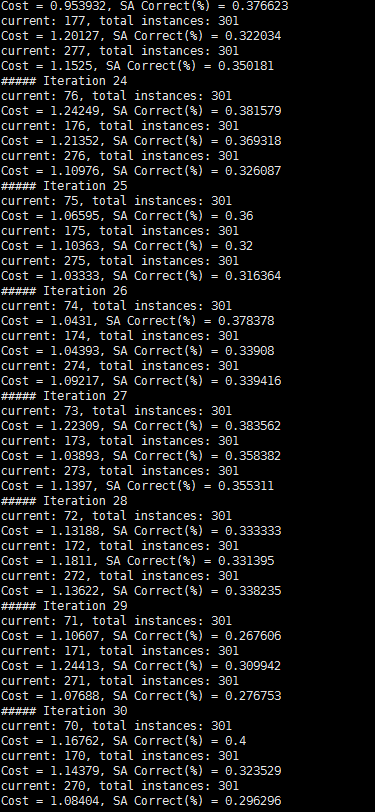
模型预测情况：准确率30%（只对训练集验证），结果如下

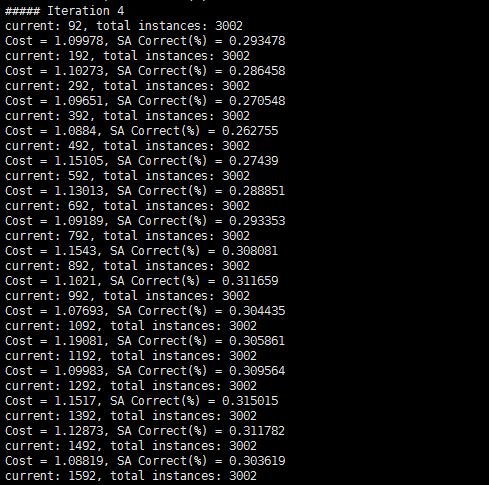
数据量：约300条评论（Positive 100,negative 100,neutural 100 ）

模型训练耗时：超过2个小时

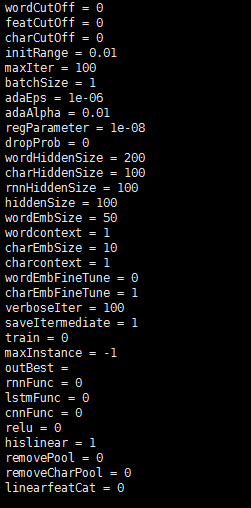
迭代次数：30次

模型预测情况：准确率30%-35%（只对训练集验证），结果如下



当训练数据稍微变大的时候，例如运行训练10000条数据时候会出现运行错误，结果如下

结果差目前总结有几点，一是可能仍不了解具体数据格式，论文中对上下文是有作分类。二是对参数没有详细说明，比较难做适当的调整。具体参数如下



还有就是在没有并行处理下运行时间过长，数千条数据迭代几次就要跑几个小时。

然后目前在github找到另一个基于上下文做情感分析的项目如下

[peace195/context-based-sentiment-analysis](https://tower.im/projects/c41e44862efa4ba488bc6b6bd62cfe72/todos/573b9cf4df844affb6fff7f10f05ba97/)

作者表示准确率为67%，数据格式是基于Sem-Eval2013的twitter数据集，约为10000条twitter 。代码是基于python的，思路是用LSTM来做基于sentence level的上下文情感分析。