Computer Science and Application 计算机科学与应用, 2017, 7(10), 944-950
Published Online October 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.710107

Hans汉斯

Research and Analysis of Textual Emotion Based on Word2Vec

Guanghua Yin

School of Computer Science, Zhongyuan University of Technology, Zhengzhou Henan
Email: 1514599792@qq.com

Received: Sep. 30th, 2017; accepted: Oct. 13th, 2017; published: Oct. 19th, 2017

Abstract

With the development of artificial intelligence, machine learning development, compared with traditional neural network and other non-linear decision and modeling theory, support vector is used in the field of text information processing to solve the classification problem, because of the simple structure and complete theories. This paper puts forward the research on the emotion classification of the depth learning text in view of the intricacies of the massive text emotion data, the inability to grasp the positive and negative information of the text accurately and accurately. Firstly, we introduce the idea of text emotion classification. Then we introduce the text emotion and TF-IDF weight calculation, and improve the TF-IDF weight and mediate the depth of learning Word2Ve word vector + LIBSVM model to train the Internet text data. Finally, the accuracy of the classification reached 92.28%.

Keywords

Machine Learning, LIBSVM, Emotional Characteristics, Word2Vec, Emotional Classification

基于Word2Vec文本情感研究与分析

尹光花

中原工学院，计算机学院，河南 郑州
Email: 1514599792@qq.com

收稿日期：2017年9月30日；录用日期：2017年10月13日；发布日期：2017年10月19日

摘　要

随着人工智能、机器学习发展，支持向量与传统的神经网络等非线性判定和建模理论相比，因结构简单，

文章引用: 尹光花. 基于 Word2Vec 文本情感研究与分析[J]. 计算机科学与应用, 2017, 7(10): 944-950.
DOI: 10.12677/csa.2017.710107

理论完备的优点被人们用到文本信息处理领域，解决分类问题。本文针对海量的文本情感数据错综复杂，不能及时准确的掌握文本正负(褒贬极性)信息，提出了深度学习文本情感分类的研究。首先，叙述了文本情感分类的算法思想；然后，引入了文本情感特征和TF-IDF权重计算，通过改进TF-IDF权重，调解优化深度学习Word2Ve词向量 + LIBSVM模型训练互联网文本数据。最后，分类的准确精度达到92.28%。

关键词

机器学习，LIBSVM，情感特征，Word2Vec，情感分类

1. 引言

随着电子商务、新闻媒体、社交网络的发展，网络社交平台也逐渐取代了以往的电话、信息，成为了朋友之间交流的主要途径，其中海量的信息都以电子文档的形式呈现在人们面前。越来越多的用户想通过网络来掌握某一事件的情感影响主流，以及某些产品的评价观点的好坏倾向。这些文本信息，基本上都含有人们对时间、地点、人物、事件的个人情感看法和态度，这些信息评价对于把握社会情感动态、舆情监控、网络安全、消费需求等多方面都具有客观意义和实用价值[1]。

虽然互联网给我们带来了很多好处，但是增加了我们对繁杂的网络信息的准确选取，比如当我们使用百度时，如果我们输入的情感关键词不够准确，我们就很难准确地获得我们想要的知识，因此如何快速、高效、准确掌握社会情感信息成为我们研究领域之一。

2. 相关研究

情感分类也可称作情感文本挖掘，帮助人们快速有效获取信息，辨别出文本的是非曲直，便于人们做出正确的价值取向。目前，李寿山[2]等结合文本情绪关联性方法，对隐含情绪进行分类研究，提升情绪分类的性能。何跃[3]等人提出基于文档频率和信息增量来构建特征向量模型，实现较佳的细粒度文本分类效果。Pang 在文档层次上利用三种不同的机器学习模型(NB、MaxEnt 和 SVM)对英文影评进行了情感分类，其中包括 Unigram、Bigram 和词性等测试集，发现支持向量机在 Unigram 这个特征上效果最好[4] [5]，准确率达到了 83%。足够说明合适的机器学习的模型能高效解决文本情感分类，但这些基本上都是浅层的机器学习分类方法。本文在浅层机器学习的基础上，融入了深度学习的 Word2Vec 实现词向量

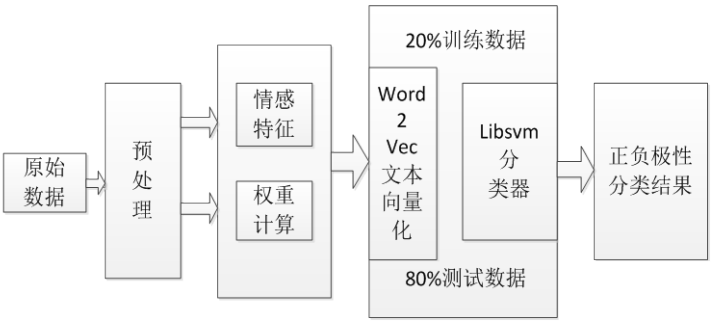

Figure 1. Text emotion analysis framework
图 1. 文本情感分析框架

特征组合，对比发现情感极性判定更明显。

3. 文本情感分析框架

如图 1 所示，首先，对原始数据简单分词、词性标注预处理[6]。然后，在情感特征选取与情感特征加权基础上，结合深度学习 Word2Vec 进行文本的向量化，最后，使用 LIBSVM 分类器分别对 20%、80%数据进行情感极性判定。

4. 情感特征选择

4.1. 特征提取

TFIDF 是一种统计算法，用于表示一个词在一个文档或者语料中的重要程度，在文本分类中经常使用到的一个特征，本文把词的 TF-IDF 特征作为文本情感特征的一个权重。TFIDF 实际上是：TF*IDF，TF 词频，IDF 逆向文件频率。TF 表示词条 t 在文档 d 中出现的频率。

$$IDF\left(t_i\right)=\lg\frac{N}{n} \tag{1}$$

其中，N 为全部的文档数，n 表示包含词条 t 的文档数量。

IDF 的主要思想是：如果包含词条 t 的数量越少，即 n 越小，则 IDF 越大，说明词条 t 具有很好的区别能力[7]。如果文档集合某一类包含词条 t 的文档数为 m，而其它类包含 t 的文档总数为 k，显然所有包含 t 的文档数 $n = m + k$，当 m 大的时候，n 也大，按照 IDF 公式得到的 IDF 的值会小，就说明该词条 t 类别区分能力不强。但是实际上，如果一个词条在一个类的文档中频繁出现，则说明该词条能够很好代表这个类的文本的特征，这样的词条应该给它们赋予较高的权重，作为区别其他类别的一个词条。这就是传统的 IDF 不足之处。

改进后的 IFIDF：针对上面的描述一个词条 t 在文档 d 中出现的频率[8]，对 IDF 值的不足之处做了修改。

$$IDF\left(t_i\right)=\lg\left(\frac{m}{n}*N\right) \tag{2}$$

其中 N 为文档总数量，m 为在文档集合中某一类 c 包含词条 t 的文档数量，n 为在文档中包含词条 t 的文档数量。与传统的 IDF 相比我们把 c 中包含词条 t 的文档数量作为了一个权重，IDF 值随着 m 的增大而增大，也就解决了传统的 IDF 的缺陷。

$$tfidf_{i.j}=tf_{i.j}\times IDF\left(t_i\right) \tag{3}$$

TF-IDF 即是 TF 与 IDF 的乘积，通过 TF-IDF 可以很好的描述文本中关键词的特征。本实验方案的特征选取使用的算法为 TF-IDF，步骤如下：

Step 1：输入数据：词集。

Step 2：输入格式：以词位基本单位存在。

Step 3：处理步骤：

1) 词频统计。

2) 特征词选取。

3) 权重计算。

Step 4：输出数据：文本特征数据。

Step 5：输出格式：索引文本特征词 TF-IDF 值。

4.2. 特征降维

语料经过预处理后，文本以词为基本单位存在，通过特征降维，去除文本中与文本主体无关的词，不仅提高了文本分类的准确率，而且减少了文本分类的时间提高了效率。TF-IDF 即可以选择语料库中出现较多的词，又可以过滤掉一些常用词，所有 TF-IDF 更加能够表示文本的原始特征。经过 TF-IDF 选择后的特征维数相对较低，但这些原始特征之间的各种深度特征还需要通过进一步的提取和降维，然后，获得低维高度可分特征信息利用分类器实现分类。

4.3. Word2Vec 特征向量

Word2vec 就是可以把一个词条用数字特征表示的工具，Word2vec 是一个神经网络，它用来在使用深度学习算法之前预处理文本。它本身并没有实现深度学习，但是 Word2vec 把文本变成深度学习能够理解的向量形式。

文本是一种非结构化的由汉字字符串构成的数据，无法利用计算机技术直接进行训练或分类。因此，进行文本分类的基础是要将文本处理成计算机能识别的形式，训练文本用语料库中的文本，词向量模型训练采用 Word2vector，尤其 word2vec 方法能够在互联网文本中提炼出一个高维度特征向量全面的表达词在互联网文本中的含义。可以修改 Word2vector 中的参数提升词向量模型的效果，训练参数设置如表 1。

表格的左边为 Word2vector 中影响词训练的有关参数，右边简单的对这些参数进行说明。

5. 基于 LIBSVM 文本情感分析

把生成的文本特征向量提交给 LIBSVM 进行构造分类器，其中 LIBSVM 中有两个非常重要的参数 c、g 对分类器预测结果的准确度有很重要的参数，本实验使用 python 脚本来得出 c、g 的最优值。

如图 2 所示：将如我们是情感句识别分类器的结果测试。

A——分类器预测正确的情感句

B——分类器预测错误的情感句

C——是情感句，且分类器没有预测出来

D——不是情感句，且分类器没有预测数来

召回率 R：用分类器预测正确的文本数量作为分子，测试数据中所有的情感句数作为分母，即

$$R(\text{召回率})=\frac{A}{A+C}$$

准确率 P：用分类器预测正确的文本数量作为分子，分类器预测出来的情感句数作为分母，即

$$P(\text{准确率})=\frac{A}{A+B}$$

F 值：综合评估性能方法，$F=\frac{2PR}{P+R}$

分类器性能评估的主要代码：

```
rValue = (double) count / result.size();
pValue = (double) count / predict.size();
doublefValue = (2 * rValue * pValue) / (rValue + pValue);
```

Table 1. Training parameter settings
表 1. 训练参数设置

参数	意义
-cbow 0	采用 Skip-Gram 语言模型
-size200	向量维度
-window5	窗口大小
-negative0,-hs1	不使用 NEG 方法，使用 HS 方法
-threads50	50 个线程并行处理
-binary	模型文件以二进制形式存储

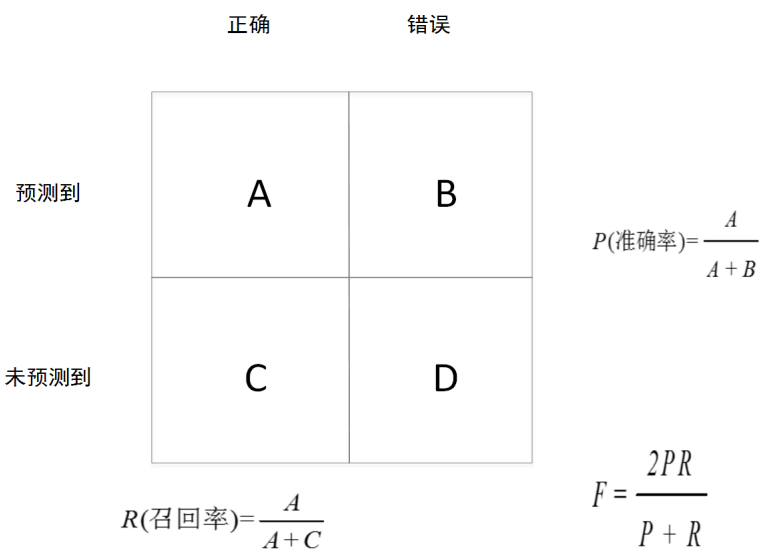

Figure 2. Emotional recognition classifier results test
图 2. 情感识别分类器结果测试

6. 实验

6.1. 数据

本文使用 COAE2014 评测[9]中提供的 9900W 条微博数据作为基础语料，从基础预料中提取 word 在语料的特征表示(TFIDF 特征、word2vec 特征向量)。情感语料库：使用 COAE 评测中提供的情感语料库，其中抽取 80%作为训练语料，其余 20%作为测试数据。

6.2. 实验过程与结果

1) 基本语料库的处理(数据清洗、分词)抽取 TFIDF 特征表、word2vec 特征向量。

2) 抽取情感语料库的 80%的数据作为训练数据，对训练数据分词处理，在 TFIDF 表中检索词的 TFIDF 特征值，在 word2vec 中获得词的 word2vec 特征向量。

3) 使用 TFIDF*word2vec 构造特征向量的方法处理互联网文本。

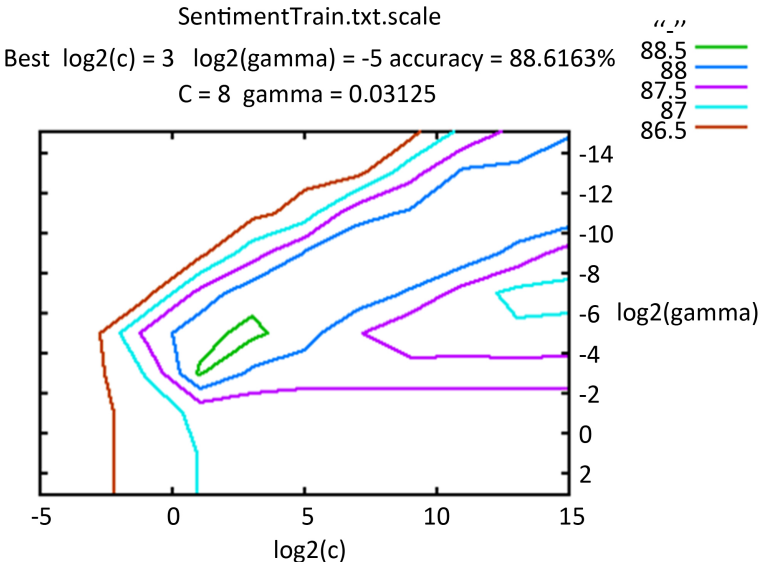

Figure 3. Emotional recognition cross validation
图 3. 情感识别交叉验证

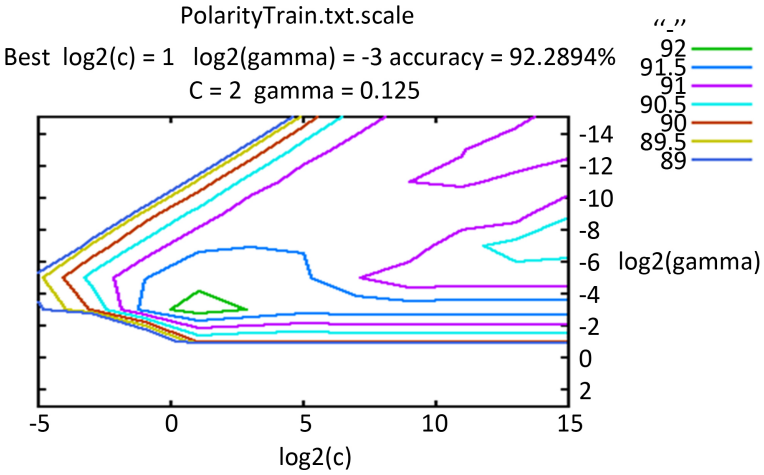

Figure 4. Emotional polarity cross validation
图 4. 情感极性交叉验证

4) 使用 libsvm 处理互联网文本特征向量，并做参数调优处理，生成情感分类器。其中 libsvm 中有两个非常重要的参数 c、g 对分类器预测结果的准确度有很重要的参数，本实验使用 python 脚本来得出 c、g。

如图 3 是使用 TFIDF*word2vec 构造特征向量方法生成的情感句识别分类器，由图可知当 c = 8、gamma = 0.03125 时，结果准确率最高为 88.6163%。因此我们使用情感句识别分类器的时候设置参数 c = 8，g = 0.03125。

如图 4 是使用 TFIDF*word2vec 构造特征向量方法生成的情感极性分类器，由图可知当 c = 2、gamma = 0.125 时，结果准确率最高为 92.2894%。所以在使用 libsvm 训练情感极性分类器的时候添加参数 c = 2、g = 0.125。

7. 总结全文

本文以特征选择和权重计算为特征提取，利用改进 TF-IDF 计算出情感值，融入 Word2Vec 调整参数

优化特征向量，实现了互联网文本情感分类。实验结果证明，基于深度 word2Vec 学习的文本情感极性分类具有可行性。下一步将进行技术优化，技术改进，以提高文本领域分类器的准确率，为互联网文本情感要素抽取作好铺垫工作。

参考文献 (References)

[1] 樊康新. 基于 SVM 的网络文本情感分类系统的研究与设计[J]. 计算机时代, 2015(12): 34-37.

[2] 李寿山, 李逸薇, 刘欢欢, 等. 基于情绪相关事件上下文的隐含情绪分类方法研究[J]. 中文信息学报, 2013, 27(6): 90-95.

[3] 何跃, 邓唯茹, 张丹. 中文微博的情绪识别与分类研究[J]. 情报杂志, 2014(2): 136-139.

[4] Pang, B. and Lee, L. (2006) Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, **1**, 91-231.

[5] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up Sentiment Classification Using Machine Learning Techniquer. *Proceedings of MNLP-02 the Conference on Empirical*, 79-86.

[6] 张勇. 基于词性与 LDA 主题模型的文本分类技术研究[D]: [硕士学位论文]. 合肥: 安徽大学, 2016.

[7] 张玉芳, 彭时名, 吕佳. 基于文本分类 TFIDF 方法的改进与应用[J]. 计算机工程, 2006, 32(19): 76-78.

[8] 宋章浩. 中文文本分类中 TF-IDF 方法的改进与应用[J]. 科技展望, 2014(22).

[9] 杨静, 徐蔚然, 谭松波. COAE2014 情感关键句评测任务和评测数据设计[J]. 山西大学学报(自然科学版), 2015, 38(1).