doi:10.3969/j.issn.1002-0802.2017.03.021

基于加权 word2vec 的微博情感分析 *

李 锐，张 谦，刘嘉勇
（四川大学 电子信息学院，四川 成都 610065）

摘 要：随着社交媒体的普及，微博情感分析受到了广大研究者的关注。为解决情感分析中词间语义关系缺失和词汇重要程度被忽略的问题，提出了一种基于加权词向量和支持向量机的情感分析方法，对微博的情感分析问题进行研究。首先用 word2vec 训练并计算得到文档词向量；然后根据 TFIDF 算法计算文档中词汇的权重，对 word2vec 词向量进行加权；最后，使用 SVM 对情感数据进行训练和分类。在微博实验数据中，与已有方法相比，所提方法分类准确率和召回率都得到了提高。

关键词：情感分析；word2vec；加权词向量；支持向量机
中图分类号：TP391.1 **文献标志码**：A **文章编号**：1002-0802(2017)-03-0502-05

Microblog Sentiment Analysis based on Weighted Word2vec

LI Rui，ZHANG Qian，LIU Jia-yong
(College of Electronics and Info.,Sichuan Univ.,Chengdu Sichuan 610065, China)

Abstract: With the popularity of social media, microblog sentiment analysis attracts more attention from most researchers. In order to solve the problem of lacking lexical semantic relation and neglecting lexical importance in sentiment analysis, a sentiment analysis method based on weighted word vector and support vector machine (SVM) is proposed, thus to analyze the microblog sentiment. Word2vec is firstly used to train and calculate the document word vector, then by using TFIDF algorithm, the weight of document word is calculated, and word2vec weighted. Finally SVM is used to train and classify the sentiment data. Microblog experimental data indicates that compared with the existing methods, the proposed method is greatly improved in classification accuracy and recall rates.

Key words: sentiment analysis; word2vec; weighted word vector; support vector machine

0 引 言

伴随着社交网络的不断发展，更多的人通过微博、博客来表达自己的情感，发表对热点事件的观点。微博平台以其灵活性、及时性，毫无疑问地成为新事件和热门话题的前沿阵地。通过分析微博内容来了解事态的变化及人们的情感倾向，成为许多学者的研究方向。

文本情感倾向性分析，是指对说话人的态度（或称观点、情感）进行分析，也就是对文本中主观信息进行分析。与传统的基于主题文本分类不一样，这种分类对象是一些主观因素。对于一个文本要得到它是否支持某种观点的信息，而不是一些简单的客观内容，这种独特的文本分类任务被称为“情感文本分类”。根据分类的粒度不同，情感文本分类可以分为短语级、句子级和篇章级[1-2]。根据在训练集中标注样本所占的比例，情感文本分类可以大致分为基于半监督学习、基于监督学习和基于无监督学习的情感分析[3]。

* 收稿日期：2016-11-09；修回日期：2017-02-12 Received date:2016-11-09;Revised date:2017-02-12

1　相关工作

目前，情感分析的主要研究方法大致分为两类。一类是基于情感词典及规则的方法，另一类是目前使用较多的基于机器学习的方法。

Turney 等 [4] 针对情感字典的不足，使用 PMI 方法对基准字典进行了扩充；李寿山等 [5] 利用标签传播算法构建覆盖领域语境的中文情感词典，用于文本情感分析；张婧等人 [6] 建立基于二元语法依赖关系的情感倾向互信息特征模型，通过机器学习方法训练分类器自动识别词语情感极性；杨经等 [7] 通过提取分析情感词的相关特征，使用 SVM 对句子进行情感识别及分类；Pang 等 [8] 尝试使用 n-grams 模型和 SVM 分类模型对情感分类，并选择 unigrams 作为特征来获取最佳分类结果；李素科等 [9] 针对监督学习分类的不足，对情感特征进行聚类，并提出了一种半监督的情感分类算法；陈昀等 [10-11] 提出了基于多特征融合的中文评论情感分析方法，通过 word2vec 和 SVM 进行训练和分类来判断情感倾向，提高了情感分类准确率。然而，文本中词汇的重要程度很少被考虑。事实上，文本分类中特征项权重的赋予对于分类效果有着较大影响 [12-13]，而 TFIDF 算法是权重计算的重要算法之一。基于此，本文提出了基于 TFIDF 的加权 word2vec 情感分析方法。

2　基于加权 word2vec 的情感分析

情感主要分为积极和消极，这样微博文本情感分析可转换为短文本的二分类问题。Word2vec 在文本分类上具有良好性能 [14-16]，主要取决于传统分类模型中的词汇都是独立的、毫无关联的。Word2vec 词向量是根据词汇所在上下文计算出的，充分捕获了上下文的语义信息。针对 word2vec 模型无法区分文本中词汇的重要度问题，进一步借助 TFIDF 算法计算短文本中词汇的权重，提出加权 word2vec 分类模型，将得到的微博文本词向量作为特征向量训练分类器，从而预测待测试微博的情感极性。

2.1　Word2vec 模型

Word2vec 是 Google 在 2013 年年中开源的一款将词表征为实数值向量的高效工具。它利用深度学习思想，通过训练把对文本内容的处理简化为 K 维向量空间中的向量运算 [17]，而向量空间上的相似度可以用来表示文本语义上的相似度。换个思路，把词当做特征，那么 word2vec 就可以把特征映射到 K 维向量空间，从而为文本数据寻求更加深层次特征表示。这使得 word2vec 输出的词向量可以被用来做很多 NLP（Natural Language Processing）相关的工作，如聚类、找同义词、词性分析和短文本分类等。

Word2vec 包含了两种训练模型：Continuous Bag of Words（CBOW）和 Skip-gram。CBOW 的目标是根据上下文来预测给定词，数学表示为：

$$P\left(W_t \mid W_{t-k}, W_{t-k-1}, \cdots, W_{t+k-1}, W_{t+k}\right) \quad (1)$$

其中，W_t 为语料词典中的一个词。CBOW 通过和 W_t 相邻上下文窗口大小为 k 的词来预测词 W_t 出现的概率。Skip-gram 刚好相反，它根据当前词语来预测上下文，数学表示为：

$$P\left(W_{t-k}, W_{t-k-1}, \cdots, W_{t+k-1}, W_{t+k} \mid W_t\right) \quad (2)$$

即通过词汇 W_t 去预测相邻窗口 k 内词汇的概率。

与 CBOW 模型相比，Skip-gram 语义准确率高，代价是模型计算复杂度高，模型训练耗时较长。CBOW 模型因为窗口大小的限制，导致窗口以外的词汇与待预测词的关系不能正确被模型所捕获。如果单纯扩大窗口，又会增加训练模型的耗时。Skip_gram 模型会通过跳跃词汇来构建词组，避免了因窗口大小限制导致丢失语义信息的问题。

考虑到 word2vec 模型无法区分文本中词汇的重要程度，借助 TFIDF 算法计算短文本中词汇的权重。

2.2　TFIDF 模型

TF-IDF 的主要思想：如果某个词或短语在一篇文章中出现的频率高，且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。实际上，TF-IDF 是 TF*IDF，TF 词频（Term Frequency），IDF 逆向文件频率（Inverse Document Frequency）。TF 表示词条在文档 d 中出现的频率。

IDF 的主要思想：如果包含词条 t 的文档越少，也就是 n 越小，IDF 越大，则说明词条 t 具有很好的类别区分能力。

词频（Term Frequency，TF）表示某一个给定的词条 t_i 在文档 d_j 中出现的频率，该词汇 t_i 的 TF 公式为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (3)$$

其中，$n_{i,j}$ 是词 t_i 在文件 d_j 中的出现次数，而分母 $\sum_k n_{k,j}$ 则是在文件 d_j 中所有字词的出现次数之和。

逆文档频率（Inverse Document Frequency，IDF）

是一个词语普遍重要性的度量，某一给定词语 t_i 的 IDF 为：

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \quad (4)$$

其中，$|D|$ 表示语料库中的文件总数，$|\{j:t_i \in d_j\}|$ 表示包含词语 t_i 的文件数目（即 $n_{i,j} \neq 0$ 的文件数目）。如果该词语不在语料库中，就会导致分母为零。因此，一般情况下使用 $1+|\{j:t_i \in d_j\}|$。

词汇 t_i 的 TFIDF 权重为：

$$tfidf_{i,j}=tf_{i,j}*idf_i \quad (5)$$

2.3 加权 word2vec 情感分析

Word2vec 情感分类方法解决了分类词间的语义关系，却忽略了词汇的重要程度；TFIDF 解决了词汇的重要程度，却忽略了词汇间的语义关系。基于此，本文采用基于 TFIDF 加权的 word2vec 情感分类方法，以期提高微博情感分类的准确率。

设有训练语料词典 vocab 和文档 $d_j=\langle w_1,w_2,\cdots w_j\rangle$，$N$ 是词向量维度：

$$\text{vocab}=\{t_i|i \in 1\cdots N\} \quad (6)$$

首先，使用 word2vec 中默认的 Skip_gram 模型训练语料，用训练得到的 word2vec 模型计算文档 d_j 中各词汇的 word2vec 向量。累加文档 d_j 中每个词汇的词向量得到 d_j 的向量表示为 $R(d_j)$：

$$R(d_j) = \sum_t word2vec(t) \text{ where } t \in d_j \quad (7)$$

其中 $word2vec(t)$ 表示词汇 t_i 的 word2vec 词向量。

然后，根据实验微博语料训练 TFIDF 模型，计算每条微博中词汇的 TFIDF 权重，将词汇的 word2vec 词向量乘以对应 TFIDF 权重得到加权 word2vec 词向量。累加文档词汇的加权 word2vec 词向量，得到文档 d_j 新的向量 $W\_R(d_j)$：

$$W\_R(d_j) = \sum_t word2vec(t) \times tfidf_{i,j} \quad (8)$$

将得到的微博文本向量作为特征训练分类器，从而预测待测试语料文本的情感极性（积极和消极）。很多研究表明，与其他分类系统相比，SVM 在分类性能上和系统健壮性上表现出很大优势[18-19]，因此实验选用 SVM 作为分类工具。

SVMperf 是 SVMlight 的开发者 Thorsten Joachims 在 SVMlight 的基础上采用更优化的内核算法得到的新型分类模型。SVMperf 相较于 SVMlight 具有 3 点优势：分类速度更快、分类精度更高、适合大数据集。因此，本文采用 SVMperf 训练测试语料。

文本分类前，一般要经过去停用词等预处理技术。停用词主要包括英文字符、数字、数学字符、标点符号以及使用频率特别高的没有实际意义字符的单汉字如“的、在、和”等。移除文本中的停用词，能改善文本分类效果[20]。

3 实验结果与分析

3.1 实验数据

实验训练 word2vec 模型的语料，来自中文维基百科网站下载的、常用的、未处理的词条正文数据集。情感分析采用来自网络中已有的用户微博情感分析语料，去重后保留 6 000 条数据，包括 3 000 条积极微博和 3 000 条消极微博。测试数据原始微博，如表 1 所示。为了进行实验，将积极和消极的数据集各分为两份，其中 80% 作为训练集，余下 20% 作为测试集。

表 1 实验数据

积极	今天手机报上那个王宝强的哪吒造型，彻底笑喷了 ~~~
	太美啊！太感谢了！为庆贺海棠花儿们的新据点诞生，小宣特此冒死献出独家私房女超人美国太浩湖美图，望花儿们天天都开心。
	股票回暖了，上午够本儿。我今天终于忍住一次没有抛，果然，下午继续飘红。
	# 飘飘龙巴厘岛 # 哦，特别佩服那位帅哥 ~//@ 飘飘龙品牌：回复 @ 闫小东：老鼠，，，我以前还养过二只小老鼠呢，嘿嘿 //@ 飘飘龙品牌：看着怕怕的 ~ 很怕蛇
消极	最讨厌下雨，最讨厌变冷，最讨厌阴霾，啊 ~~~，总之这样天气的珠海，我最最最讨厌了！
	蝴蝶效应？关我们什么事儿啊？？？难道韩国人过来收购白菜啊？太思密达鸟韩国泡菜危机使我国部分地区白菜大涨价。
	牙龈肿痛，连带着整个半边腮帮子还有脖子一侧都很疼，嘴只能张开一公分，水果、米饭、好吃的菜一概吃不了，我只能喝稀饭度日了。哭求大伙儿献出良方搭救搭救我吧！！

3.2 评价标准

本文情感分类的评价指标采用精度（Precision）、召回率（Recall）、F-score。表 2 是两分类分类器的混淆矩阵（Confusion Matrix），其中 TP（True Positive）表示实际为正类、预测也为正类的文本数量；FN 表示实际为正类、预测为反类的文本数量；FP 表示实际为反类、预测为正类的文本数量；TN 表示实际为反类、预测也为反类的文本数量。

表 2　混淆矩阵

数据类别	预测正例	预测反例
实际正例	TP	FN
实际反例	FP	TN

准确率定义为：

$$precision = \frac{TP}{TP+FP} \quad (9)$$

召回率定义为：

$$recall = \frac{TP}{TP+FN} \quad (10)$$

实际应用时，需要平衡准确率和召回率。通常，使用两者的调和平均数作为一个综合的评价指标，称之为 F-score：

$$F\text{-}score = \frac{2 \times precision \times recall}{precision + recall} \quad (11)$$

3.3　微博情感分析

实验训练 word2vec 模型，将下载的维基百科语料经过中文繁体转简体、文本中噪音过滤等处理后，经 ICTCLAS 分词，共提取出词汇 672 135 个。Word2vec 模型参数向量维度为 400，窗口大小为 20，其他参数都为默认。

实验针对微博语料，先删除非用户微博内容，保留用户个人的微博正文；采用 ICTCLAS 分词后，将各条微博文本分词去停用词。测试数据处理前后，对比如表 3 所示。

表 3　测试数据处理前后对比

原始微博文本	# 飘飘龙巴厘岛 # 哦，特别佩服那位帅哥 ~//@ 飘飘龙品牌: 回复 @ 闫小东: 老鼠，，，我以前还养过二只小老鼠呢，嘿嘿 //@ 飘飘龙品牌：看着怕怕的 ~ 很怕蛇
处理后微博文本	特别 佩服 那位 帅哥

将处理后的微博文本作为语料训练TFIDF模型。先计算微博文本中各个词汇的 TF 值及该词汇在语料中的 IDF 值，相乘后得到文本中各词汇的 TFIDF 权重。然后，使用训练得到的 word2vec 模型计算词汇的 word2vec 向量，将词汇 word2vec 词向量乘以对应的 TFIDF 权重，得到加权 word2vec 词向量，累加加权 word2vec 词向量表示每一条微博文本。

将微博文本的加权词向量作为特征向量，采用 SVM 情感分类算法训练分类器，对已标注的情感数据进行分类预测。为了测试该方法的性能，先采用 TFIDF 和未加权的 word2vec 方法对微博数据进行情感分类，再采用基于 TFIDF 加权的 word2vec 方法分类并对比情感分类效果，对比结果如表 4 所示。

表 4　对比实验结果

方法	积 / 消极	准确率	召回率	F 值	正确率
TFIDF	积极	0.788	0.795	0.791	0.791
	消极	0.793	0.786	0.789	
word2vec	积极	0.835	0.849	0.842	0.841
	消极	0.846	0.833	0.839	
加权 word2vec	积极	0.885	0.904	0.894	0.894
	消极	0.902	0.883	0.892	

从实验结果可以看出，基于加权 word2vec 和 SVM 的情感分类方法取得了较好的分类效果。TFIDF 模型忽略了词汇间的语义关系；word2vec 模型的分布式词向量及联系上下文的特点能很好地解决微博文本特征稀疏等问题，但不能解决模型中词汇的权重问题。因此，本文结合 TFIDF 模型，提出加权 word2vec 模型，通过文本分类来分析微博内容的情感倾向。从实验结果可以看出，与 TFIDF、word2vec 模型相比，加权 word2vec 模型在分类精确率、召回率、*F* 值和正确率方面都有所提高。

4　结　语

对比已有的情感分析方法，加权 word2vec 方法主要根据词汇间的语义信息和词汇在语料中的权重，采用 SVM 的分类方法，对微博内容进行训练分类，取得了较好的实验结果。如果能考虑到微博之间的文本相似度，相信能取得更好的情感分析结果，这有待后续进一步的研究和实验。

参考文献：

[1] 杨立公，朱俭，汤世平 . 文本情感分析综述 [J]. 计算机应用，2013,33(06):1574-1607.
YANG Li-gong,ZHU Jian,TANG Shi-ping.Review on Text Emotion Analysis[J].Journal of Computer Applications,2013,33(06):1574-1607.

[2] 魏韡，向阳，陈千 . 中文文本情感分析综述 [J]. 计算机应用，2011,31(12):3321-3323.
WEI Wei,XIANG Yang,CHEN Qian.Survey on Chinese Text Sentiment Analysis[J].Journal of Computer Applications,2011,31(12):3321-3323.

[3] 黄胜 .Web 评论文本的细粒度意见挖掘技术研究 [D]. 北京：北京理工大学，2014.
HUANG Sheng.Research on Fine-grained Opinion Mining Technologies of Web Review Texts[D]. Beijing:Beijing Institute of Technology,2014.

[4] Turney P D,Littman M L.Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J].

ACM Transactions on Information Systems(TOIS),2003,21(04):315–346.

[5] 李寿山 , 李逸薇 , 黄居仁等 . 基于双语信息和标签传播算法的中文情感词典构建方法 [J]. 中文信息学报 ,2013,27(06):75–82.
LI Shou–shan,LI Yi–wei,HUANG Ju–Ren,et al.Construction of Chinese Sentiment Lexicon using Bilingual Information and Label Propagation Algorithm[J].Journal of Chinese Information Processing,2013,27(06):75–82.

[6] 张靖 , 金浩 . 汉语词语情感倾向自动判断研究 [J]. 计算机工程 ,2010,36(23):194–196.
ZHANG Jing,JIN Hao.Study on Chinese Word Sentiment Polarity Automatic Estimation[J].Computer Engineering,2010,36(23):194–196.

[7] 杨经 , 林世平 . 基于 SVM 的文本词句情感分析 [J]. 计算机应用与软件 ,2011,28(09):225–228.
YANG Jing,LIN Shi–ping.Emotion Analysis on Text Words and Sentences based on SVM[J].Computer Applications and Software,2011,28(09):225–228.

[8] Pang B,Lee L,Vaithyanathan S.Thumbs up?:Sentiment Classification Using Machine Learning Techniques[C].Proceedings of the ACL–02 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics,2002:79–86.

[9] 李素科 , 蒋严冰 . 基于情感特征聚类的半监督情感分类 [J]. 计算机研究与发展 ,2013,50(12):2570–2577.
LI Su–ke,JIANG Yan–bing.Semi–Supervised Sentiment Classification based on Sentiment Feature Clustering[J].Journal of Computer Research and Development,2013,50(12):2570–2577.

[10] 陈昀 , 毕海岩 . 基于多特征融合的中文评论情感分类算法 [J]. 河北大学学报 : 自然科学版 ,2015,35(06):651–656.
CHEN Yun,BI Hai–yan.A Sentiment Classification Algorithm of Chinese Comments based on Multi Features Fusion[J].Journal of Hebei University(Natural Science Edition),2015,35(06):651–656.

[11] Zhang D,Xu H,Su Z,et al.Chinese Comments Sentiment Classification based on Word2vec and SVM Perf[J].Expert Systems with Applications,2015,42(04):1857–1863.

[12] 施聪莺 , 徐朝军 , 杨晓江 .TFIDF 算法研究综述 [J]. 计算机应用 ,2009(S1):167–170,180.
SHI Cong–ying,XU Chao–jun,YANG Xiao–jiang.Study of TFIDF Algorithm[J].Journal of Computer Applications,2009(S1):167–170,180.

[13] 侯艳钗 . 基于词语权重的中文文本分类算法的研究 [D]. 天津 : 河北工业大学 ,2011.
HOU Yan–chai.Term Weight–based Chinese Text Classification Algorithm[D].Tianjin:Hebei University of Technology,2011.

[14] Lilleberg J,Zhu Y,Zhang Y.Support Vector Machines and Word2vec for Text Classification with Semantic Features[C].Cognitive Informatics & Cognitive Computing(ICCI*CC),2015 IEEE 14th International Conference on,2015:136–140.

[15] Wolf L,Hanani Y,Bar K,et al.Joint Word2vec Networks for Bilingual Semantic Representations[J].International Journal of Computational Linguistics and Applications,2014,5(01):27–44.

[16] 苏增才 . 基于 word2vec 和 SVMperf 的网络中文文本评论信息情感分类研究 [D]. 石家庄 : 河北科技大学 ,2015.
SU Zeng–cai.Research on Sentiment Classification for Chinese Online Comment Texts based on word2vec and SVMperf[D].Shijiazhuang:Hebei University of Science & Technology,2015.

[17] Mikolov T,Chen K,Corrado G,et al.Efficient Estimation of Word Representations in Vector Space[J].Computer Science,2013,25(05):213–219.

[18] 徐易 . 基于短文本的分类算法研究 [D]. 上海 : 上海交通大学 ,2010.
XU Yi.Research of Text Classification Algorithm based on Short Text[D].Shanghai:Shanghai Jiaotong University,2010.

[19] 李玲俐 . 数据挖掘中分类算法综述 [J]. 重庆师范大学学报 : 自然科学版 ,2011(04):44–47.
LI Ling–li.A Review on Classification Algorithms in Data Mining[J].Journal of Chongqing Normal University(Natural Science),2011(04):44–47.

[20] Patel B,Shah D.Significance of Stop Word Elimination in Meta Search Engine[C].Intelligent Systems and Signal Processing(ISSP),2013:52–55.

作者简介：

李　锐（1992—），男，硕士，主要研究方向为数据挖掘、信息安全；

张　谦（1987—），男，博士，主要研究方向为数据挖掘、信息安全；

刘嘉勇（1962—），男，博士，教授，主要研究方向为信息安全理论与应用、网络信息处理与信息安全、大数据分析。