Computer Science and Application 计算机科学与应用, 2018, 8(5), 669-686
Published Online May 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.85076

Hans汉斯

Research and Analysis of Textual Multi-Emotion Based on Deep Learning

Nan Chen, Jincai Chen, Ping Lu

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan Hubei
Email: 2550566043@qq.com

Received: Apr. 29th, 2018; accepted: May 16th, 2018; published: May 23rd, 2018

Abstract

Text emotional analysis is mainly based on text mining technology for emotional analysis and processing of the text with tendentiousness, which is the subjective text tendency recognition process of positive and negative, neutral. This text about emotional particle division is not sufficient, not comprehensive, is too stiff and violence, and not only cannot effectively reflect the text sentiment granules with different strength and size, but also needs a lot of manual annotation. This paper proposes and constructs the multiple text sentiment data Co-Training based on semi supervised training set, and combines with the emotion of frequency, emotion dictionary, emotion semantic information to construct the D & W, T & W, SSW three kinds of emotion word vector. Finally, CNN and LSTM neural network structure model is used to construct multivariate data sets that were compared with the training and optimization model of emotion word vector, which verifies the validity of the emotional word vector, but also improves the accuracy of text sentiment classification.

Keywords

Emotion Analysis, Text Classification, Multiple Emotion, Emotional Word Vector

基于深度学习的多元文本情感研究与分析

陈　楠，陈进才，卢　萍

华中科技大学武汉光电国家研究中心，湖北　武汉
Email: 2550566043@qq.com

收稿日期：2018年4月29日；录用日期：2018年5月16日；发布日期：2018年5月23日

摘　要

文本情感分析主要是通过文本挖掘技术对带有倾向性的文本进行情感分析和处理，识别其中主观性文本

文章引用: 陈楠, 陈进才, 卢萍. 基于深度学习的多元文本情感研究与分析[J]. 计算机科学与应用, 2018, 8(5): 669-686.
DOI: 10.12677/csa.2018.85076

的倾向是正面、负面、中性的过程，这种关于文本情感颗粒的划分是不充分的，不全面的，显得过于生硬和暴力，不仅不能有效地体现出不同的文本情感颗粒的强度和大小，而且还需要大量的人工标注。本文针对此问题，提出和构建了基于Co-Training半监督训练的多元文本情感数据集，并且结合情感词频、情感词典、情感语义信息构建了D & W、T & W、SSW三种情感词向量，最后利用CNN和LSTM神经网络结构模型分别对构建的多元数据集进行了情感词向量的对比训练和模型优化，从而验证了情感词向量的有效性，而且提升了文本情感分类的准确度。

关键词

情感分析，文本分类，多元情感，情感词向量

1. 引言

随着移动互联网和大数据时代的到来，爆炸式的海量信息急需要我们去处理分析，基于文本的情感分析技术也越来越成为研究热点，作为互联网的主体，每一条文本都带着我们的主观情绪、每种情绪也是各不相同的，比如喜爱，愤怒，悲伤，难受，赞扬，中立等。情感分析又称作情感挖掘或者意见挖掘，它主要包含的研究内容是情感信息分类任务、情感信息抽取任务等，传统的情感分析方法主要包括基于机器学习算法的研究和基于情感词典的构建的研究，基于情感词典的情感分析方法存在覆盖率不足的缺点，基于机器学习算法研究的情感分析会存在特征选择困难，人工标注训练集困难，模型简单，系统可扩展性不足，准确率低等缺点。

为了定量地得到文本内容的情感倾向，避免情感二元极性划分带来的情感不充分的，不全面的问题，本文构建了针对多元情感分析的语义情感数据集，结合 Co-Training 标注方法，减少了单纯进行人工标注带来的误差，增强了模型数据的健壮性和泛化能力，为了更加有效地结合文本的情感信息，抽取更加深层次的语义情感信息，本文构建了更加有效的情感词向量模型算法，将情感词典和情感词频-逆文档概率，情感词向量三者有效地结合起来，充分地考虑到了三者的优点，提出了结合情感词向量和情感词典的 D & W 词向量、结合情感词向量和情感词频的 T & W 词向量、结合情感词频和情感词典以及情感词向量的 SSW 情感语义词向量，并且将其应用在深度学习模型的词向量特征表示上面，进行模块叠加训练，实验证明对比基准词向量模型，其能够在文本情感分类任务中取得更加优秀的效果。

2. 相关研究

2.1. 深度学习

近些年，深度学习在语音识别、自然语言处理、机器视觉、图像处理等领域取得了巨大成功，在 1986 年，Hinton [1]提出了非常著名且沿用至今的反向传播算法，使得基于深度神经网络的深度学习(Deep Learning)的方法应运而生，从此神经网络变得非常流行起来，利用神经网络来建立语言模型的研究思路逐渐走向成熟，大大提升了文本的特征质量。2003 年，Bengio 等提出用神经网络的方法去构建二元语言模型[2]。2008 年，Ronan Collobert 和 Jason Weston 推出 SENNA 系统[3]，并将其应用到自然语言处理领域中，利用词向量的方法完成了其中的词性标注、命名实体识别、短语识别、语义角色标注等多种任务。

2013 年，随着 Hinton 提出 word embedding 的概念[4]，以及 Mikolov 对该理论的进一步实现[5]，这种全新的文本特征表示已经被越来越多的研究者所认可。基于 word embedding 的特征表示方法不但能够避免“维度灾难”现象，还能够从更高的语义层面上描述词与词之间的关系。梁军等人通过采用自动编码器，实现了利用半监督学习的方法对微博的文本数据进行情感分析，大量减少了人工标注的工作量[6]。陈翠平引入了深度学习的思想来完成文本分类任务，利用深度信念网络自动提取文本特征[7]。Yoon 等尝试利用卷积神经网络结构来解决情感分析和问题分类等若干自然语言处理任务，获得了非常好的效果[8]。

2.2. 情感分析

情感分析的目的是将具有情感倾向的主观性文本识别出来，并且分为褒义和贬义两类。其中在传统的情感分析方法中，主要采用基于规则的方法，需要相当一部分人力和物力作为支撑，所以，现在情感分析研究领域的学者纷纷转向了基于统计的学习方法，该方法主要根据特征的分布对文本的情感类别做出正确的判断。Pang 等在对电影评论数据进行褒贬二分类的研究中，使用了包括一元词、二元词、词性标注等若干特征[9]，Davidov 等利用在 Twitter 中的标签元素和笑脸符号来作为特征，从而对 Twitter 进行情感分类[10]，李婷婷等尝试从文本数据中人工构建若干特征，再利用传统的机器学习方法进行文本分类[11]；李荣陆等人利用最大熵模型实现了中文文本分类[12]。Taboada 等[13]采用的是基于词库的方法，文本的最后情感值采用集约化的方法计算，进而确定文本的最后情感倾向。Hu 等[14]在文章中提出采用 Bootstrapping 策略，句子中所有情感词的情感倾向性分数总和决定最后该句子的情感倾向。以上方法本质上均属于机器学习范畴，其分类效果严重依赖所构建特征的质量和模型参数的调优，整个过程非常耗时耗力，往往需要大量的领域内知识，因此最终的分类效果并不稳定。

3. 多元情感数据集的构建

3.1. 多元情感分析

文本情感分析是指对于网络用户的喜爱、观点和意见的分析和挖掘，获得用户对于事件的主观性情感倾向评论，我们通常用情感权重来代表某个词语的情感极性，如果情感权重大于 0 则代表用户的积极正面的情感倾向，如果情感极性权重小于 0 则代表主观用户的负面消极的情感倾向，这样简单高效的划分情感便于我们直接的得到情感极性和做出最终的判断任务，同样也带来了关于情感划分颗粒不充分，不全面的问题，情感之间的划分过于暴力和生硬，没有有效的区分不同的文本情感颗粒强度和大小。

关于情感(Sentiment)狭义的定义：情感是对于一个实体或者事件等事物评价的极性，又称极性情感，情感的主要类别包含正面、负面、中性情感颗粒，但是情感之间出现相互影响，相互交割带来的情感交叉现象，对于我们最终的情感分类模型造成相应的误判，基于此问题，本文在基于 Ekman *et al.* (1982)的基础上面考虑加入了相关的中性极性情感，提出了基于情感极性的细颗粒情感分类模型，在基本的正向，负向，中性三种极性情感的基础上将情感极性强度划分成 8 中基本的情感颗粒，分别是高兴、生气、厌恶、悲伤、害怕、吃惊、轻视、中性 8 中基本情感强度，如表 1 所示，最大化的减少文本极性情感强度之间的干扰和误判，我们提出了基于多元情感分析的细颗粒划分指标，充分反应情感持有者的情感强度，不仅包含文本情感的极性分析，而且包含了文本情感的极性强度，将二者充分的进行了融合。

3.2. 情感文本数据集构建

情感数据集对于最终的多元情感分类任务有着重要的表现，考虑到单纯的情感语料库的标签都是针对二元情感分类任务的，不能够满足多元情感分类的要求，因此我们首先要满足构建多元情感任务的情感文本标签。

为了构建多元情感的极性文本，我们提出了借助表情包文字识别技术来构筑多元的情感极性文本，表情已经充斥在网络的各个角落，现在用户都流行 “能发图就不打字”的习惯，表情包文字本身充满了多元的情感极性，我们需要借助这些表情包，从形形色色的网络表情中找出对应的文字，从这些表情中提取出的文字，可用于我们后续的文本分析，情感预测，语义理解等任务。

目前主流的文字识别方法都差不多。主要分为两个模块，一个模块定位文字位置，另外一个模块针对定位后的文字进行识别。针对这两个模块，我们使用的是 Faster RCNN + CTC 的方案。文字定位部分本文使用了 Faster RCNN 技术，如图 1 所示，其是从 RCNN 逐渐演变过来的。相对于它的前辈 RCNN 以及 Fast RCNN，Faster RCNN 提出了 RPN (Region Proposal Network)网络。通过 RPN 输出 Anchor Box Proposals，再通过 NMS 和其他一些方法进行 Proposals Reduction。该方法对比以往的方案，性能更优，减少了 selective search 里面繁琐的计算。目前目标检测还有其他 state of art 的定位方案，例如 YOLOv2，Mask RCNN 等，其中 Mask RCNN 更多聚焦在 image segmentation 上。关于文字识别部分的结果，如图 2 和图 3 所示。

Table 1. Basic emotion and extended emotion
表 1.基本情感和扩充情感

Basic	Extended
高兴(Happiness)	Depressed(郁闷)
生气(Anger)	boring(无聊)
厌恶(Disgust)	Lonely(孤独)
悲伤(Sadness)	irritated(烦躁)
害怕(Fear)	envy(嫉妒)
吃惊(Surprise)	Disappointed(失望)
轻视(Contempt)	regret(后悔)
中性(Neutral)	Like(喜欢) hope(希望)

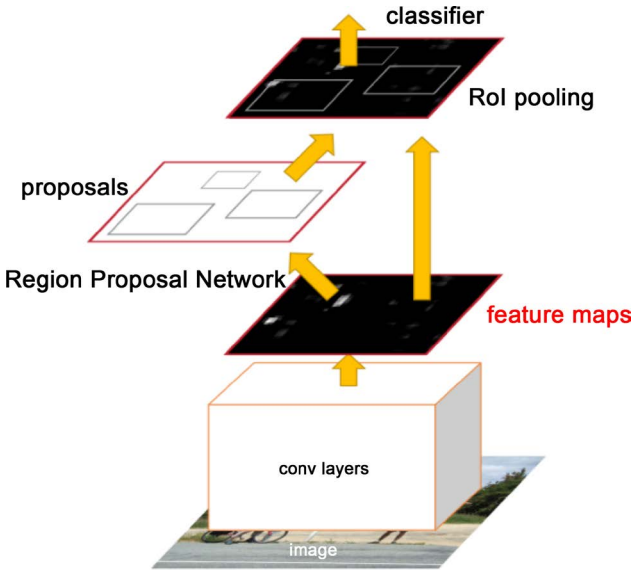

Figure 1. The basic structure of Faster RCNN
图 1. Faster RCNN 的基本结构

Figure 2. Text captured after positioning
图 2. 文字定位后截取的图片

```
filename=8c4cdedc7ff5fb244555dbe4faefcd3c_crop1.jpg,ocr=给公子们请安,probs=1.000,1.000,1.000,1.000,0.999,1.000,trust=True
filename=2208517827a73f92b2d56ee05ecd3532_crop1.jpg,ocr=面对疾风吧.,probs=1.000,1.000,1.000,0.999,1.000,0.994,trust=True
filename=acd96debcfe24b81a5765f53dfb73a67_crop1.jpg,ocr=你恶心到我,probs=1.000,1.000,1.000,1.000,1.000,trust=True
filename=2be3bf2321d0c27ce8df4d46054f39f4_crop1.jpg,ocr=都输光,probs=0.912,1.000,1.000,trust=True
filename=5049c9292d2f430db7708e210ed5fe97_crop1.jpg,ocr=有种报警,probs=1.000,1.000,1.000,1.000,trust=True
filename=3cf9ca606a2718c629458e879b51349e_crop1.jpg,ocr=棒棒棒,probs=0.931,0.790,0.792,trust=False
filename=c76be8a9ef46273611b12252f2a1a29c_crop1.jpg,ocr=只有学习能让我快乐,probs=1.000,1.000,1.000,0.984,1.000,0.987,0.663,0.998,1.000,trust=False
filename=043b809934a8bc016c58359663c66220_crop1.jpg,ocr=我是仙女,probs=1.000,1.000,1.000,0.999,trust=True
filename=6f559c022990d66e831b5a63ff9033e2_crop1.jpg,ocr=你在逗我吗,probs=1.000,1.000,0.999,1.000,1.000,trust=True
filename=81be8a99e38956ae88f33c18ebd837d2_crop1.jpg,ocr=决斗吧.,probs=0.534,0.944,1.000,0.874,trust=False
filename=7d7cf31579f7dd8a7432eb047eed5585_crop1.jpg,ocr=拿起电话联系我,probs=1.000,0.993,1.000,1.000,1.000,1.000,1.000,trust=True
filename=40c0d2b82108668e7cfd0c783ea056c0_crop1.jpg,ocr=只有吃能让我快2,probs=0.999,1.000,1.000,1.000,1.000,1.000,0.999,0.528,trust=False
filename=a8b96bed7ba18929774000f0362e6a68_crop1.jpg,ocr=你真的很不错,probs=1.000,1.000,1.000,1.000,0.996,1.000,trust=True
filename=950191b5fe14af04abf40b8869d74c5d_crop1.jpg,ocr=扎心不老铁,probs=0.841,0.997,0.999,0.992,0.620,trust=False
filename=bccb4f6ee83fbd8f604509ee54f410a5_crop1.jpg,ocr=对对对,probs=0.999,0.999,0.999,trust=True
filename=0cd1f2a84fcba85e51948bf6b374c05d_crop1.jpg,ocr=老婆哔我错了,probs=1.000,1.000,0.829,1.000,1.000,1.000,trust=False
filename=4eacf38bb7baa3b74ed9d9e5d837685c_crop1.jpg,ocr=巨龙盯上你了,probs=0.994,1.000,0.989,1.000,1.000,1.000,trust=True
filename=f68b3f97e1831d44ed11729b1af3bb1e_crop1.jpg,ocr=打死你,probs=1.000,1.000,1.000,trust=True
filename=70fc7ff37795ece5f95039f62209b8e1_crop1.jpg,ocr=群主在泡好,probs=1.000,0.992,1.000,0.999,0.523,trust=False
filename=c447f9af9df0e4cafc1f86b5507cf90f_crop1.jpg,ocr=你坏坏：,probs=0.999,0.999,0.992,0.392,trust=False
```

Figure 3. CTC model text output
图 3. CTC 模型文字输出

3.3. 基于半监督的多元情感数据集的构建

由于直接使用无监督聚类机器学习算法带来的人工工作量大，耗时的问题，本文提出了借助 Co-training 思想和半监督思想来进行数据集的构建。

Co-training 是目前很流行的一种半指导机器学习的方法，它的基本思想是：构造两个不同的分类器，利用小规模的标注语料，对大规模的未标注语料进行标注的方法。Co-training 方法最大的优点是不用人工干涉，能够从未标注的语料中自动学习到知识。Co-training 方法，是无监督和有监督机器学习两者的一个折中办法，它的原则是：在不牺牲性能的前提下，尽量多的使用未带标数据，它从一个小规模的带标的语料库开始，同时使用大规模的未带标语料来进行学习。这里面，我们 Co-training 使用的文本特征

分别是 Word2vec 语义特征样本选择和加权的 Word2vec 文本语义特征：Word2vec * TF-idf，这两种文本特征选择来分别对文本进行特征选择过程，分类器使用的是 SVM 分类器，得到的情感类别比重如图 4 所示。

其中要先选择表情包图片数据集和种子图片的类别和数量；表情包的图片包含了二次元、斗图、纯文字、动物、彩字祝福、真人、未知、七个大的类别，一共大约有 25 万张，每种表情包的数量如表 2、表 3 和表 4 所示：

具体的 Co-training 针对文本特征选择过程和分类标签如下：

1) 先选择一些标签种子文本，这部分文本可以通过手工标注的方式获得；
2) 然后根据预训练的 Word2vec 词向量特征对种子文本进行分类训练；
3) 得到了训练好的分类器；
4) 接着根据训练的 Word2vec * TF-IDF 特征结合训练好的分类器接着对剩下的文本进行标签预测；
5) 所使用的分类器是 SVM 分类器；
6) 这样就得到了待分类文本的标签集。

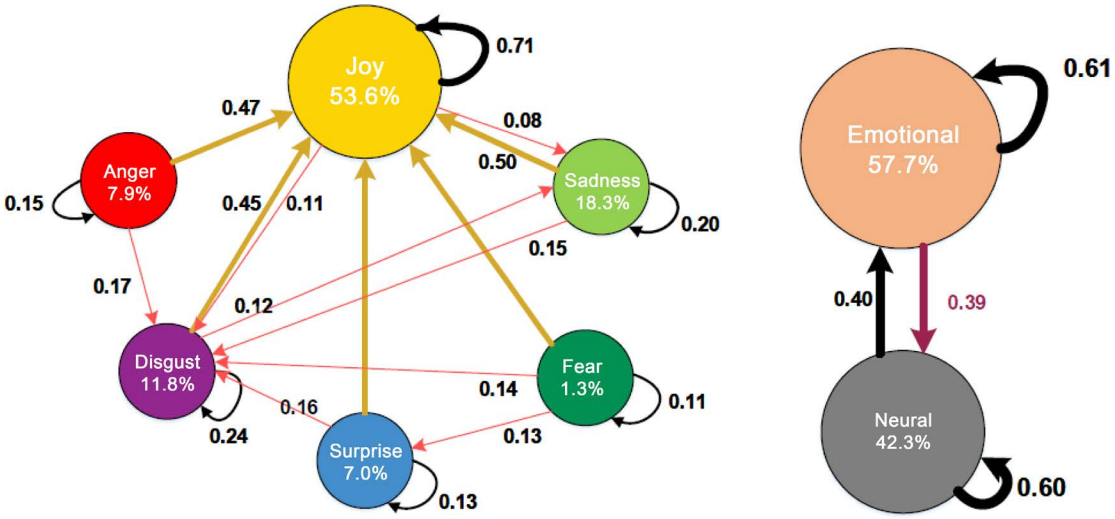

Figure 4. Unsupervised emotional label distribution
图 4. 无监督情感标签分布示意图

Table 2. Expression package picture data set
表 2. 表情包图片数据集

一级分类	数量	比例
二次元	31,639	12.48%
斗图	70,585	27.85%
纯文字	991	0.39%
动物	20,408	8.05%
彩字祝福	14,985	5.91%
真人	52,685	20.79%
未知	62,157	24.52%
合计	253,504	100%

Table 3. Emoticons seed pictures data set
表 3. 表情包种子图片数据集

一级分类	数量	比例
二次元	167	4.2%
斗图	176	4.4%
纯文字	17	0.43%
动物	155	3.96%
彩字祝福	79	2.02%
真人	431	1.10%
未知	2891	73.82%
合计	3916	100%

Table 4. Text seed dataset
表 4. 文本种子数据集

Emotion	Size	Percent
despise	293	7.5%
neutral	1656	42.3%
angry	178	4.6%
sad	413	10.6%
disgust	266	6.81%
surprise	158	4.04%
happy	1211	30.9%

4. 情感语义词向量的构建

在传统的词向量模型基础上，为了更加有效地结合文本的情感信息，抽取更加深层次的语义情感信息，本文构建了更加有效的情感词向量模型算法，将情感词典和情感词频-逆文档概率，情感词向量三者有效地结合起来，充分地考虑到了三者的优点，本文提出了结合情感词向量和情感词典的D & W词向量、结合情感词向量和情感词频的T & W词向量、结合情感词频和情感词典以及情感词向量的SSW情感语义词向量。

4.1. 基于机器算法的情感词典的构建

目前的情感词典主要是包含正向和负向两种情感，为了满足多元情感分析任务，本文根据之前提出的在基本的正向，负向，中性三种极性情感的基础上将情感极性强度划分成8中基本的情感颗粒，结合机器学习的相应方法，减少人工手工标注的工作量，自动化构建相应的多元情感词典。

本文提出了结合词向量的相似性大小，用词聚类的方法来判断文本词语的情感极性大小，具体的，就是先通过人工选择一些情感词语作为我们的种子词语，然后通过词语之间的相似性判断来对情感词典进行扩充，利用种子词语和新的词语之间的关系来计算词语的情感极性大小。

本文的多元情感词典参考了大连理工大学的中文情感词典的构建方法，具体的种子词典的构建是先选择8种情感大类，在这8种基本的大类情感下面再进一步的选择其中的情感文本关键词作为我们的情

感词典基准词，每种情感类别下面的情感基准词的大小数量不一样，但是他们的情感极性得分是统一的，得分的范围是：[−0.8, 0.8]，情感文本基准词的最终构建结果如表 5 所示。

4.2. 基于 Word2vec 与 Tf-idf 的 T & W 词向量研究

我们在文本情感特征的选择和文本情感特征的加权基础上面，结合基于机器学习的特征选择和基于深度学习词向量的特征选择方法，对于文本情感特征的特征进行加权，构建了基于 Word2vec 和 tf-idf 的加权词向量特征构建方法，然后对文本的情感极性进行判断。

这里面把整个 tf-idf 作为衡量整个词语的重要程度权重，Tf-IDF 表示的是一个词语在整个文档中的重要程度，其主要有两个部 c 分组成：词频 TF 和逆文档概率 IDF，主要公式：

$$\text{Tf-IDF} = \text{TF} * \text{IDF}$$

对于 word2vec 词向量部分，我们显示使用 word2vec 中的 CBOW 模型训练语料库，得到相应的词向量模型，然后将文档中的所有对应的相同单词的词向量进行叠加求和得到所有词语的词向量表示：

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

其中：$\text{word2vec}\left(t\right)$表示词语的词向量表示。

接下来我们先后分别取得对应语料库的 Word2vec 词向量和 TF-IDF 词向量的特征表示方法，再将二者有效的进行结合即可得到新的基于 Word2vec 与 Tf-idf 的词向量表示。

$$\text{W\_R}\left(d_j\right) = \sum_t \text{word2vec}\left(t\right) * \text{tfidf}_{i,j}$$

这样，我们就结合了 Word2vec 的词向量表示的词语语义信息，同时，我们又将 tf-idf 表达的词频信息进行了有效的结合起来，这样新的词向量用来提高文本的多元情感分类的准确率。

Table 5. Emotion dictionary examples
表 5. 情感词典实例

情感类别	情感词语	情感极性	情感数量
高兴(Happiness)	喜悦、开心、欢喜、称心、振奋、快活、欢乐、快乐、起劲、兴奋、愉悦、安乐、	0.8	12
生气(Anger)	愤怒、起火、生机、活气、发怒、不满、负气、活力、朝气、发火、赌气、动怒	-0.6	14
厌恶(Disgust)	憎恶、嫌恶、恨恶、厌烦、可恶、讨厌、憎恨、腻烦	-0.8	8
悲伤(Sadness)	心酸、悲哀、哀思、酸楚、难过、痛苦、伤心、痛心、	-0.5	8
害怕(Fear)	惊恐、畏惧、畏缩、畏怯、	-0.7	12
吃惊(Surprise)	惊讶、惊异、惊奇、受惊、惊诧、惊呀、惊愕、诧异	0.2	8
轻视(Contempt)	小看、渺视、鄙夷、看不起、藐视、轻蔑、忽视、歧视、小瞧、轻茂、鄙视、看轻、	-0.4	24
中性(Neutral)	生活、爱情、活动、情况…	0	100

4.3. 基于 Word2vec 与情感词典结合的 D & W 词向量研究

本文基于上面介绍的情感词典得到了情感词语的情感极性得分大小，同时，基于情感的词向量得到了情感词语语义表示，为了将二者有效的进行结合，本文提出了基于情感词向量和基于情感词典的词向量构建过程，将情感词语的情感极性强度和情感词语的词向量结合起来，改进了传统的词向量只包含情感词语单纯的语义信息的构建模型，以此来希望提升我们最后的文本情感分析准确率。

本文结合词向量的相似性大小，用词聚类的方法来判断文本词语的情感极性大小，具体的，就是先通过人工选择一些情感词语作为我们的种子词语，然后通过词语之间的相似性判断来对情感词典进行扩充，利用种子词语和新的词语之间的关系来计算词语的情感极性大小得到了文本词语的情感词典语料，然后结合情感词语的词向量构建了最终的情感语义词向量。

具体的词向量构建过程为：

1) 将语料库的文本词语进行预训练过程；

2) 对于语料库里面的词语选中一部分作为种子词典，并且给出这些种子语料的情感极性得分词典；

3) 对于每种情感分类的情感种子词语进行加和求平均得到每种情感的中心词向量；

4) 将每种情感得到的平均词向量作为虚拟的中心词语聚类中心，计算每个新词与聚类中心的词向量相似度大小，

5) 并且将每个新词与聚类中心的词向量相似度大小进行从小到大的排序；

6) 选择上一步得到的最大相似度的情感类别作为新词的情感类别；

7) 并且根据新词与聚类中心词语的相似度大小与种子文本得情感极性得分得到最终的新词的情感极性权重得分；

8) 这样就得到了新词的情感极性得分。

4.4. 基于 Word2vec 与 Tf-idf、情感词典结合的 SSW 词向量研究

传统的词向量模型根据中心词的上下文来表示中心词，表达出了词向量的语义信息，这样的分布式词向量可以简单的处理基本的文本任务，但是在基于文本的情感分析任务处理中并不能够有效的结合文本的情感信息，所以这样的词向量语义信息是一种不包含情感信息的浅层语义信息，所以为了更加有效的结合文本的情感信息，抽取更加深层次的语义情感信息，本文提出的基于情感词典和 Tf-idf (词频-逆文档)概率，word2vec 词向量三者有效的结合起来，充分的考虑到了三者优点，将词频，词义，词性三者结合起来，并且将其应用在深度学习模型的词向量特征表示上面，实验证明对比以前的词向量模型，有效的提升了文本分类效果。

Word2vec 的词向量模型虽然在更高的层次上面得到了文本的语义特征词向量，但是 Word2vec 的两种模型：基于 CBOW 和 Skip-gram 的词向量算法都没有能够将文本的情感信息有效的蕴含起来，所以，在构建新的语义情感的词向量上面需要先将情感词典，文本词向量的语义信息有效的结合起来，本文先分别基于机器学习算法构建了新的情感词典，对比以前的情感词典构建过程更加的直接有效，然后结合基本的词频-逆文档特征选择算法，Word2vec 算法将文本的词频信息，情感信息，语义信息有效的结合起来并且将新构建的情感语义词向量与原来的词向量进行了对比试验，有效的证明了本文提出的情感语义词向量的有效性。

具体的词向量构建过程为：

1) 将语料库的文本词语进行预训练过程；选择相应的文本特征进行特征训练过程，分别得到对应的 TF-IDF 和 Word2vec 特征词向量；

2) 对于语料库里面的词语选中一部分作为种子词典，并且给出这些种子语料的情感极性得分词典；同时改进相应的 TF-IDF 特征选择过程，得到新的改进的 TF-IDF 特征选择词向量；

3) 分别将得到的文本的改进的 TF-IDF 特征词向量与 Word2vec 特征词向量的结果保存起来；

4) 对于每种情感分类的情感种子词语进行加和求平均得到每种情感的中心词向量；

5) 将每种情感得到的平均词向量作为虚拟的中心词语聚类中心，计算每个新词与聚类中心的词向量相似度大小，并且进行从小到大的排序

6) 选择上一步得到的最大相似度的情感类别作为新词的情感类别；

7) 并且根据新词与聚类中心词语的相似度大小与种子文本得情感极性得分得到最终的新词的情感极性权重得分。

8) 将之前得到的情感极性得分与改进的 TF-IDF 特征词向量、Word2vec 特征词向量的乘积作为最终的情感语义特征词向量。

5. 基于深度学习的多元情感分析实验

传统的 CNN 和 RNN 模型只能解决简单的句子级别文本情感分类，而我们的神经网络模型需要结合文本的情感和语义信息，进行单词词向量的联合训练，因此本文提出了新的修改升级的神经网络模型，将文本的情感和语义信息都包含进行，进行叠加模块训练，因此新的 CNN 和 RNN 神经网络包含 3 个层次模块的文本信息，分别是文本的词频特征信息、情感特征信息、语义特征信息、对应的相关新的 CNN 模型分别是 CNN-DW、CNN-TW、CNN-SSW，同样的，相应的 RNN 模型分别是 RNN-DW、RNN-TW、RNN-SSW，同时改进了相关的卷积神经网络结构和递归神经网络结构，提升了实验效果，提高了准确率，从而验证了结合情感词向量和情感词典的 D & W 词向量、结合情感词向量和情感词频的 T & W 词向量、结合情感词频和情感词典以及情感词向量的 SSW 情感语义词向量的有效性。

5.1. 实验数据集

为了比较本文提出的三种词向量模型在多元情感分类模型任务上的效果，我们结合前面第四章得到的多元情感文本数据集进行了相关的实验分析，结合深度学习的 CNN 和 RNN 等模型进行了对比实验效果。

该实验数据为多元的情感类别文本，具体的是多元情感的语义情感数据集标签是与我们之前提出的多元情感分类一致的，一共包含 8 中基本的情感颗粒，最大化的减少文本极性情感强度之间的干扰和误判，这里面我们提出了基于多元情感分析的细颗粒划分指标，充分反应情感持有者的情感强度，不仅包含文本情感的极性分析，而且包含了文本情感的极性强度分析，将二者充分的进行了融合。

我们需要先定义好各种情感极性对应的情感标签，其对应过程如表 6、表 7 和表 8 所示。

其中统计得到相应的情感标签数量如下。

5.2. 实验基准系统

为了比较本文提出的将 CNN 和 RNN 模型与文本的情感和语义信息都包含进来，进行叠加模块训练效果，需要将新的 CNN 和 RNN 神经网络包含 3 个层次模块的文本信息，分别是文本的词频特征信息、情感特征信息、语义特征信息、对应的相关新的 CNN 模型分别是 CNN-DW、CNN-TW、CNN-SSW，同样的，相应的 RNN 模型分别是 RNN-DW、RNN-TW、RNN-SSW。为了比较相应的词向量模型效果，需要先设置基准系统，方便我们进行对比观察结果，相关的实验标准参照词向量模型是将如下两个基本词向量模型进行平均计算：

Table 6. Emotional tag classification
表 6. 情感标签分类

情感标签	情感内容
1	高兴
2	生气
3	厌恶
4	悲伤
5	害怕
6	吃惊
7	轻视
8	中性

Table 7. Various emotional tag statistics
表 7. 各种情感标签统计图

Emotion	Size
Despise	10,309
Neutral	109,634
Angry	13,305
Sad	31,811
Disgust	7397
Surprise	10,936
Fear	2358
Happy	82,442

Table 8. Emotional polarity distribution
表 8. 情感极性分布图

情感极性	文本数量
中性	109, 634 (34%)
非中性	158,558 (66%)

1) CBOW：根据输入的上下文来预测中心词构建的语义词向量模型，

2) Skip-gram：根据输入的中心词的上下文来构建中心词对应的上下文语义的词向量模型。

5.3. 基于 CNN 模型的多元情感分析实验

针对基于卷积神经网络 CNN 模型的多元文本情感分类实验，我们采用的卷积神经网络结构依次如下是：Embedding → Convection → Pooling → Activation，其中 Embedding 和 Convection 之间的 Dropout、以及 Pooling 和 Activation 之间的 Dropout，我们都将其大小设置为 0.20，其中输入层词向量维度大小是 100，并且卷积层的滤波器将相应的输入层词向量矩阵分成 3 个 region，每个 region 分别设置 32 个滤波器，一共采用的数量是 96 个滤波器，滤波器的大小是(2, 3, 4)，就是说将滤波器设置成 3 种状况，分别是 2 * 3 * 32，2 * 4 * 32，3 * 4 * 32 的 3 种状况，每种状况下的滤波器设置得到的每个 region 是 32 维度的

特征向量，池化层，采用的是最大池化层策略，激活层通过 Softmax 激活函数得到 8 中情感类别的概率大小，再与标准的情感类别比较得到误差反向传播即可，如图 5 所示。

5.4. 基于 RNN 模型的多元情感分析实验

因为 CNN 神经网络模型无法获取对应的很长的文本序列信息，这里面我们选择使用双向的 LSTM 长短记忆单元神经网络模型来进行文本词向量的对比试验，Bi-LSTM 在序列标注、命名体识别、seq2seq 等模型有很多场景都有应用，它能够更好的表达文本上下文信息，而且 Bi-LSTM 可以获取变长且双向的的 n 元语法信息，更加适合拟合文本序列信息，如图 6 所示。

我们的网络结构依次是：Embedding → Convection → Pooling → Activation，其中相应的 Embedding 和 Convection 之间的 Dropout、以及 Pooling 和 Activation 之间的 Dropout，我们都将其大小设置为 0.25，其中输入层词向量维度大小是 100，激活函数同样选择 Softmax 激活函数，分别结合 RNN-DW、RNN-TW、RNN-SSW 三种词向量模型得到相应的情感分类结果。

图 7 是 Bi-LSTM 用于分类问题的网络结构原理示意图，其中 LSTM 节点分别是前向和后向的双向 RNN 的输出表示，这里面需要注意的是 LSTM 单元之间的相互连接和 FC 层的表示，同样，最终的输出层使用的是 Softmax 函数输出。

5.5. 实验结果对比

通过与基准实验的对比，获取了多元情感类别在 3 种词向量下面分别通过深度学习实验得到的相应结果如下图 8~图 15 所示。

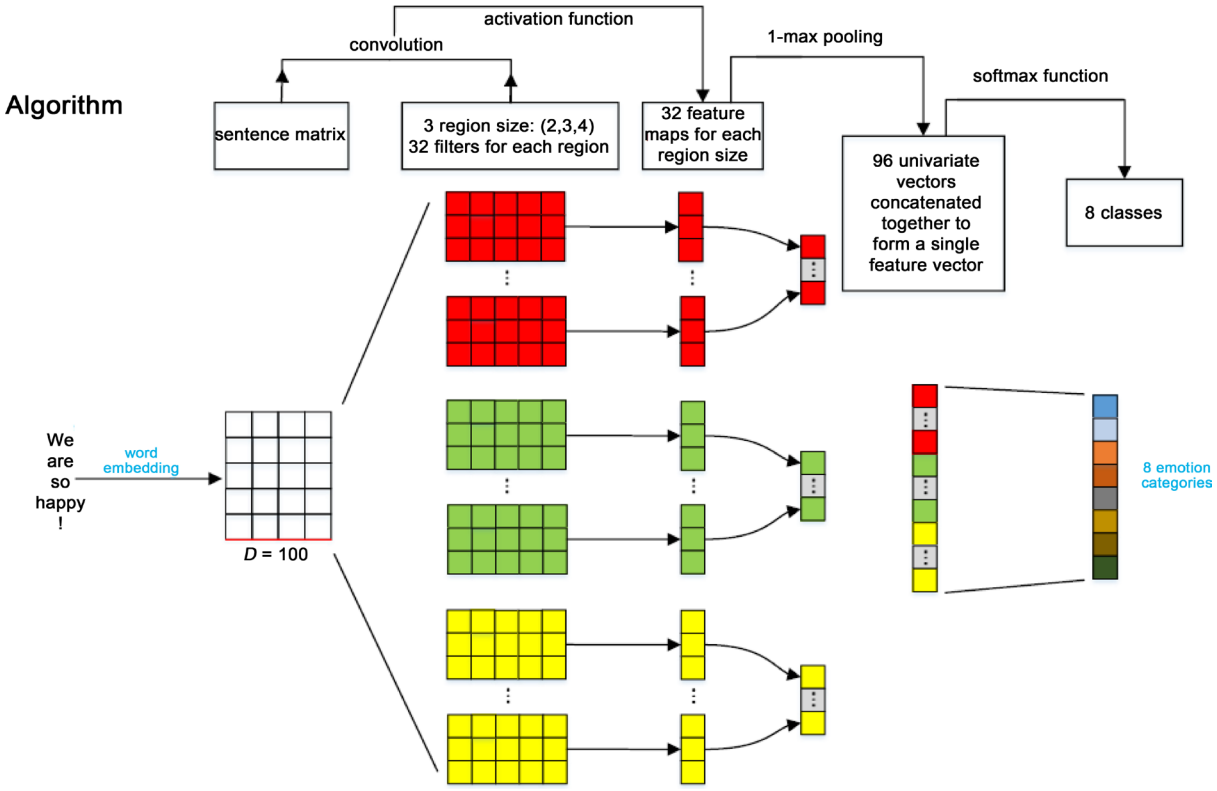

Figure 5. Multivariate sentiment analysis experimental setup based on CNN model
图 5. 基于 CNN 模型的多元情感分析实验设置

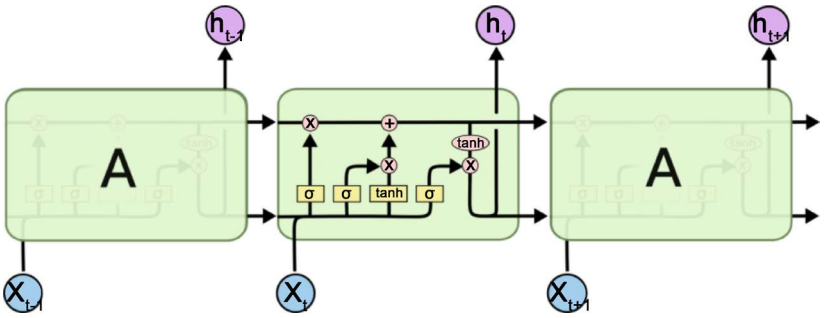

Figure 6. LSTM structure diagram
图 6. LSTM 结构示意图

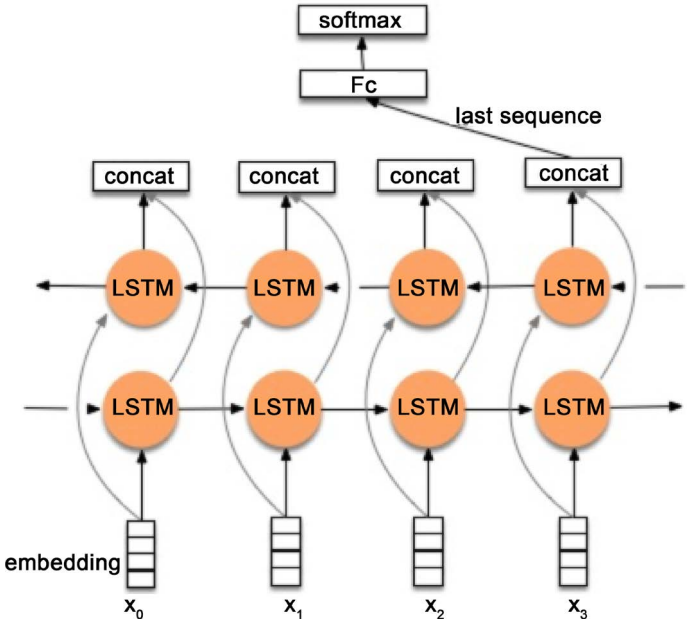

Figure 7. Multivariate sentiment analysis experiment setup based on Bi-LSTM model
图 7. 基于 Bi-LSTM 模型的多元情感分析实验设置

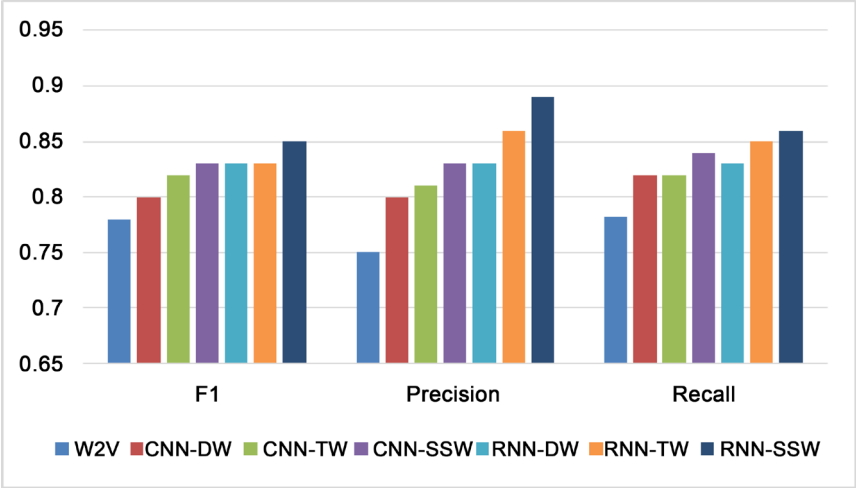

Figure 8. Despise experiment comparison results
图 8. Despise 实验对比结果

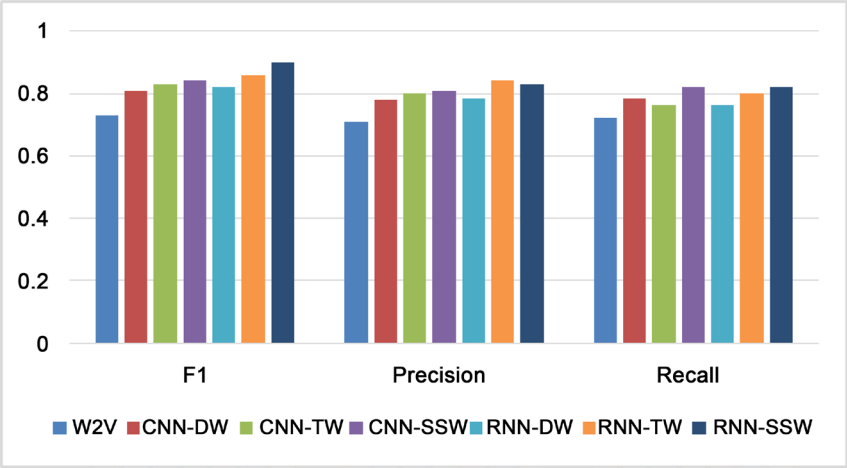

Figure 9. Angry experiment comparison results
图 9. Angry 实验对比结果

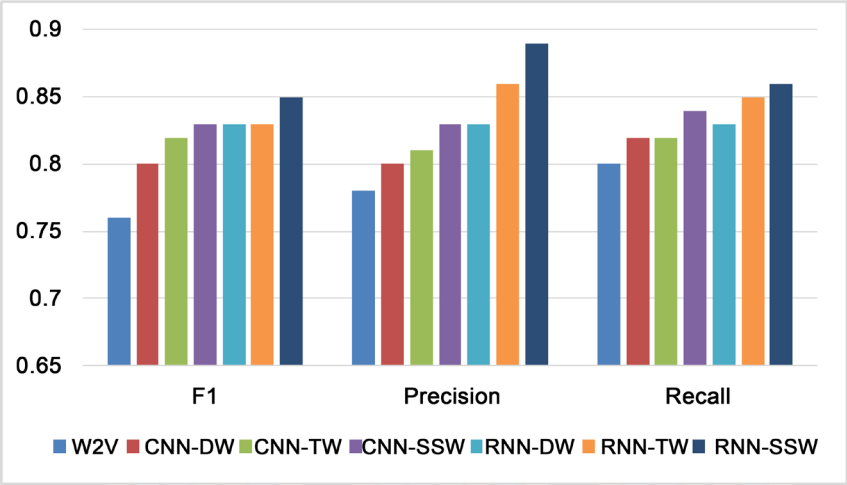

Figure 10. Neutral experiment comparison results
图 10. Neutral 实验对比结果

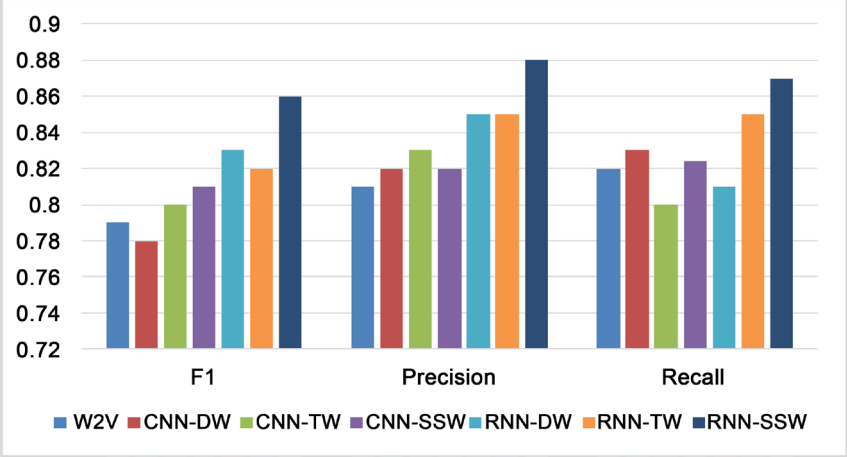

Figure 11. Sad experiment comparison results
图 11. Sad 实验对比结果

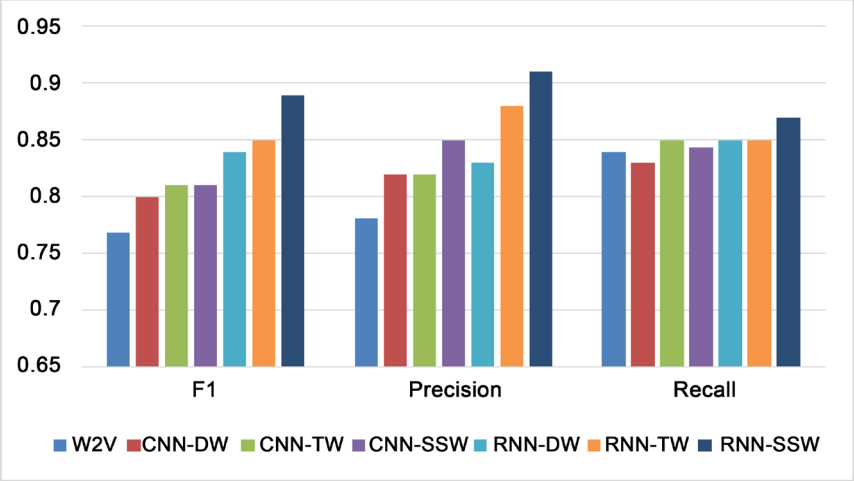

Figure 12. Disgust experiment comparison results
图 12. Disgust 实验对比结果

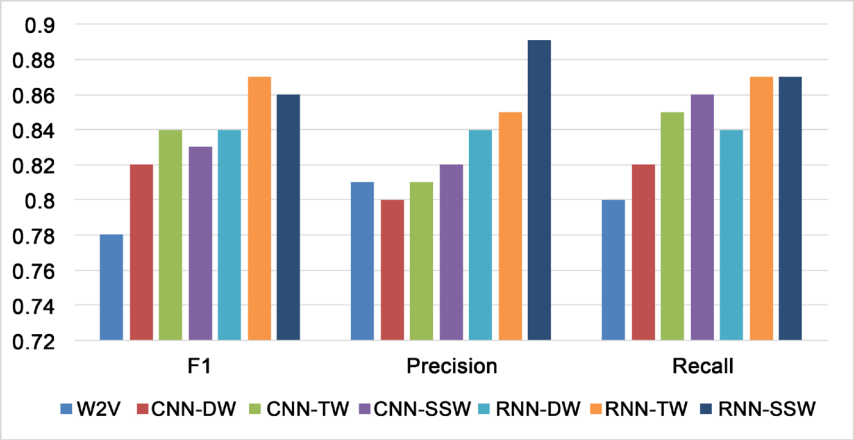

Figure 13. Surprise experiment comparison results
图 13. Surprise 实验对比结果

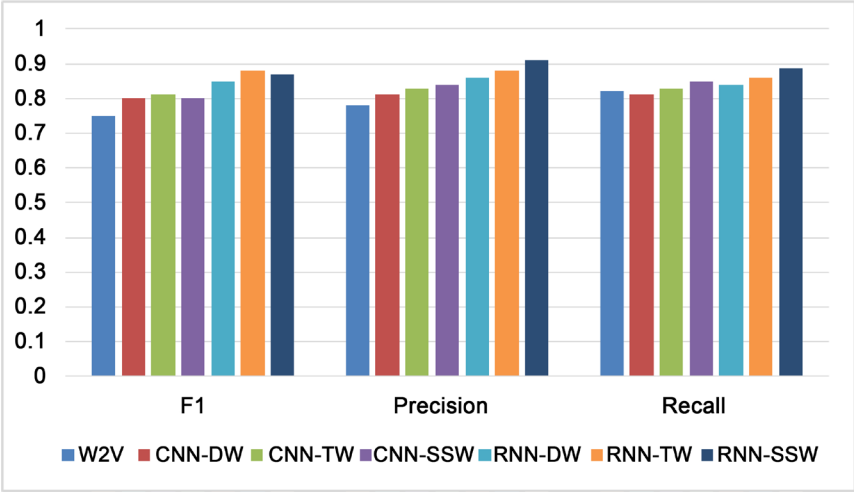

Figure 14. Fear experiment comparison results
图 14. Fear 实验对比结果

为了更加方便的突出我们的实验对比结果，我们计算了多元情感类别在 3 中词向量模型下面的深度学习文本分类实验中的平均 F1 值、Precision、Recall 值作为我们的实验对比指标，得到了相应的实验对比分析结果如下图 16~图 18 所示。

5.6. 实验结论总结

根据上面的词向量实验对比分析，可以简单的得出以下结论：

1) 在构建多元情感数据公开集上面，本文提出的基于情感词典和情感语义的几种词向量在 CNN 和 RNN 深度学习模型上面都取得比基准词向量要好的分类效果，说明将情感信息和词向量信息有效的结合起来，能够在文本情感分类任务中取得更加优秀的效果。

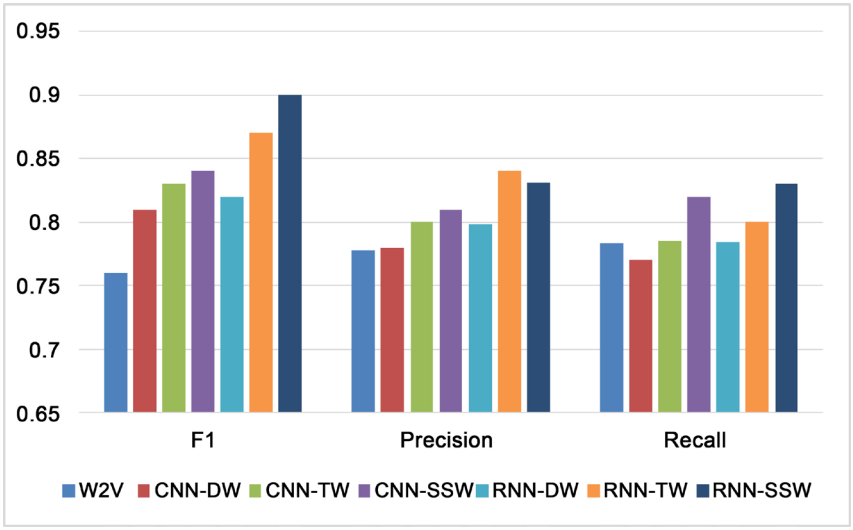

Figure 15. Happy experiment comparison results
图 15. Happy 实验对比结果

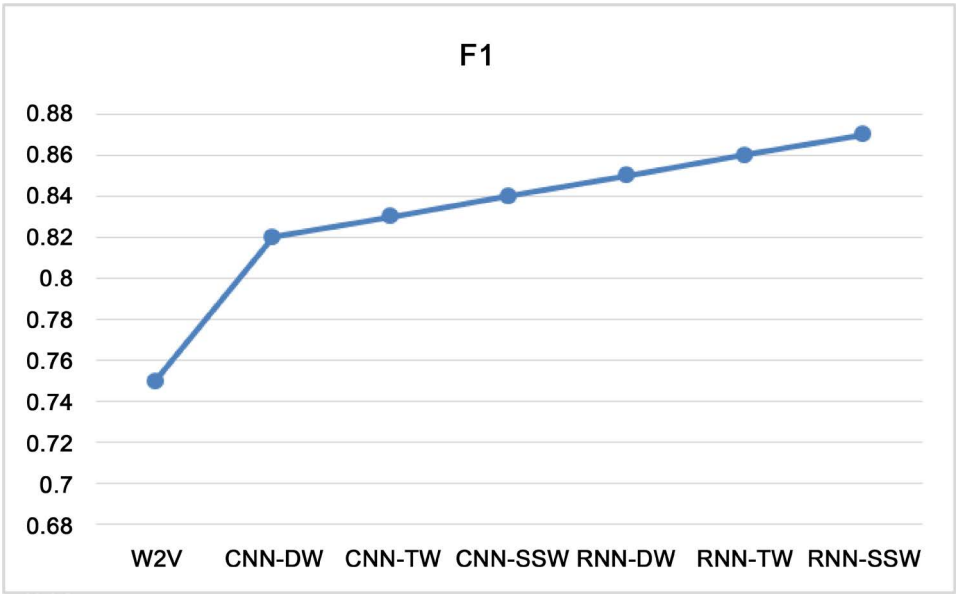

Figure 16. Depth learning effect comparison of f1 values of text emotion classification under 3 word vectors
图 16. 深度学习分别在 3 种词向量下的文本情感分类 F1 值效果对比图

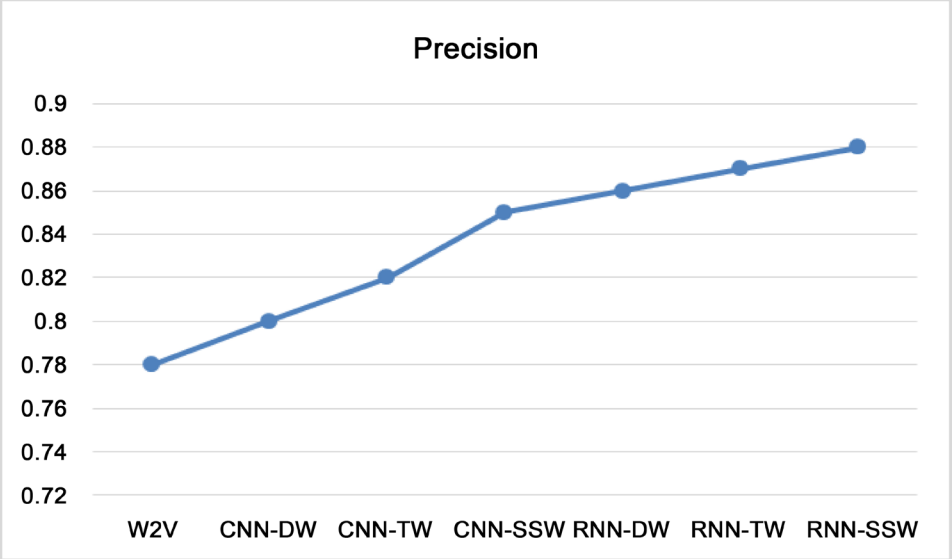

Figure 17. Depth learning effect comparison of precision values of text emotion classification under 3 word vectors
图 17. 深度学习分别在 3 种词向量下的文本情感分类 Precision 值效果对比图

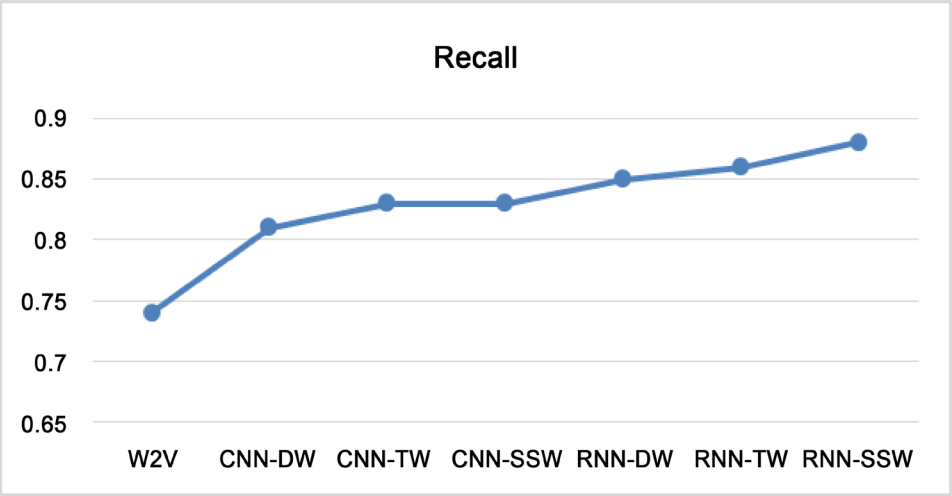

Figure 18. Depth learning effect comparison of recall values of text emotion classification under 3 word vectors
图 18. 深度学习分别在 3 种词向量下的文本情感分类 Recall 值效果对比图

2) 在本章提出的基于深度学习的几种词向量上面：结合情感词向量和情感词典的 D & W 词向量、结合情感词向量和情感词频和 T & W 词向量、结合情感词频和情感词典以及情感词向量的 SSW 情感语义词向量对比中，可以看出 SSW 的效果要比其他两种词向量的效果好很多，说明将文本情感信息和语义信息结合起来对比其他的词向量模型对于文本情感分类任务可以提升分类效果。

3) 可以看出在基于深度学习的几种词向量模型的情感分类任务中，基于语义词向量学习模型的分类效果是最差的，说明仅仅将文本的词义信息作为情感特征去进行情感分类的效果是最差的，不能够有效的得到情感的上下文语序语义信息以及情感信息，因此其分类效果对比其他词向量进行的特征选择效果相比很差。

4) 在深度学习实验中，对比基于情感词典和情感词向量的 D & W 词向量模型、结合情感词向量和情感词频的 T & W 词向量模型，二者相对比基本的基于词频的特征词向量 BOW 模型和单纯的词向量

CBOW 和 Skip-gram 提升了分类效果，因为二者加入了情感信息和词频信息，比单纯的词频信息和词向量信息好很多，说明需要在情感分类任务中加入多层次的词向量信息。

6. 全文总结

本文根据构建的多元文本情感数据集，结合深度学习对其进行特征抽取和多元情感分类任务，同时改进了相关的卷积神经网络结构和递归神经网络结构，提升了实验效果，提高了准确率，从而验证了结合情感词向量和情感词典的 D & W 词向量、结合情感词向量和情感词频的 T & W 词向量、结合情感词频和情感词典以及情感词向量的 SSW 情感语义词向量的有效性。

参考文献

[1] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Internal Representations by Error Propagation. MIT Press, Nature, 3l8-362.

[2] Bengio, Y., Ducharme, R., Vincent, P., *et al.* (2003) A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, **3**, 1137-1155.

[3] Collobert, R. and Weston, J. (2008) A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. *Proceedings of the Twenty-Fifth International Conference on Machine Learning* (*ICML* 2008), Helsinki, Finland, 160-167.

[4] Hinton, G.E. (1986) Learning Distributed Representations of Concepts. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, **1**, 12.

[5] Mikolov, T., Sutskever, T., *et al.* (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, **26**, 3111-3119.

[6] 梁军, 柴玉梅, 原慧斌, 等. 基于深度学习的微博情感分析[J]. 中文信息学报, 2014, 28(5): 155-161.

[7] 陈翠平. 基于深度信念网络的文本分类算法[J]. 计算机系统应用, 2015, 24(2): 121-126.

[8] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, 25-29 October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[9] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Sentiment Classification Using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNILP*), July 2002, Philadelphia, 79-86.

[10] Davidov, D., Tsur, O. and Rappoport, A. (2010) Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. *International Conference on Computational Linguistics*: *Posters*, Paris, 241-249.

[11] 李婷婷, 姬东鸿. 基于 SVM 和 CRF 多特征组合的微博情感分析[J]. 计算机应用研究, 2015, 32(4): 978-981.

[12] 李荣陆, 王建会, 陈晓云, 等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展, 2005, 42(1): 94-101.

[13] Taboada, M., Brooke, J., Tofiloski, M., *et al.* (2011) Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, **37**, 267-307.

[14] Hu, M. and Liu, B. (2004) Mining and Summarizing Customer Reviews. *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington DC, USA, August, 168-177.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org