

用于微博情感分析的一种情感语义增强的 深度学习模型

何炎祥^{1),2)} 孙松涛¹⁾ 牛菲菲¹⁾ 李 飞¹⁾

¹⁾(武汉大学计算机学院 武汉 430072)

²⁾(武汉大学软件工程国家重点实验室 武汉 430072)

摘 要 基于神经语言模型的词向量表示技术能够从大规模的未标注文本数据集中自动学习词语的有效特征表示,已经在许多自然语言处理任务及研究中取得重要进展.微博中的表情符号是微博情感分析最重要的特征之一,已有大量的研究工作在探索有效地利用表情符号来提升微博情感分类效果.借助词向量表示技术,为常用表情符号构建情感空间的特征表示矩阵 R^E ;基于向量的语义合成计算原理,通过矩阵 R^E 与词向量的乘积运算完成词义到情感空间的映射;接着输入到一个 MCNN(Multi-channel Convolution Neural Network)模型,学习一个微博的情感分类器.整个模型称为 EMCNN(Emotion-semantics enhanced MCNN),将基于表情符号的情感空间映射与深度学习模型 MCNN 结合,有效增强了 MCNN 捕捉情感语义的能力.EMCNN 模型在 NLPCC 微博情感评测数据集上的多个情感分类实验中取得最佳分类性能,并在所有性能指标上超过目前已知文献中的最好分类效果.在取得以上分类性能提升的同时,EMCNN 相对 MCNN 的训练耗时在主客观分类时减少了 36.15%,在情感 7 分类时减少了 33.82%.

关键词 微博;情感分析;深度学习;卷积神经网络;自然语言处理;社交网络

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2017.00773

A Deep Learning Model Enhanced with Emotion Semantics for Microblog Sentiment Analysis

HE Yan-Xiang^{1),2)} SUN Song-Tao¹⁾ NIU Fei-Fei¹⁾ LI Fei¹⁾

¹⁾(Computer School, Wuhan University, Wuhan 430072)

²⁾(State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072)

Abstract Word embedding based on neural language model can automatically learn effective word representation from massive unlabeled text dataset, and has made essential progress in many natural language processing tasks. Emoticons in microblog are important emotion signals for microblog sentiment analysis. There have been a lot of research works exploiting emoticons to improve sentiment classification performance for microblog effectively. Commonly used emoticons are adopted to construct an emotion space as feature representation matrix R^E from their word embedding. On the basis of vector based semantic composition, the projection to emotion space is performed as matrix-vector multiplication between R^E and other embedding. Then, the results are forward to MCNN to learn a sentiment classifier for microblog. This new model is named as EMCNN, short for Emotion-semantic enhanced MCNN, which seamlessly integrates emotion

收稿日期:2016-04-23;在线出版日期:2016-09-27. 本课题得到国家自然科学基金(61303115,61373039,61472290)、高等学校博士学科点专项科研基金(2013014111002512)资助. 何炎祥,男,1952年生,教授,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为编译系统、可信软件、软件工程和自然语言处理. E-mail: yxhe@whu.edu.cn. 孙松涛,男,1986年生,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为自然语言处理、情感分析和机器学习. 牛菲菲,女,1992年生,硕士研究生,主要研究方向为自然语言处理和情感分析. 李 飞,男,1980年生,博士,副教授,主要研究方向为分布并行处理、计算机网络和社交网络.

space projection based on emoticon into deep learning model MCNN to enhance its ability of capturing emotion semantic. On the datasets of NLPCC microblog sentiment analysis task, EMCNN achieves the best performance in several sentiment classification experiments and surpass the state-of-the-art results on all the performance metrics. Comparing to MCNN, EMCNN not only improve the classification performance, but also reduce the training time, i. e. 36.15% for subject classification and 33.82% for 7-class sentiment classification.

Keywords microblog; sentiment analysis; deep learning; convolution neural network; natural language processing; social networks

1 引言

在以微博和微信等为代表的新兴社交媒体上,人们可以自由地以文本、图片和视频等形式发布个人内容,进行信息交流和意见表达.对这些海量用户生成的数据进行情感分析,将有助于达到多方面的目的,例如获取大众对于有关政策的意见^[1]、相关产品的满意程度^[2]和各种社会事件的看法^[3]等等.因而,针对社交媒体的情感分析已成为国内外相关领域的研究热点,特别是微博内容的特殊性为微博情感分析带来新的机遇和挑战.

情感分析,又称意见挖掘,是分析人们对于产品、服务、组织、个人、事件、主题及其属性等实体对象所怀有的意见、情感、评价、看法和态度等主观感受的研究领域^[4].任何可以传达人类情感的事物都是情感分析的研究对象,例如文本、图片、声音和视频等.微博情感分析一般通过分析微博的文本内容来判断其情感倾向,因而大多采用文本分析技术.然而,现实中微博的内容除了文本之外,还具有其他多种多样的表现形式,例如包含大量的主题标签、表情符号、URL、图片和视频等,同样有助于分析微博的情感倾向^[5-6].

值得注意的是,语言学家发现表情符号正逐渐成为网络文本语言的重要特征之一^[7],不仅增强了主观文本的情感表达,甚至为“客观”文本增添了主观情感^[8].近来,许多微博平台甚至手机输入法,都为用户提供了各种图画形式的表情符号,从而进一步提高了其使用频率.大量观察表明,表情符号所处的上下文中,常常伴随与其情感倾向一致的其它表情符号和文本语义单元,因而可以使用词向量表示技术学习表情符号的语义表示向量. Skip-gram^[9-10]是词向量表示技术的一种,每个语义单元 w_i 对应一个 d 维的实数向量 e_i ,通过训练一个神经网络,使其

具有预测输入 e_i 所处上下文的能力,即给定窗口大小 n 的情况下语义单元 $w_{i-n}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+n}$ 对应的向量 $e_{i-n}, \dots, e_{i-1}, e_{i+1}, \dots, e_{i+n}$. 如果将文本语义单元和表情符号一起作为语义单元集合,其中的每个元素都将以向量的形式映射到同一个 d 维的语义空间中.将这些向量统称为词向量,它们能够很好地捕捉语义单元之间潜在的语义关系^[11].不加区分地使用这些词向量作为特征进行情感分析,许多研究都表明可以提高情感分类性能.

然而,人们在发布微博内容时,使用表情符号的方法与一般文本语义单元之间还是存在很大不同.自然语言中的文本语义单元都具有语法功能,构建文本内容时必须遵守相应的语法规则.情感词指的是文本语义单元中包含情感信号的词语,同样需要按照语法规则使用,并且还具有一些特殊的情感表达语法结构,例如情感的否定、增强和转移等语法现象.但是,表情符号不具有语法功能,在微博中通常以独立的形式来表达情感信号.例如,微博“很绿很环保,很好很强大[good]”中,“好”和“强大”都是情感词,副词“很”增强了它们的情感,如果移除它们的话整个句子就不再完整;“[good]”是表情符号,与两个情感词的情感一致,但移除后不影响句子的完整性.这将会导致两者的上下文环境出现差异,进而影响其词向量,因为词向量表示技术同时具有捕捉语法和语义的能力^[11].还有一点是,情感词相比情感符号通常含有更加复杂的情感信号^[12].这些不同导致了表情符号的词向量相比情感词的词向量具有更加优秀的情感表达能力,由其构建的情感空间也更加优良.因此,文献[12]提出 ESM (Emoticon Space Model),利用表情符号的词向量构建表情符号空间,并利用此空间完成所有词向量的重映射,之后使用 SVM 进行情感分类,在 NLPCC 微博情感评测任务的数据集上取得了已知最好的情感分类结果.

Skip-gram 的训练是一个非监督学习过程,通

常需要海量的文本语料数据,得到的词向量具有普遍性,适用于各种 NLP 任务.这个过程又通常称为词向量的预学习.如果一个 NLP 任务是有监督分类问题,采用合适的深度学习模型,就可以在监督训练过程中进一步更新这些词向量,使其语义更加明确,从而提高分类性能^[13].此外,适用于 NLP 的深度学习模型通常包含语义合成函数,可以完成多个词向量间的语义合成计算,因而比传统分类器更能发掘词向量的语义特性,进而提高分类效果.

本文利用深度学习模型进行中文微博的情感分类研究.结合基于向量表示的语义合成计算原理,通过表情符号空间矩阵与词向量的矩阵乘法将普通词向量映射到情感空间中,使得每个词向量的情感语义更加明确.采用多通道卷积神经网络(Multi-channel Convolutional Neural Network, MCNN)模型^[14]进行有监督学习,不仅能够根据情感标注进一步更新词向量,而且可以在不同大小的窗口内进行语义合成操作,从而识别更多的情感分类模式.因此,本文模型可以称为一种情感语义增强的深度学习模型,具体为情感语义增强的多通道卷积神经网络(Emotion-semantics enhanced Multi-channel Convolutional Neural Network, EMCNN).本文的主要贡献有以下几点:

(1)探索了表情符号在微博情感分析中的作用.通过人工标注常用表情符号的情感倾向,发现情感符号的使用频率越高,具有情感指示作用的可能性就越高,其情感明确的可能性也会越高;利用词向量技术和有标注数据集的情感标签,对比研究表情符号和情感词的情感表达能力,发现前者相比后者能够更好地区分多种情感语义.

(2)引入 MCNN 模型进行中文微博的多情感分类,利用表情符号的词向量增强 MCNN 的情感语义,并提出向量范数保持算法来简化模型的学习过程.采用 2014 年 NLPCC 微博情感评测的公开数据集,在 Subj F1、Micro F1 和 Macro F1 这 3 个指标上相对当前已知最好的方法 ESM 分别提高 2.23%、2.98%和 1.54%的同时,主客观分类训练耗时减少 36.15%,情感 7 分类训练耗时减少 33.82%.

(3)在情感评测任务的实验中,本文模型超过 ESM 模型取得最好的多情感分类性能,对于含表情符号和不含表情符号微博的情感识别都有提升;通过组织更多的实验,发现本文模型能够更好地同时利用非标注数据集和标注数据集,并且充分发挥情感符号的情感指示作用.

本文第 2 节介绍微博情感分析和深度学习等相关工作;第 3 节重点描述本文提出的 EMCNN 模型;第 4 节介绍实验数据集和基本流程,详细描述实验结果和分析;最后对本文研究工作进行总结和展望.

2 相关工作

微博情感分析通常采用文本情感分析^[15]的相关技术,同样具有情感信息抽取、情感信息分类(简称情感分类)和情感信息的检索与归纳等多种研究任务,本文主要研究微博文本的情感分类任务.

根据分类目标的不同,情感分类又可以分为主客观分类、正负情感极性分类和多情感分类等.主客观分类是情感分类的基本任务,主要负责从大规模原始数据集中识别包含主观信息的文本,构建主观文本数据集进行下一步的情感分析研究.一般而言,研究人员将主观文本的极性分为正向和负向两类,正向表示文本的情感语义为褒义,负向表示文本的情感语义为贬义.这种分类方法尽管简单,却可以满足很多现实应用的需求,例如判断消费者对于商品和服务是好评还是差评,人们对于一些观点和政策是支持还是反对.然而,人类的情感是复杂的,不是简单的正负极性可以描述的,至今对情感的细粒度划分还没有一个公认标准.已有研究将情感划分为 4、8 乃至 20 多个基本类别,不同情感之间又具有相关性和合成性,甚至可以按照一定层次结构来描述.因而,仅从情感本身的特点来看,多情感分类是 3 个分类任务中最困难的一个任务.

用于情感分类的方法,大体上可以分为两类:基于情感知识的启发式判别方法和基于特征分类的机器学习方法.情感知识主要来自于情感文本中有价值的情感信息单元,例如情感词、评价对象和观点持有者等,其中情感词的指示能力最强.情感词指的是那些带有情感倾向性的词语,是文本含有情感倾向的重要信号之一,因此构建情感词典是文本情感分析的一个基础且重要的工作^[16].情感词典可以人工构建,也可以利用情感语料库自动构建.由于微博成为人们日常情感表达的主流平台,微博文本也成为很好的情感语料库,表情符号在其中发挥着重要作用.文献[17]提出的情感词典构建框架,能够同时利用情感种子词和情感符号来更加准确地捕捉候选情感词的细粒度情感倾向分布,在公开情感评测数据集上取得了最好性能.

近年来,基于特征分类的机器学习方法被越来越

越多地应用到情感分类任务中. 这种方法通常需要首先选取大量有意义的特征作为输入, 然后采用各种分类器模型完成分类. 文献[18]最早应用机器学习进行情感分类, 尝试了 n -gram 和词性两种文本特征, 并且对比了 NB、ME 和 SVM 这 3 种分类器的情感分类效果. 文献[19]还引入了位置特征和评价词特征来完成正负情感极性分类. 对于微博的情感分类, 文献[20]引入了微博文本特有的一些特征, 例如转发、hashtag、回复、URL、标点(感叹号和问号)和表情符号等. 同大多数的情感分类研究一样, 这些文献都采用了有监督学习的分类器, 需要提供一定规模大小的有标注的数据集作为训练集. 通过利用情感词、表情符号和用户打分等情感信号, 不需要人工标注数据集就可以构建无监督学习的分类器^[21-22]. 研究人员发现, 含有“:)”的微博有很大可能表达正向情感, 含有“:(”的微博则倾向于负向情感. 基于这个观察, 文献[23]最早提出使用表情符号“:)”和“:(”分别作为微博的正负向情感极性指示, 收集大量的微博文本作为训练集, 进行远距离有监督学习(distantly supervised method), 可以认为介于有监督学习和无监督学习之间. 此后, 文献[24]通过人工选取具有正负情感倾向的两组表情符号, 提高微博收集效率. 这种使用表情符号作为微博情感标签的方法, 构建的情感语料库充满了噪声样本, 被称为有噪标注数据集, 对分类器的性能有一定的影响. 因而文献[25]提出了 ESLAM(Emoticon Smoothed LAnguage Model), 首先在有标注的数据集上训练一个语言模型, 然后利用表情符号的有噪标注数据集进行平滑处理, 能够很好地同时结合两个数据集中的文本特征, 最终提高情感分类效果.

深度学习是机器学习的一种范式, 近年来引起工业界和学术界的广泛关注^[26]. 深度学习模型通常是具有多层结构的神经网络, 强调从大规模的数据集中学习各种现实事物的特征表示, 并且这些特征能够被计算机直接应用于各种计算模型中. 同样使用表情符号构建有噪标注数据集, 文献[27]采用 DBN(Deep Belief Network)算法学习整条微博的特征表示, 结合其它文本特征输入到 SVM 进行情感四分类. DBN 和词向量技术都是深度学习的一种, 都可以对文本进行无监督的特征表示学习. 不同于文献[27]将整条微博作为特征表示学习的基本单元, 词向量技术将词语作为基本单元, 学习到的特征表示是多维的稠密向量, 因而将其简称为词向量. 已有的研究提出了多种词向量学习模型, 被广泛认可

和使用的是 Skip-gram 模型. 词向量可以有效地捕捉词语的语法和语义信息, 被研究者应用到 NLP 领域的各种任务中, 取得了许多令人瞩目的成果.

情感信息来源于文本的语义表达, 因而词向量中也含有一定的情感语义. 文献[28]对比传统的文本特征 TF-IDF 和词向量特征对中文商品评论的情感分类性能, 结果表明了后者的优越性. 然而, 从无监督数据集中学习到的词向量, 其情感语义并不明确, 例如情感极性相反的“好”和“坏”, 在词向量空间中距离很近. 因此, 文献[29]提出词向量的再学习, 利用有监督数据集进一步学习词向量, 从而使得词向量具有任务相关的语义, 应用到情感分析任务中就可以增强词向量的情感语义; 文献[30-31]提出多个词向量学习算法, 在有噪标注数据集上能够有效地学习情感语义更加明确的词向量.

使用词向量进行情感分类, 以上大多数的研究都是将词向量作为特征输入传统分类器(例如 SVM); 如果结合计算语义学的知识, 利用词向量完成语义合成运算, 提高语义识别能力, 就可以进一步提高分类性能. RNTN(Recursive Neural Tensor Network)模型^[32]利用情感树库在二元化的句法树结构上进行语义合成, 文献[33]在此模型基础上, 通过对依存句法树进行变换, 完成句中指定评价对象的情感极性分析. 文献[34]通过添加语义合成函数的 pooling 层, 使得 RNTN 可以自适应地选择最适合当前节点的语义合成函数, 从而大幅提高情感分类效果. DCNN(Dynamic Convolutional Neural Network)模型^[35]通过学习句子结构来提高情感语义合成性能, CharSCNN(Character to Sentence Convolutional Neural Network)模型^[36]为英语单词增加了字符表示层的特征, 可以提高对于英文 Twitter 中不规则单词的表示能力. 由于微博文本的不规范性, 各种句法分析结果通常不准确, 因而 CNN(Convolutional Neural Network)模型成为微博情感分析的首选模型.

在参照国内外多种分类体系的基础上, 通过综合已有的情感词汇资源, 文献[37]构建了一个中文情感词汇本体库, 将情感分为 7 大类和 20 小类的层次结构, 并且包含了情感强度. 本文的实验数据集采用文献[37]的 7 大类情感类别, 分别是乐(happiness)、好(like)、哀(sadness)、恶(disgust)、怒(anger)、惧(fear)和惊(surprise). 文献[7]使用词向量技术, 研究表情符号与多种情感类别之间的映射关系, 发现词向量空间可以很好地体现表情符号在情感语义上

相关性.受文献[7]和文献[17]的启发,本文将考察选定情感词和表情符号在词向量空间中与多情感类别的相关性以及它们在训练集中与多情感标签概率分布间的关系(详见 3.1 节和 3.2 节).文献[14]采用的 MCNN 是 CNN 的一种,能够对多个窗口大小内的词向量执行语义合成操作,在多个不同的情感数据集上都取得了很好地分类性能,是本文提出的 EMCNN 的基础模型.相比于文献[12],本文进一步探索了表情符号空间的情感表达能力,通过与 MCNN 结合构建的 EMCNN 将情感语义的重映射与情感分类两个过程统一到一起,体现出具有语义合成能力的深度学习模型相对传统分类器在情感分类任务中的优势.

3 主要模型描述

3.1 表情符号的情感标注

人们越来越多地在微博中使用表情符号来表达情感,微博平台也提供了越来越多的表情符号来满足用户表达各种不同情感的需要.以新浪微博为例,2015 年 12 月 08 日的统计显示,平台为用户提供了 1008 个不同的表情符号.如果计入经常被使用的网络原创表情,这个数目将达到 1500 个以上.正如文献[12]的观察,这些表情符号的情感表达能力各不相同.采用 NLPCC2014 的训练集(详见 4.1 节),收集得到 420 个表情符号,并按照使用次数进行降序排列.两名研究人员对表情符号列表的前 300 个进行标注,一个标注任务是判断表情符号是否表达情感,另一个是进一步判断表情符号表达的情感是否明确.按照表情符号的排列顺序,分别统计包含情感和情感明确表情符号所占百分比,得到如图 1 所示结果,其中 Top- n 表示列表的前 n 个表情符号.两个任务的标注一致性 Kappa 值分别为 0.7065 和

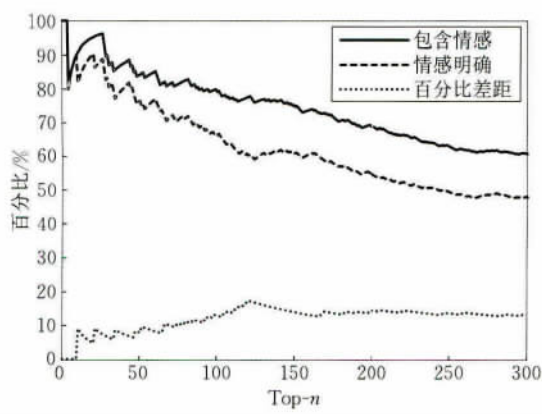


图 1 表情符号的情感表达能力

0.5818,表明后者困难度大于前者.

从图 1 中可以明显看出,随着 Top- n 的增大,包含情感或者情感明确的表情符号所占比例不断下降;由于后者是基于前者的标注,因而后者所占比例始终较低,开始时两者之间的差距也随着 Top- n 的增大而不断加大,而在大约 125~170 之间两者差距快速缩小,之后开始保持基本不变.这些现象表明,使用频率越高的情感符号,含有情感或者情感明确的可能性就越大;使用频率较低的情感符号,这两种可能性仍然存在,并且开始保持稳定,因而也不能完全忽视.人工标注过程中的经验表明,判断情感符号的多种情感倾向是非常困难的,同时为了与文献[12]保持一致,本文也就同样采用 Top- n 的方式选择情感符号用于 EMCNN 模型.

3.2 表情符号与情感词对比

根据对 NLPCC 数据集的统计发现(详见 4.1 节),7 种情感的使用频率差别很大,其中乐、好、哀和恶这 4 种情感最常用,且乐和好为积极情感,哀和恶是消极情感.参考文献[17]以及前一节对表情符号的人工标注,为这 4 种情感分别选取了 5 个最有代表性的表情符号和情感词,如表 1 所示.

表 1 选取的表情符号和情感词

	表情符号	情感词
乐	[哈哈],[嘻嘻],[太开心],[大笑],[开心]	高兴,喜悦,欢喜,放心,踏实
好	[鼓掌],[爱你],[花心],[good],[赞]	尊敬,赞扬,相信,喜欢,祝愿
哀	[泪],[委屈],[伤心],[大哭],[失望]	忧伤,伤心,绝望,内疚,悲伤
恶	[鄙视],[吐],[闭嘴],[困],[t 中枪]	讨厌,麻烦,恶心,嫉妒,可耻

参考文献[7]使用词向量分析表情符号与情感之间映射关系的方法,本文分别对表 1 中情感符号列表和情感词列表对应的词向量空间进行 PCA,得到它们在二维空间中的映射关系分别如图 2 和图 3 所示.由于通常使用词向量间的 cosine 值作为相似度的衡量^[11],这里将 PCA 后的词向量归一化到圆周上,根据坐标点之间的距离就可以判断语义相似度.可以看出,“乐”中的情感符号和情感词在两图中都比较分散,而其它 3 种情感的情感符号和情感词则比较集中,即至少有 3 个同类别的聚在一起.从积极和消极情感的角度来看,图 2 中有一个“恶”的表情符号被映射到积极情感区域,其他情感符号都被正确映射到相应区域;然而图 3 中被错误映射的情感词很多,“乐”、“好”和“恶”各有一个情感词被映射

到同一个邻域中. 这表明词向量空间具有一定的情感语义区分能力, 且表情符号的区分度更高.

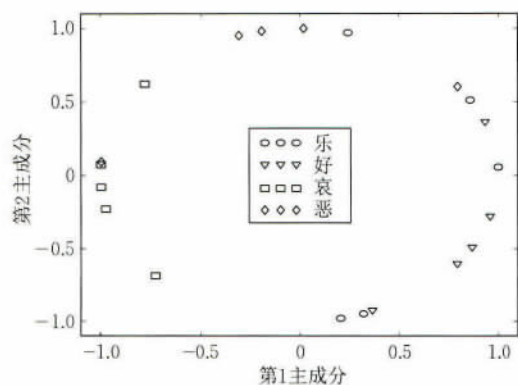


图 2 表情符号的 PCA

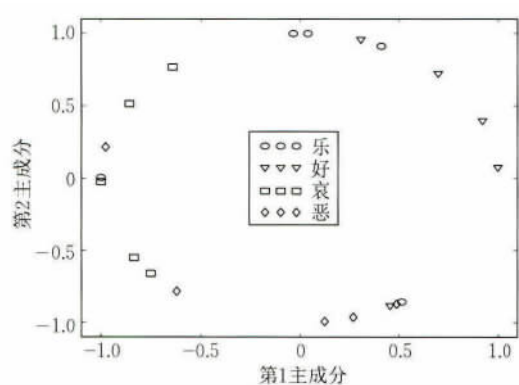


图 3 情感词的 PCA

更进一步, 可以利用 NLPCC 的训练集考察这些表情符号和情感词与微博情感标注之间的关系, 通过下面的贝叶斯公式计算:

$$P(c_j | w_i) = \frac{P(c_j)P(w_i | c_j)}{\sum_{j=0}^7 P(c_j)P(w_i | c_j)} \quad (1)$$

其中: w_i 代表一个表情符号或者情感词; c_j 代表第 j 个情感标签 (7 种情感和无任何情感共 8 个标签). 在 NLPCC2014 的训练集上统计, 得到的结果如图 4

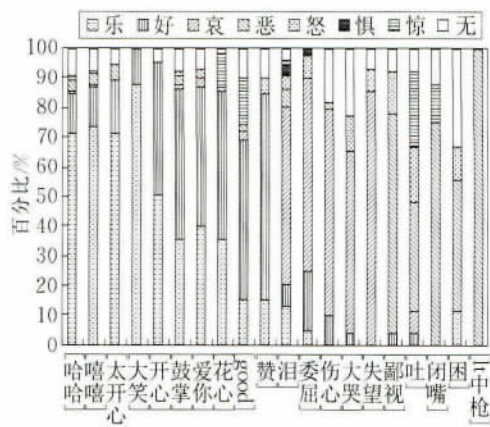


图 4 表情符号的情感标注分布

和图 5 所示. 可以明显看出, 大部分情感词出现在“无”情感微博中的比例很高, 说明其情感指示作用不强; 而表情符号大多出现在有情感的微博中, 并且表现出很强的多情感区分能力.

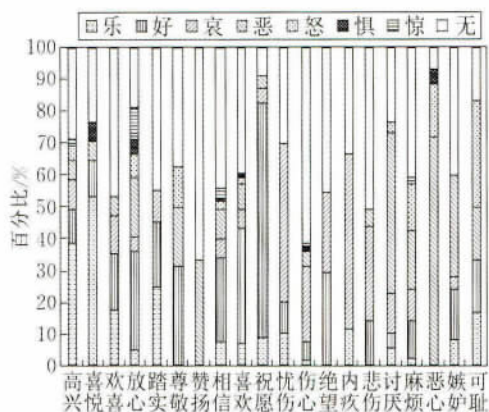


图 5 情感词的情感标注分布

3.3 EMCNN 模型

前两节的研究表明, 绝大部分的常用表情符号都包含情感, 并且大多数的情感语义明确; 与情感词的对比发现, 前者不仅具有更强的情感指示作用, 而且还可以很好地区分多种不同情感. 因而, 本文采用表情符号的词向量来增强 MCNN 模型的情感语义.

本文提出的多情感分类 EMCNN 模型整体架构如图 6 所示, 是一个自底向上的多层前向神经网络

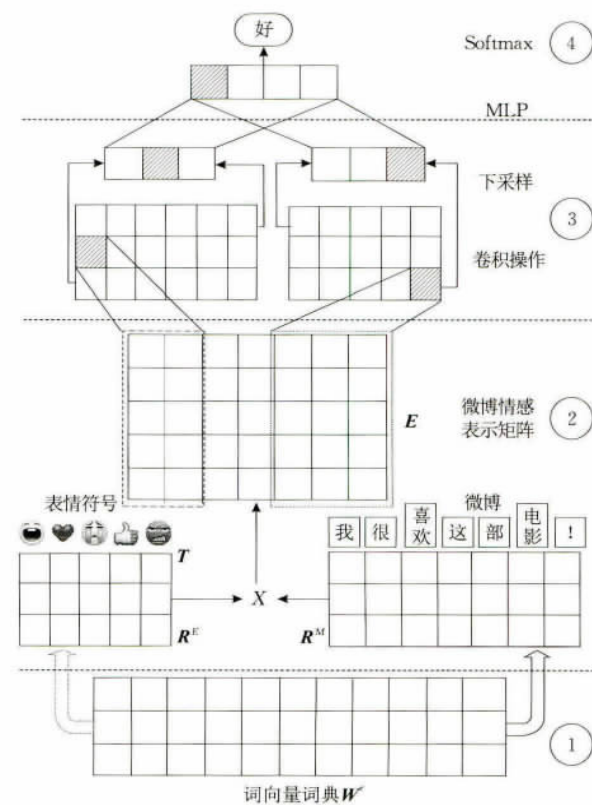


图 6 EMCNN 模型整体架构

络结构. 根据功能的不同, 本文将 EMCNN 模型划分为 4 个层次, 自底向上依次为词向量输入层、情感语义增强层、卷积采样层和情感得分输出层.

3.3.1 词向量输入层

最底部的词向量词典是 EMCNN 模型的输入层, 由整个数据集中的词语和表情符号(统称为词典元素)组成, 每一列为一个词向量表示. 假设词典中元素个数为 N , 单个词向量的维度为 d , 则 $W^e \in \mathbb{R}^{d \times N}$ 代表整个词向量词典的表示矩阵. 对于一个由词典元素构成的文本序列 $S = (c_1, c_2, \dots, c_k)$, 第 j 个词典元素的词向量 r_j 通过下面操作获得:

$$r_j = W^e v_j, 1 \leq j \leq k \quad (2)$$

其中, $v_j \in \mathbb{R}^N$ 为一个词典元素 c_j 对应的维度取值为 1, 其它维度全为 0 的 N 维指示向量, 程序实现中常常采用一个整型值来进行稀疏表示即可. 这个过程可以看做是查字典(Lookup), 词向量表示矩阵 W^e 是模型一个重要的参数, 在训练过程中将不断被学习.

通过将文本序列中对应位置词典元素的词向量依次拼接起来, 就得到整个文本序列的词向量表示矩阵:

$$R^S = r_1 \oplus r_2 \oplus \dots \oplus r_k \quad (3)$$

其中: \oplus 表示行向量方向的拼接操作, 图 6 中的 R^E 和 R^M 分别是选定的表情符号序列和微博文本序列通过上面的操作得到的. 这里, R^E 即为情感空间的矩阵表示, 通过选择不同数量的表情符号, 可以改变此矩阵的维度; R^M 即为微博的词向量矩阵表示, 是对微博文本序列的另外一种表示形式, 方便后面使用矩阵运算进行各种语义合成操作.

3.3.2 情感语义增强层

为了得到整条微博在情感空间中的特征表示 E , 基于向量的语义合成计算原理^[38], 执行下面的矩阵相乘运算:

$$E = (R^E)^T \times R^M \quad (4)$$

其中, $(R^E)^T$ 表示 R^E 的转置. 如果选取了 d_e 个表情符号, 那么就有 $R^E \in \mathbb{R}^{d_e \times d_e}$. 由于数据集中微博的长度各不相同, 这里就选取最大微博长度 d_m , 那么就有 $R^M \in \mathbb{R}^{d_e \times d_m}$; 对于长度小于 d_m 的微博, 在其末端添加全 0 的列向量直到长度达到 d_m . 这样, 每条微博在情感空间的特征表示 $E \in \mathbb{R}^{d_e \times d_m}$.

文献[12]中提出的 ESM 模型同样使用表情符号的词向量构建情感空间, 映射过程则采用词向量间的 cosine 值来完成, 计算公式如下:

$$\cos(p_i, q_j) = \frac{p_i \cdot q_j}{\|p_i\| \cdot \|q_j\|} \quad (5)$$

其中: p_i 代表词向量词典中的所有词向量; q_j 则代表

选取的表情符号列表中的第 j 个表情符号对应的词向量. 对比式(4)发现, 两者相差式(5)中的分母部分, 即两个词向量的 L2 范数. 为了与文献[12]中的情感映射过程保持一致, 并且简化 EMCNN 模型的计算量, 本文提出向量范数保持算法, 确保在使用反向传播算法进行模型训练过程中, 词向量的 L2 范数始终保持不变.

算法 1. 向量范数保持算法.

输入: 待更新词向量 e_i , 对应的更新量 Δ_i

输出: 更新后的词向量 e_i^*

1. $e'_i = e_i - \Delta_i$
2. $\Delta'_i = e_i - e'_i \times \|e_i\| / \|e'_i\|$
3. $e_i^* = e_i - \Delta'_i$

3.3.3 卷积采样层

随后, 对微博的特征表示 E 执行卷积操作, 可以在指定窗口内完成微博文本的语义合成. 设卷积算子为 $W^f \in \mathbb{R}^{d_e \times d_{win}}$, 其中 d_{win} 表示卷积窗口的大小. 那么, 卷积操作可以表示为

$$\begin{cases} r_{ij}^f = W_i^f \otimes z_j + b_i^f \\ z_j = E_{j-d_{win}+1} \oplus \dots \oplus E_j \end{cases} \quad (6)$$

其中: r_{ij}^f 表示第 i 个卷积算子 W_i^f (又称作过滤器) 在第 j 个词语处的卷积结果; z_j 表示在窗口 d_{win} 范围内特征表示 E 的分块矩阵; b_i^f 表示第 i 个卷积算子的偏置值; \otimes 表示矩阵的内积操作; E_j 表示 E 的第 j 个列向量; \oplus 表示行向量方向的拼接操作. 根据对特征表示 E 左右两端卷积操作方式的不同, 卷积窗口有两种滑动类型, 即窄型(narrow, 图 7(左))和宽型(wide, 图 7(右)), 对应卷积结果的维度分别为 $d_m - d_{win} + 1$ 和 $d_m + d_{win} + 1$, 本文所有模型的卷积操作都采用后者. 由于本文采用了多通道的网络结构, 因而将有多组卷积操作的参数, 每组都包含 W^f 、 b^f 和 d_{win} .

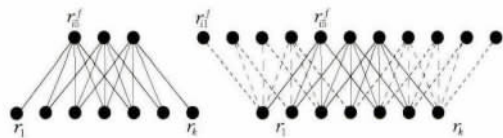


图 7 两种卷积类型

接着, 对卷积后的特征矩阵 r^f 执行最大化下采样操作, 获取每个维度上的最大值, 即获取每个卷积算子对整条微博操作后的最强激励信号. 运算结果如下:

$$x_i = \max_j [r_{ij}^f]_{i,j}, 0 < j < k + d_{win} - 1, 0 < i < H \quad (7)$$

其中, H 表示卷积算子的个数, 是 CNN 的重要超参数之一, 需要人工调节选取.

如图 6 中为双通道的网络结构, d_{win} 分别为 2 和 3, 它们的卷积和下采样结果将拼接为一个向量 x 输入到下一层. 本文实验采用了三通道的网络结构, 具体配置详见 4.3.3 节.

3.3.4 情感得分输出层

然后, 将前一层得到的向量 x 输入到一个多层感知器 (MLP), 得到更高层的特征表示. 本文的模型选择不含任何隐含层的 MLP, 对其输出向量进行非线性函数 g 变换后, 得到情感标签的得分向量:

$$Score(S) = g(W^h x + b^h) \quad (8)$$

其中: $Score(S) \in \mathbb{R}^{|L|}$ 为情感标签的得分向量; L 表示情感标签集合; $W^h \in \mathbb{R}^{|L| \times H}$ 和 $b^h \in \mathbb{R}^H$ 分别为 MLP 的参数矩阵和偏置量, 本文实验采用 Rectifier 函数完成非线性变换.

为了得到模型在参数 $\theta = \{W^e, W^f, W^h, b^f, b^h\}$ 条件下一条微博 S 的多情感分类概率, 需要对情感得分向量执行 Softmax 运算:

$$p(l|S, \theta) = \frac{e^{Score(S)_l}}{\sum_{i \in L} e^{Score(S)_i}} \quad (9)$$

在训练集 C 中, 通过使得式 (9) 的对数似然函数的负值最小化, 可以学习得到最优化的模型参数 θ :

$$\theta \mapsto \frac{1}{|C|} \sum_{(S, y) \in C} -\log p(y|S, \theta) \quad (10)$$

其中 (S, y) 表示训练集中的一条微博及其对应的情感标签. 同文献 [14] 一样, 模型的训练使用 SGD 优化技术, 采用 AdaDelta 算法 [39] 动态更新每个参数的学习率, 对 MLP 的参数 W^h 执行 dropout 操作 [40] 防止模型过拟合, 并根据各个训练集的规模选择合适的 mini-batch 大小.

3.4 EMCNN 模型的复杂度

由于 EMCNN 模型具有多层前向神经网络结构, 那么就可以按照其层次划分来进行时间复杂度的分析; 人工神经网络模型代码的实现通常采用矩阵运算, 因而采用矩阵基本运算的时间复杂度作为基础. 这样, EMCNN 模型与 MCNN 模型的前向计算时的时间复杂度可以表示为

$$\begin{cases} T_{EMCNN} = T_1 + T_2 + T_3 + T_4 \\ T_{MCNN} = T'_1 + T'_3 + T_4 \\ T_1 = O(d \times N \times (d_e + d_m)) \\ T'_1 = O(d \times N \times d_m) \\ T_2 = O(d_e \times d \times d_m) \\ T_3 = O(d_e \times d_{win} \times (d_m + d_{win}) \times H) \\ T'_3 = O(d \times d_{win} \times (d_m + d_{win}) \times H) \\ T_4 = O(H \times |L|) \end{cases} \quad (11)$$

其中: T_1 和 T'_1 是词向量输入层查字典操作; T_2 是 EMCNN 模型独有的情感语义增强操作; T_3 和 T'_3 是单通道的卷积采样操作; T_4 是情感得分和 Softmax 操作. 对于 T_1 和 T'_1 , 这里按照式 (2) 和式 (3) 来计算, 而实现时常常可以将式 (2) 简化为 $O(d)$ 的数据查找操作, 那么就有

$$\begin{cases} T_1 = O(d \times d_m) \\ T'_1 = O(d \times (d_e + d_m)) \end{cases} \quad (12)$$

同时, 考虑到 d, d_e, d_m 和 H 通常属于 100 这个数量级, 而 d_{win} 则通常为 10 以内, 因此 EMCNN 和 MCNN 的时间复杂度就相差在 T_2, T_3 和 T'_3 上, 而 T_3 和 T'_3 的差异主要由词向量维度 d 和选取的情感符号数目 d_e 的不同所引起的. 详细的实验耗时比较参见 4.4.6 节. 可以看出, 由于采用了向量范数保持算法, 仅使用式 (4) 就可以完成情感空间转换, 即仅仅需要计算式 (5) 的分子部分, 避免计算 $d_e + d_m$ 个向量的 L2 范数, 从而提高了 EMCNN 模型前向计算速度.

对于两个模型的空间复杂度, 就主要考察模型参数 θ 的矩阵大小, 那么就有

$$\begin{cases} M_{EMCNN} = M_1 + M_2 + M_3 \\ M_{MCNN} = M_1 + M'_2 + M_3 \\ M_1 = O(d \times N) \\ M_2 = O(d_e \times d_{win} \times H) \\ M'_2 = O(d \times d_{win} \times H) \\ M_3 = O(H \times |L|) \end{cases} \quad (13)$$

其中: M_1 为词向量词典 W^e 的大小; M_2 和 M'_2 为单通道卷积算子的大小; M_3 为 MLP 的大小. 可以看出, 两个模型的空间复杂度就相差在 M_2 和 M'_2 之上, 同样是主要由词向量维度 d 和选取的情感符号数目 d_e 的不同引起的.

4 实验与分析

4.1 数据集

本文采用 NLPCC 在 2013 年和 2014 年情感评测任务的两个公开数据集组织实验, 每个数据集中都包含一个训练集和一个测试集, 微博内容来自新浪微博, 每条微博都有一个情感标签, 共有 8 种不同的情感标签, 即 none、happiness、like、sadness、disgust、anger、fear 和 surprise. 两个数据集的统计信息如表 2 所示. 其中, 2014 年的训练集 (14train) 中有两条微博的文本内容为空, 本文选择将其移除, 最终得到 13998 条微博.

表 2 有标注数据集 (单位:条)

	13train	13test	14train	14test
none	1828	4873	6591	3603
happiness	370	1116	1459	441
like	595	1558	2204	1042
sadness	385	744	1173	189
disgust	425	935	1392	389
anger	235	436	669	128
fear	49	102	148	46
surprise	113	236	362	162
总计	4000	10000	13998	6000

由表 2 可以看出,8 种标签的比例很不均衡,这对于机器学习算法是一种挑战.为了充分测试 EMCNN 模型的性能,除了 NLPCC 正式的情感评测任务,本文还组织了类别平衡的主客观分类、正负情感分类和情感四分类这 3 个实验,其中主客观分类将 none 作为客观标签,其它 7 种情感作为主观标签;正负情感分类将 happiness 和 like 作为正向标签,sadness、anger、disgust 和 fear 作为负向标签;情感四分类选取了 like、happiness、sadness 和 disgust 的 4 个标签,因为其它 3 个情感标签的数量比较少.最终,3 个实验的数据集划分情况如表 3 和表 4 所示.

表 3 有标注数据集的划分一 (单位:条)

	客观	主观	正向	负向
2013	6701	7299	3639	3660
2014	10194	9804	5146	4658

表 4 有标注数据集的划分二 (单位:条)

	like	happiness	sadness	disgust
2013	2153	1486	1129	1360
2014	3246	1900	1362	1781

为了使用 word2vec 预学习词向量,本文通过新浪微博的 API 接口随机获取了 2012 年到 2015 年之间的大约 3 千万条微博文本作为非监督学习的无标注数据集.经过简单的去噪和预处理之后,形成的数据集称为 full-set,将其中含有情感符号的微博提取出来形成的数据集称为 emo-set,将其中 2012 和 2013 年发布的微博提取出来形成的数据集称为 part-set,统计信息如表 5 所示,其中 E 代表情感符号,“微博总数”和“含 E 微博”的单位为“条”,“E 总

表 5 无标注数据集(M 为百万,B 为十亿)

	Dataset1	full-set	emo-set	part-set
微博总数	20M	31.2M	1.09M	26.8M
含 E 微博	5M	1.09M	1.09M	0.26M
E 总数	—	2.23M	2.23M	0.45M
E 个数	—	1154	1154	808
总词数	—	1.08B	34.2M	925M
平均长度	—	34.6	31.4	34.5
词典大小	—	347886	68662	310046

数”、“E 个数”、“总词数”和“词典大小”的单位为“个”,“平均长度”代表平均每条微博中含有的词语数,单位为“个/条”.Dataset1 为文献[12]中的无标注数据集,采集时间为 2012 年 3 月到 2012 年 12 月.除了 4.4.4 节,本文其它所有实验中使用的预学习词向量,都是通过 full-set 数据集训练得到的.

4.2 对比实验

(1) MNB. 文献[21]的实验表明,在长度较短的文本情感分析中,MNB(Multinomial Naïve Bayes)通常取得优于 SVM 的分类效果;文献[24]也发现 MNB 在英文微博的情感分类中存在优势.选取微博文本中词语的 1-gram、2-gram 和 3-gram 一起作为分类特征,并且忽略文档频率小于 2 的特征.所有模型超参数的选择参见 4.3.3 节.

(2) SVM. SVM(Support Vector Machine)在许多情感分类任务中取得最好的结果^[18,23],采用类似文献[12]中的方法提取微博文本特征,并按照二元向量的形式进行表示.

(3) EMB. 使用从非监督数据集中学习到的词向量,经过下采样之后直接输入 SVM 分类器中.本文实验采用最大、最小和平均的采样策略,将此模型记为 E-EMB.此方法被许多研究(例如文献[14]、文献[30]、文献[32]、文献[34]和文献[35]等)选为对比模型,是词向量技术与传统分类器 SVM 的常用结合方式.

(4) ESM. 文献[12]提出的表情符号空间模型,在如上 EMB 方法的基础上,通过计算微博中词语的词向量与表情符号词向量的 cosine 值,完成词向量到情感空间的转换.采用求和策略的模型记为 B-ESM,采用最大、最小和求和策略的模型记为 E-ESM.此模型在情感评测数据集上取得了目前已知的最好性能,其情感空间转换方法是本文模型创新思路的来源,因而将其选为对比模型.

(5) MCNN. 本文 EMCNN 模型的基础模型^[14],直接对预学习的词向量执行卷积操作以及后续的运算,即使用 R^M 直接替代 E 作为微博特征表示.选用此模型进行对比,可以更好地体现出使用情感符号进行词向量情感语义增强对于情感分类的提升作用.

4.3 实验流程

4.3.1 文本预处理

新浪微博的文本内容使用特殊的格式标识表情符号,例如“大笑”表情符号的文本格式为“[大笑]”,

即使用中括号包含表示表情符号语义的文本. 本文使用 NLPIR 工具对微博文本进行分词, 通过这些文本格式的表情符号作为用户词典, 使得它们在分词结果中作为基本的语言单元. 此外, 如果词性属于集合 $\{\text{url}, \text{x}, \text{m}, \text{nr}\}$, 则将对应的字符串当做基本语言单元对待. 其中, 词性 url 的对应字符串, 统一转化为语言单元 $\langle \text{URL} \rangle$; 以 @ 开头且词性为 nr 的字符串, 统一转化为语言单元 $\langle \text{UNAME} \rangle$; 词性为 m 的字符串中的每个数字, 统一转化为 D 表示; 词性为 x 的字符串, 如果以 www. 开头, 则转化为 $\langle \text{URL} \rangle$, 否则保持原样.

4.3.2 词向量的预学习

使用 word2vec 工具^①中的 Skip-gram 模型进行词向量的预学习, 采用默认的参数设置. 经过文本预处理之后, 未标注数据集中每条微博被转化成空格分隔的词语序列, 一条微博为一行输入 word2vec 学习词向量.

4.3.3 模型参数设置

采用 14train 数据集的 5 倍交叉验证来调节各个分类模型中的超参数, 选择最优值用于其他所有实验. 其中, EMCNN 选用 Top-100 的情感符号 (详细讨论见 4.4.5 节), 其它超参数与 MCNN 保持一致, 具体如表 6 所示 (d_{win} 和 H 分别有 3 个数值对应 3 个通道). 这 3 个参数先于 Top- n 进行调节, 其中 d 的取值范围为 100、200、300 和 400, d_{win} 的组合为 (3, 4, 5)、(4, 5, 6) 和 (5, 6, 7), H 在 3 个通道上保持一致, 取值范围为 50、100、150 和 200, 从而可以构成 48 种不同的参数组合, 采用网格搜索 (grid search) 方法选择使得平均准确率最优的一组. MNB 模型的超参数 alpha (平滑因子) 设置为 1.0, 所有使用 SVM 的模型统一将超参数 C (正则化常数) 设置为 1.0.

表 6 EMCNN 超参数设置

超参数	说明	取值
d	词向量维度	200
d_{win}	卷积算子宽度	3, 4, 5
H	卷积算子个数	100, 100, 100

4.4 结果与分析

4.4.1 情感评测任务结果

按照 2013 年 NLPCC 情感评测任务的评价标准^②, 本文分别对 2013 年 2014 年的数据集进行的实验, 最终的评测结果分别如表 7 和表 8 所示.

表 7 NLPCC13 的评测结果

NLPCC13	Subj F1	Micro F1	Macro F1
BEST	0.7290	0.3520	0.3130
B-ESM*	0.7820	0.4160	0.3290
E-ESM*	0.7880	0.4390	0.3500
MNB	0.7652	0.3208	0.3105
SVM	0.7732	0.3624	0.3052
E-EMB	0.7827	0.4209	0.3395
E-ESM	0.7854	0.4348	0.3490
MCNN	0.7890	0.4365	0.3462
EMCNN	0.7931	0.4422	0.3517

表 8 NLPCC14 的评测结果

NLPCC14	Subj F1	Micro F1	Macro F1
MNB	0.6534	0.3595	0.2783
SVM	0.6607	0.4053	0.3051
E-EMB	0.6878	0.4324	0.3703
E-ESM	0.6860	0.4425	0.3786
MCNN	0.6946	0.4534	0.3829
EMCNN	0.7083	0.4723	0.3940

其中, B-ESM* 和 E-ESM* 的数据摘自文献 [12]. 本文的 E-ESM 实验结果在 3 个指标上都略小于 E-ESM*, 主要是由于非监督数据集的不同引起的, 4.4.4 节将安排实验来说明. 最终, EMCNN 模型在所有指标上都取得了最好的结果; MCNN 在 Subj F1 上取得了第 2 好, 但是在 Micro F1 上弱于 E-ESM* 排在第 3 好, 在 Macro F1 上弱于 E-ESM* 和 E-ESM 排在第 4 好.

同样, 在 2014 年的数据集上, EMCNN 模型在所有指标上都取得了最好的结果, 并且明显高于第 2 好的 MCNN 模型, 相比 E-ESM 模型依次提高了 2.23%、2.98% 和 1.54%. 此外, MCNN 模型在所有指标上都取得了第 2 好的结果, 不同于在 2013 年数据集的结果. 经过进一步的实验发现 (详见 4.4.2 节), MCNN 需要较大的训练集, 而 EMCNN 需要的训练集规模相对要小一些. 由于 2013 年的训练集只有 4000 条微博, 导致 MCNN 没有得到充分的训练.

由于表情符号是微博情感表达的重要信号之一, 这里将 2014 年的测试集分为含表情符号 (2069 条) 和不含表情符号 (3931 条) 两部分, 来考察各个模型的性能, 得到如图 8、图 9 和图 10 所示结果. 从整体上看, 含表情符号部分的性能指标始终高于不含表情符号部分, 表明表情符号的情感指示能力有利于微博的情感分类, 特别是图 8 中各个模型在含表情符号部分的 Subj F1 指标基本一致. 然而, 表情

① <https://code.google.com/p/word2vec/>

② <http://tcci.ccf.org.cn/conference/2013/dldoc/ev02.pdf>

符号的情感指示能力对于情感多分类的作用明显不如主客观分类,因而需要更好的模型来发现更多的情感模式,特别是占整个测试集 2/3 的不含情感符号部分。从图 9 和图 10 可以看出,情感多分类性能的整体提升来自两个部分分类性能的同时提升。值得注意的是,E-ESM 在这两个指标上的提升主要得益于不含表情符号部分,表明其情感空间转换方法对于这部分微博更加有效。然而,MCNN 相对 E-ESM 的性能提升,Micro F1 主要来自不含表情符号部分,Macro F1 则主要来自含表情符号部分,原因是前者强调体现整体分类性能的提升,而后者则重点体现不同类别分类性能的平均提升,即表现出区分不同类别能力的提升。这个现象表明,MCNN

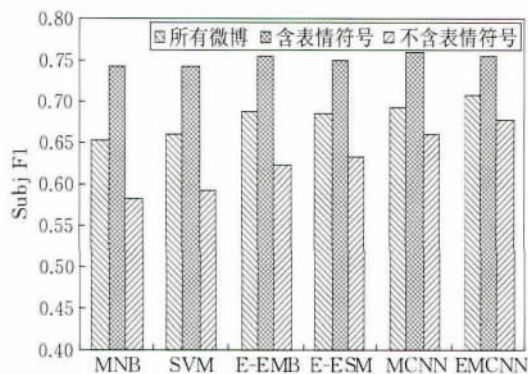


图 8 2014 年测试集的 Subj F1

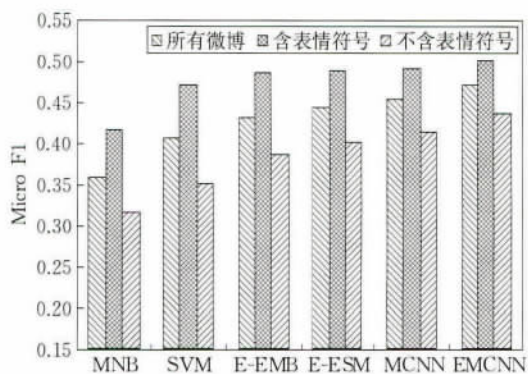


图 9 2014 测试集的 Micro F1

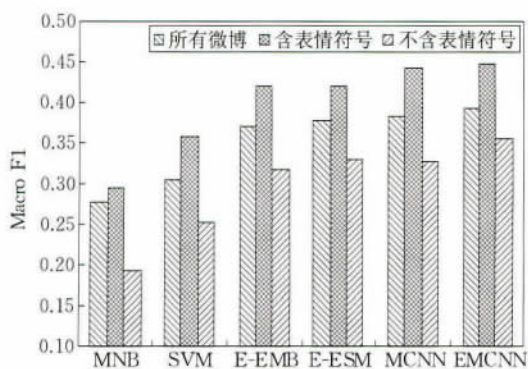


图 10 2014 年测试集的 Macro F1

相比 E-ESM 可以更好地识别不同类别间的情感模式。作为两者有效结合的 EMCNN 模型,很好地结合了两种方法的优势,在除了 Subj F1 的不含表情符号部分之外,都取得了最好性能。导致这点异常的原因将在 4.4.7 节进行详细分析。

4.4.2 类别平衡数据集的实验

为了充分测试在类别平衡数据集中各个模型的性能,本文还组织了主客观分类、正负情感分类和情感四分类的 3 个实验,数据集的具体划分见 4.1 节。这里主要测试训练集大小变化对于各个模型的影响,评测指标都选用分类准确率 (Accuracy)。假设某次实验指定的训练集大小为 X ,则每种标签随机选择 X/n (n 为标签个数) 个训练样本和 $L_{\min} - X/n$ (L_{\min} 为数量最少标签的样本数) 个测试样本。

(1) 主客观分类

此实验为二分类问题 ($n=2$),其中 2013 年的 $L_{\min}=6701$,2014 年的 $L_{\min}=9804$,结果分别如图 11 和图 12 所示。从性能曲线的趋势来看,EMCNN 始终优于 MCNN;两者在训练集增加到一定程度 (图 11 中分别为 2000 和 4000,图 12 中分别为 6000 和 8000) 之后,开始持续优于 E-EMB 和 E-ESM,且 EMCNN 需要更少的训练数据;这 4 个模型都始终

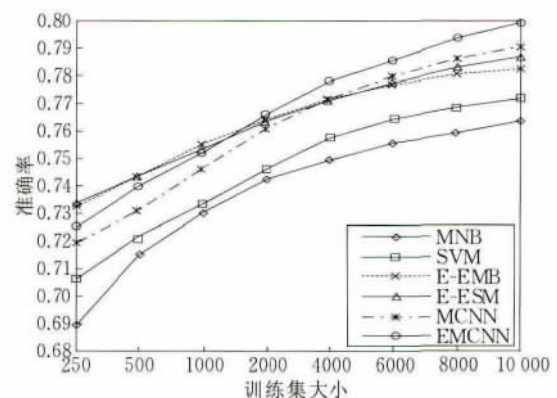


图 11 2013 年主客观分类

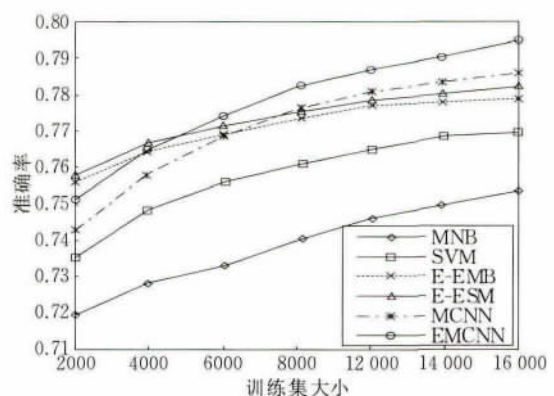


图 12 2014 年主客观分类

优于 MNB 和 SVM. 同文献[12]的 Fig. 3 对比发现, 本文实现的 E-ESM 模型以及 MNB 和 SVM 的实验结果都与其比较一致. 值得注意的是, 图 9 中 MNB 与 SVM 之间的差距增大, 且所有分类器的准确率都有所降低, 表明 2014 年的数据集在主客观分类上更加困难.

(2) 正负情感分类

此实验同样为二分类问题($n=2$), 其中 2013 年的 $L_{\min}=3639$, 2014 年的 $L_{\min}=4658$, 结果分别如图 13 和图 14 所示. 显而易见, 此实验结果的性能曲线与主客观分类实验具有相同结论. 不同之处有: 训练样本很少时, EMCNN 和 MCNN 相对于 E-ESM 和 E-EMB 的性能差距更大; E-ESM 和 E-EMB 的性能差距更加明显; EMCNN 和 MCNN 超越 E-ESM 和 E-EMB 时, 需要更少的训练样本(图 13 中分别为 1000 和 3000, 图 14 中分别为 2000 和 3000). 可以看出, 采用了表情符号空间的 E-ESM 和 EMCNN, 相对于 E-EMB 和 MCNN, 在正负情感分类时可以取得更高的性能增益. 需要注意的是, 同文献[12]的 Fig. 2 对比发现, 本文实现的 E-ESM 在训练样本小于 2000 时, 分类的准确率明显小于原文献, 这是由于采用不同的非监督训练集引起的(详细讨论见 4.4.4 节).

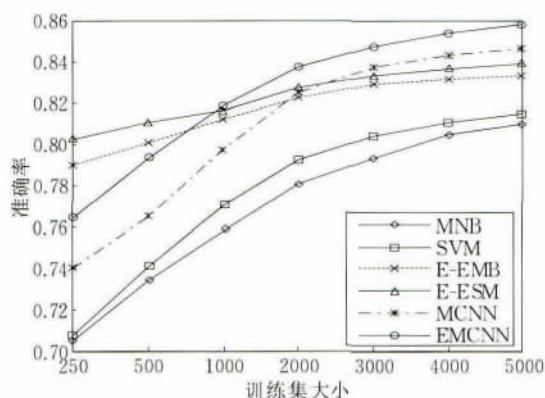


图 13 2013 年正负情感分类

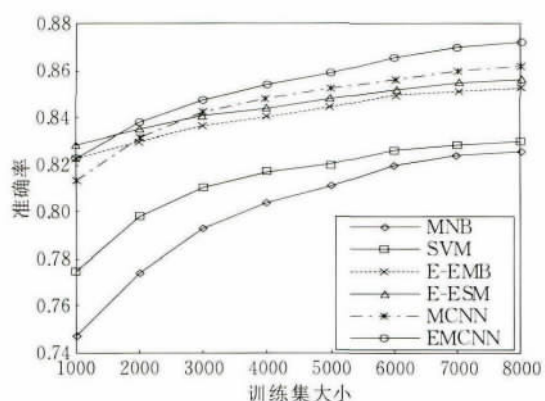


图 14 2014 年正负情感分类

(3) 情感四分类

此实验是四分类问题($n=4$), 2013 年的 $L_{\min}=1129$, 而 2014 年的 $L_{\min}=1362$, 结果分别如图 15 和图 16 所示. 显而易见, 此实验结果的性能曲线与前两个实验具有相同结论, 采用表情符号空间的模型仍然获得显著的性能增益; EMCNN 和 MCNN 超越 E-ESM 和 E-EMB 时, 需要的训练样本进一步减小(图 15 中分别为 1200 和 2400, 图 16 中分别为 1400 和 2000).

为了定量地分析实验结果, 特定义以下两个量化指标.

$$\begin{cases} \Delta_{\text{avg}} = \frac{1}{K} \sum_{i=1}^K acc_i^a - acc_i^b \\ ACC_{\text{span}} = acc_K - acc_1 \end{cases} \quad (14)$$

其中: K 为某次实验训练集大小的划分次数; acc^a 和 acc^b 分别为模型 a 和 b 的准确率; acc_i 为第 i 个训练集划分下的准确率, 那么 Δ_{avg} 表示模型 a 相对模型 b 的平均准确率增益; ACC_{span} 表示单个模型的最大准确率增益. 由于所有实验结果显示准确率随着训练集大小的增大而增大, 这里 acc_K 和 acc_1 就分别为准确率的最大值和最小值. 这里定义的两个量化指标, Δ_{avg} 能够指示出两个模型中较优的一个, 并得出增益大小; ACC_{span} 则可以表现出训练集大小对于不同实

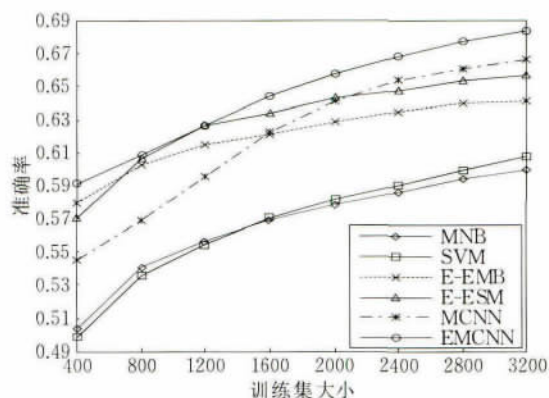


图 15 2013 年情感四分类

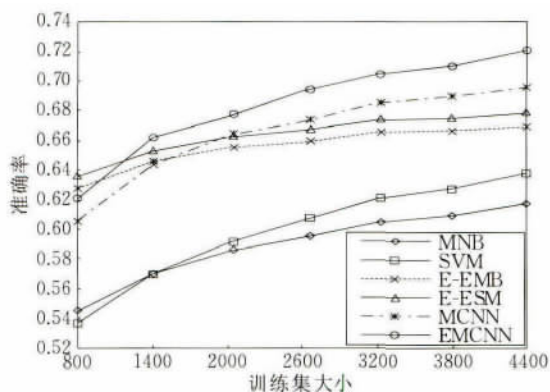


图 16 2014 年情感四分类

验和模型的影响程度,或者说不同实验和模型对于训练集大小的依赖程度。

本文实验的 6 个模型中,E-ESM 和 EMCNN 分别为 E-EMB 和 MCNN 的情感增强模型,因而选择 E-ESM 和 E-EMB、EMCNN 和 MCNN 计算平均准确率增益,分别记为 ESM-EMB 和 EMC-MCN,得到如图 17 和图 18 所示量化结果。可以看出,EMC-MCN 始终大大优于 ESM-EMB,表明 EMCNN 相比 ESM 具有更强的情感增强能力;纵观主客观分类、正负情感分类和情感四分类这 3 种不同实验,可以发现平均准确率增益依次增大,表明情感增强模型对于情感分类特别是多类别的情感分类更加有效。

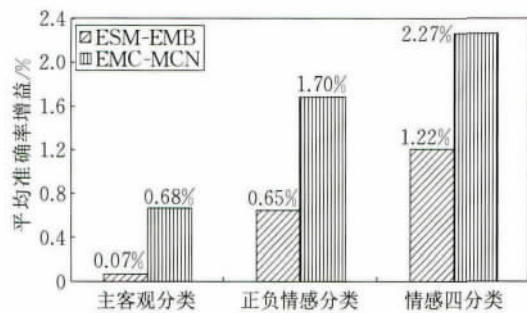


图 17 2013 年平均准确率增益

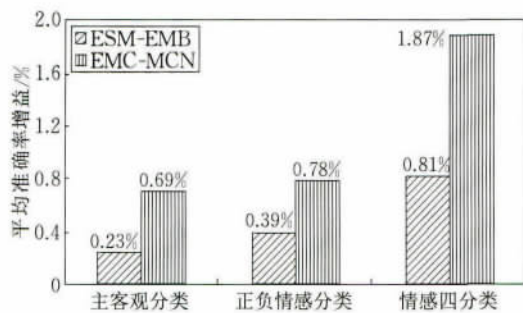


图 18 2014 年平均准确率增益

对于单个模型的最大准确率增益,得到如图 19 和图 20 所示量化结果。可以明显看出,E-EMB 和 E-ESM 对于训练集的依赖程度最小,同文献[12]中的结论一致;MCNN 和 EMCNN 对于训练集的依赖度,与传统机器学习模型 MNB 和 SVM 比较接近,因而 E-EMB 和 E-ESM 能够更加有效地利用有标注数据集。

纵观 3 个实验还可以发现,EMCNN 相对其它分类模型(特别是 MCNN)的增益不断增加,超越 E-ESM 需要的训练样本也越来越小。这些都表明利用表情符号空间增强 MCNN 模型是必要且成功的,同时也体现了深度学习模型可以更好地利用词向量技术。

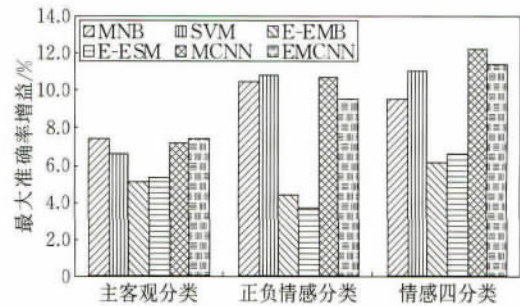


图 19 2013 年最大准确率增益

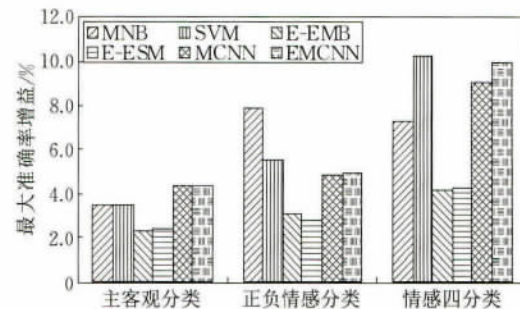


图 20 2014 年最大准确率增益

4.4.3 词向量的作用

之前的实验中,MCNN 和 EMCNN 都使用了预学习的词向量,并且在训练过程中对词向量进行更新(再学习)。为了研究词向量对于两个模型的影响,设计两个方案:(1) MCNN-r 和 EMCNN-r,使用随机生成的词向量(采用平均值为 0、方差为 0.25 的均匀分布),训练过程中进行更新;(2) MCNN-s 和 EMCNN-s,使用预学习的词向量,训练过程中不更新。选用 2014 年类别平衡数据集的情感四分类实验,同样是在不同训练集大小上各进行 10 次实验。为了更好的比较,实验结果分为图 21 和图 22,其中后者的性能曲线全面优于前者。从图 21 可以看出,采用随机生成的词向量使得模型性能严重下降,特别是需要更多的训练样本才能优于 MNB 和 SVM。

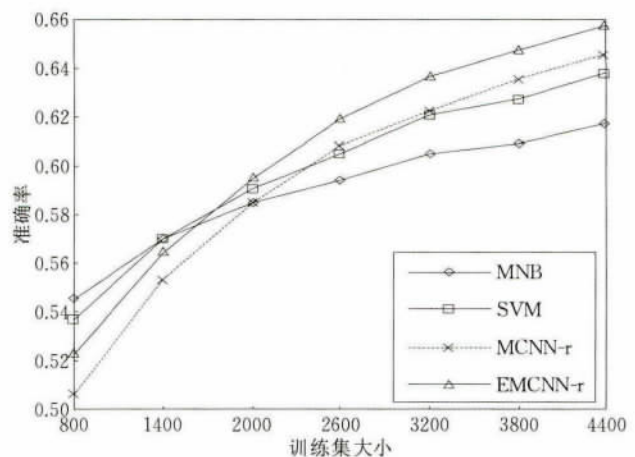


图 21 随机词向量的 MCNN

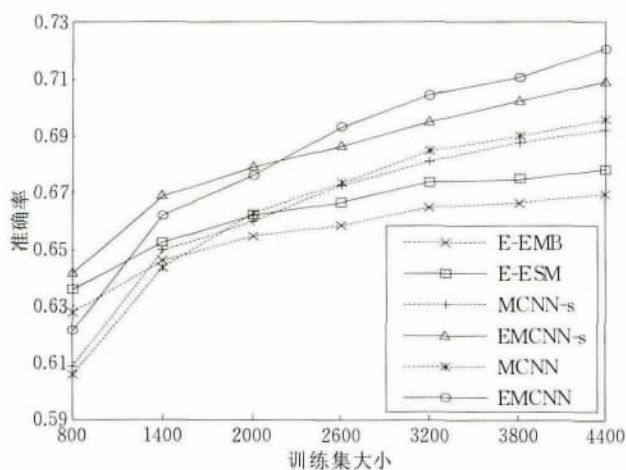


图 22 静态词向量的 MCNN

然而,图 22 显示的情况更为复杂. 首先,EMCNN-s 在训练样本很少时就性能优异,甚至超过了 E-ESM,但是随着训练样本的增加,逐渐被 EMCNN 超越,反映出词向量两个方面的作用:预学习的词向量在训练样本很少时发挥着重要影响;随着训练样本的增加,词向量需要充分地更新,才能更好地提升模型性能.

4.4.4 无标注数据集的影响

前一个实验表明,采用随机词向量和预学习词向量将对分类性能产生很大的影响,本节研究不同数据集预学习到的词向量对相关模型的影响. 之前实验采用的词向量都是从 full-set 数据集中学习得到的,这里将 emo-set 数据集和 part-set 数据集学习到的词向量应用到相关模型,则在模型名称后面分别添加“-e”和“-p”来区分. 仍然采用 2014 年类别平衡数据集的情感四分类实验,图 23 对 E-EMB 和 E-ESM 进行对比,图 24 则对 MCNN 和 EMCNN 进行对比.

从图 23 可以看出,E-EMB 对于数据集的大小比较敏感,因为更大规模的文本数据集学习到的词向量包含更多的语法和语义信息. E-ESM 同样需要更大规模的数据集,数据集规模越大则最终的准确率越高. 但是包含表情符号的文本对于 E-ESM 更加重要,具体表现为:训练集较小时,E-ESM-e 和 E-ESM-p 的准确率比较接近;而且,无论训练集的大小如何变化,E-ESM 与 E-ESM-p 之间准确率的差距,始终远远大于 E-EMB 与 E-EMB-p 之间准确率的差距.

从图 24 可以看出,训练集规模较小时,MCNN 和 EMCNN 都受无标注数据集影响较大,特别是包含表情符号的文本可以提高 EMCNN 的性能;但是随着训练集规模的增大,标注数据集的影响逐渐增强,每个模型的分性能也不断接近. 对比图 23 可

以发现,由于采用的深度学习模型做基础,EMCNN 相比 ESM 能够更好地利用标注数据集提升分类性能.

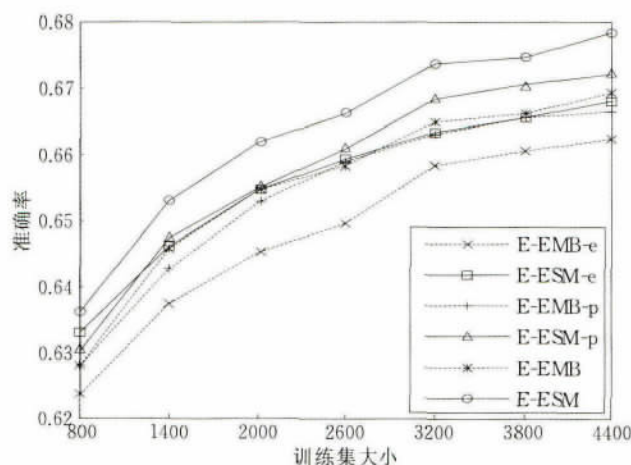


图 23 不同非监督数据集的 ESM

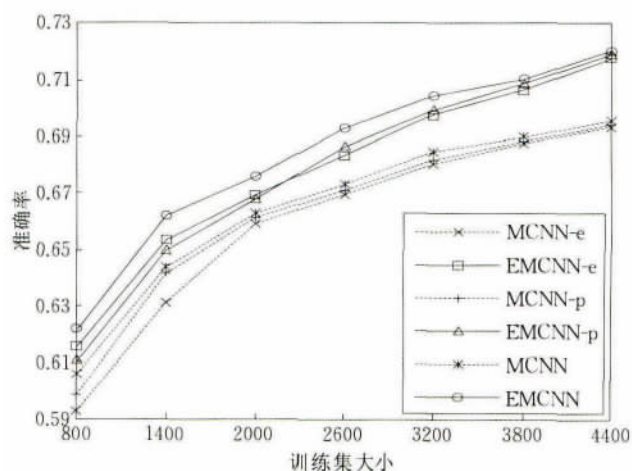


图 24 不同非监督数据集的 EMCNN

4.4.5 选取表情符号的数目

对于数据集中的表情符号,按照其使用频率降序排列,选取列表中的前 n 个(记为 Top- n)构建表情符号空间. 采用 2014 年类别平衡数据集的情感四分类,只选用训练样本大小为 4400 的实验,引入 B-ESM 同 E-ESM 和 EMCNN 进行比较,结果如图 25 所示. 可以看出,Top- n 需要选择一个适中的数值,太小的 Top- n 构建的表情符号空间,由于维度太少而判别力不足;太大的 Top- n ,容易引入判别力较弱的表情符号,从而降低情感语义的区分度. 本文前面的实验都选用 Top- $n = 100$,使得对比模型 B-ESM 和 E-ESM 达到最优,同时确保 EMCNN 的模型复杂度适中且性能优良.

4.4.6 模型训练耗时比较

通过 3.4 节的理论分析可以看出,MCNN 和

EMCNN 在时间复杂度上的不同,主要由于超参数 d 和 d_e 的不同引起,因而本节进行模型训练耗时比较时,固定除 d 和 d_e 之外的所有参数(d_{win} 和 H 参见表 6). 采用 SGD 训练人工神经网络模型,通常需要多次遍历训练集,实验中统一遍历 10 次取平均耗时,且 mini-batch 大小固定为 100 个样本. 实验平台和环境配置具体如表 9 所示.

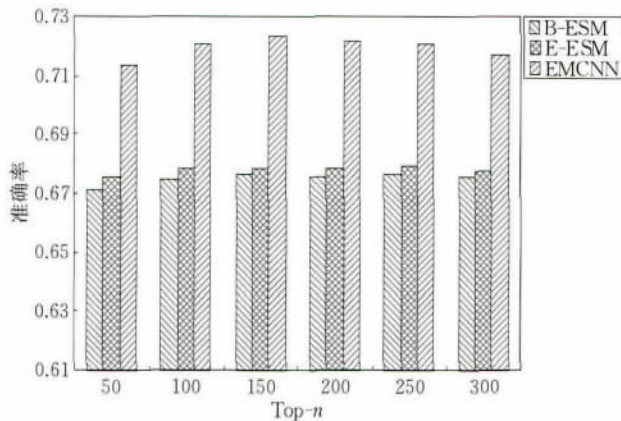


图 25 表情符号的选取

表 9 实验平台和环境配置

软硬件	配置
CPU	i5-3470, 3.20 GHz
内存	DDR3, 8GB, 1600 MHz
GPU	GeForce GTX 970 (4GB 显存)
操作系统	Ubuntu 14.04.4
开发环境	Python 2.7.11 Theano 0.8.0

对于 MCNN 模型,超参数 d 取值为 100 和 200, 分别记为 MCNN100 和 MCNN200;对于 EMCNN 模型,超参数 d 固定为 200,而超参数 d_e 取值为 100 和 200, 分别记为 EMCNN100 和 EMCNN200. 在 2014 年训练集上依次进行主客观分类和情感 7 分类,具体的实验结果如图 26 所示. 由于主客观分类的训练集大小为 13998,而情感 7 分类的训练集大小仅为 7407,因而后者的耗时整体较少. 正如理论分析的一样,超参数 d 和 d_e 对模型的训练耗时影响很大. 当 d 和 d_e 相等时,EMCNN 相对 MCNN 增加的耗时比较少,即 EMCNN 模型的 T_2 部分增加的耗时. 特别重要的是,4.4.1 节到 4.4.4 节的实验采用的 MCNN 和 EMCNN 就是这里的 MCNN200 和 EMCNN100,可以明显看出,EMCNN100 相比 MCNN200 不仅提高了分类性能,同时还减少了训练耗时,具体是主客观分类减少 36.15%,情感 7 分类减少 33.82%. 此外,对于 EMCNN100 和 MCNN200,式(13)中 M_2 是 M'_2 的一半,即空间复杂度也有所降低.

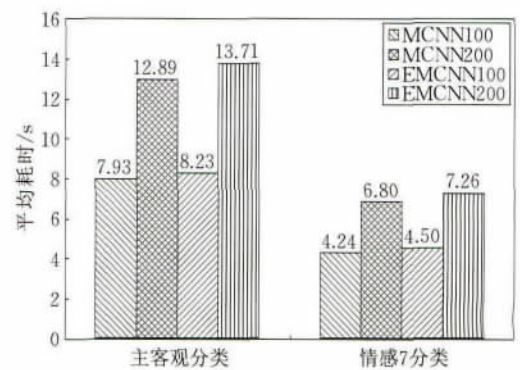


图 26 模型训练平均耗时

4.4.7 实例分析

如表 10 为从 14test 数据集中选取的微博实例,这里已对微博的文本内容进行了预处理. 对于编号为 241 和 956 的微博,情感标签都是“无”. 然而,这两条微博中都包含了情感词和表情符号,从而导致所有分类器都将它们判断为主观文本. 简单统计发现,这是导致 14test 主客观分类性能下降的原因之一. 对于 417 号微博,文本中存在多种不同情感的特征,例如“变态”和“畜生”表达“恶”的情感,“祝愿”表达“乐”的情感,“[怒]”则表达“怒”的情感. 这样的情况就要求分类器不仅能够区分这些特征的情感表达倾向,而且还必须根据上下文判断出主要表达的情感倾向. MNB 将此微博错分为“乐”,SVM、E-EMB 和 E-ESM 则错分为“怒”,只有 MCNN 和 EMCNN 分类正确,表现出以上两种特点. 对于 793 号和 4229 号微博,则只有 EMCNN 分类正确,采用文献[36]中 4.6 节分析 CNN 模型特征识别的方法可以发现,EMCNN 相比 MCNN 更多地选择了“美”、“[花心]”和“可怕”这些词语的卷积结果.

表 10 微博实例

微博编号	情感	文本内容
241	无	该是耐克表现伟大的时候了[泪]//<NAME>:[嘻嘻]
417	恶	心里变态,必须严惩!!祝愿男孩早日康复//<NAME>:这两个畜生一定要找出来严惩![怒]
793	喜	逆生长了!!!!//<NAME>:美哭了[花心][花心][泪][泪]
956	无	发发围脖碰碰运气,希望我是最幸运的那一个
4229	惧	对<NAME>说:请问我现在在天坛体育西路这边,应该怎么去八达岭长城,感觉北京到处都是骗子,好可怕,不知道怎么去了

5 结 论

本文提出了一个情感语义增强的深度学习模型

EMCNN 进行中文微博的多情感分类. 通过对表情符号列表的人工情感标注, 表明常用表情符号具有很好的情感指示作用; 尽管表情符号和情感词在词向量空间中都具有很好的情感语义区分度, 但是表情符号在情感语义空间中的判别度更好. 因此本文提出利用表情符号列表对应的词向量构建情感空间, 对其他词向量进行重映射之后, 直接输入 MCNN 进行情感分类. EMCNN 将情感语义的重映射与情感分类过程无缝结合起来, 增强了 MCNN 捕捉情感语义的能力, 并且能够充分利用无标注数据集和标注数据集. 在 NLPCC 的微博情感分类数据集上, EMCNN 取得最好分类效果, 并在多项性能指标上超过已知最好结果. 由于数据集的限制, 本文的情感空间是全局唯一的, 无法体现单个用户的情感表达特点. 一个思路是在 EMCNN 模型中为每个用户单独构建情感空间, 从而可以从用户的角度进行情感分类. 因此, 本文下一步的工作需要首先构建一个基于用户的情感数据集, 然后实验并改进 EMCNN 来更好地进行用户导向的情感分类.

参 考 文 献

- [1] Tumasjan A, Sprenger T O, Sandner P G, et al. Predicting elections with Twitter: What 140 characters reveal about political sentiment//Proceedings of the 4th International AAAI Conference on Web and Social Media (ICWSM-10). Washington, USA, 2010: 178-185
- [2] Jansen B J, Zhang M, Sobel K, et al. Micro-blogging as online word of mouth branding//Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI 2009): Extended Abstracts Volume. Boston, USA, 2009: 3859-3864
- [3] O'Connor B, Balasubramanyan R, Routledge B R, et al. From tweets to polls: Linking text sentiment to public opinion time series//Proceedings of the 4th International AAAI Conference on Web and Social Media (ICWSM-10). Washington, USA, 2010: 122-129
- [4] Liu B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1-167
- [5] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using Twitter hashtags and smileys//Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010): Posters. Beijing, China, 2010: 241-249
- [6] Kouloumpis E, Wilson T, Moore J D. Twitter sentiment analysis: The good the bad and the OMG !//Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM-11). Barcelona, Spain, 2011: 538-541
- [7] Schlichtkrull M S. Learning affective projections for emoticons on Twitter//Proceedings of the 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2015). Győr, Hungary, 2015: 539-543
- [8] Derks D, Bos A E, Von Grumbkow J. Emoticons and social interaction on the Internet: The importance of social context. Computers in Human Behavior, 2007, 23(1): 842-849
- [9] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168, 2013
- [10] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality//Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS 2013). South Lake Tahoe, USA, 2013: 3111-3119
- [11] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations//Proceedings of the 2013 Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2013). Atlanta, USA, 2013: 746-751
- [12] Jiang F, Liu Y, Luan H, et al. Microblog sentiment analysis with emoticon space model. Journal of Computer Science and Technology, 2015, 30(5): 1120-1129
- [13] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning//Proceedings of the 25th International Conference on Machine Learning (ICML 2008). Helsinki, Finland, 2008: 160-167
- [14] Kim Y. Convolutional neural networks for sentence classification //Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014). Doha, Qatar, 2014: 1746-1751
- [15] Zhao Yan-Yan, Qin Bing, Liu Ting. Sentiment analysis. Journal of Software, 2010, 21(8): 1834-1848(in Chinese) (赵妍妍, 秦兵, 刘挺. 文本情感分析. 软件学报, 2010, 21(8): 1834-1848)
- [16] Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 2005, 39(2-3): 165-210
- [17] Song K, Feng S, Gao W, et al. Build emotion lexicon from microblogs by combining effects of seed words and emoticons in a heterogeneous graph//Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT 2015). Guzelyurt, Northern Cyprus, 2015: 283-292
- [18] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques//Proceedings of the 2002 Conference on Empirical Methods on Natural Language Processing (EMNLP 2002). Philadelphia, USA, 2002: 79-86
- [19] Kim S M, Hovy E. Automatic identification of pro and con reasons in online reviews//Proceedings of the International Conference on Computational Linguistics (COLING) and the Association for Computational Linguistics (ACL): Posters. Sydney, Australia, 2006: 483-490
- [20] Barbosa L, Feng J. Robust sentiment detection on Twitter

- from biased and noisy data //Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010): Posters. Beijing, China, 2010: 36-44
- [21] Bermingham A, Smeaton A F. Classifying sentiment in microblogs: Is brevity an advantage? //Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010). Toronto, Canada, 2010: 1833-1836
- [22] Hu X, Tang J, Gao H, et al. Unsupervised sentiment analysis with emotional signals //Proceedings of the 22nd International Conference on World Wide Web(WWW 2013). Rio de Janeiro, Brazil, 2013: 607-618
- [23] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. Stanford, California, USA. 2009
- [24] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining //Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010). Valletta, Malta, 2010: 1320-1326
- [25] Liu K, Li W, Guo M. Emoticon smoothed language models for Twitter sentiment analysis //Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI-12). Toronto, Canada, 2012: 1678-1684
- [26] Yu Kai, Jia Lei, Chen Yu-Qiang, et al. Deep learning: Yesterday, today, and tomorrow. Journal of Computer Research and Development, 2013, 50(9): 1799-1804(in Chinese)
(余凯, 贾磊, 陈雨强等. 深度学习的昨天, 今天和明天. 计算机研究与发展, 2013, 50(9): 1799-1804)
- [27] Tang D, Qin B, Liu T, et al. Learning sentence representation for emotion classification on microblogs //Proceedings of the 2nd Natural Language Processing and Chinese Computing (NLPCC 2013). Chongqing, China, 2013: 212-223
- [28] Zhang D, Xu H, Su Z, et al. Chinese comments sentiment classification based on word2vec and SVM^{perf}. Expert Systems with Applications, 2015, 42(4): 1857-1863
- [29] Lebre R E M, Legrand J E L, Collobert R. Is deep learning really necessary for word embeddings? No, EPFL-REPORT-196986. Idiap, 2013
- [30] Tang D, Wei F, Yang N, et al. Learning sentiment-specific word embedding for Twitter sentiment classification //Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Baltimore, USA, 2014: 1555-1565
- [31] Tang D, Wei F, Qin B, et al. Building large-scale Twitter-specific sentiment lexicon: A representation learning approach //Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014). Dublin, Ireland, 2014: 172-182
- [32] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank //Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). Seattle, USA, 2013: 1642
- [33] Dong L, Wei F, Tan C, et al. Adaptive recursive neural network for target-dependent Twitter sentiment classification //Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Baltimore, USA, 2014: 49-54
- [34] Dong L, Wei F, Zhou M, et al. Adaptive multi-compositionality for recursive neural models with applications to sentiment analysis //Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI-14). Québec City, Canada, 2014: 1537-1543
- [35] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences //Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Baltimore, USA, 2014: 655-665
- [36] Dos Santos C N, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts //Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014): Technical Papers. Dublin, Ireland, 2014: 69-78
- [37] Xu Lin-Hong, Lin Hong-Fei, Pan Yu, et al. Constructing the affective lexicon ontology. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185(in Chinese)
(徐琳宏, 林鸿飞, 潘宇等. 情感词汇本体的构造. 情报学报, 2008, 27(2): 180-185)
- [38] Mitchell J, Lapata M. Composition in distributional models of semantics. Cognitive Science, 2010, 34(8): 1388-1429
- [39] Zeiler M D. ADADELTA: An adaptive learning rate method. arXiv preprint arXiv:1212.5701. 2012
- [40] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580. 2012



HE Yan-Xiang, born in 1952, Ph.D., professor, Ph.D. supervisor. His main research interests include compiler system, trusted software, software engineering and natural language processing.

SUN Song-Tao, born in 1986, Ph. D. candidate. His main research interests include nature language processing, sentiment analysis and machine learning.

NIU Fei-Fei, born in 1992, M. S. candidate. Her research interests include natural language processing and sentiment analysis.

LI Fei, born in 1980, Ph. D., associate professor. His main research interests include distributed parallel processing, computer network and social network.

Background

Microblog sentiment analysis has been hot topic for research ever since social media such as Twitter became popular. Many machine learning methods were adopted to treat sentiment analysis as a learning task. Feature engineering is essential to these methods, so researchers have explored a lot of characteristics in microblog as emotion features, such as hashtag, emoticon etc. Deep learning is a new paradigm of machine learning and has developed rapidly recent years. Its main idea is learning feature representation automatically from raw signals. Word embedding is a deep learning technique which widely used in NLP recently, and has been proved to be a better feature representation than n-gram. Word embedding can be used as traditional feature for classifier like SVM, or combined with other deep learning models for NLP. These

models usually have computational layers for semantic composition which is essential in nature language understanding. This paper proposed a new deep learning model EMCNN for Chinese microblog sentiment analysis. EMCNN is an enhanced version of MCNN by using emoticon embedding as emotion space projection operator. EMCNN is more capable of capturing emotion semantic than MCNN, so it can improve the sentiment classification performance. EMCNN achieved the state-of-the-art results on the NLPCC datasets.

This work is supported by the National Natural Science Foundation of China (Nos. 61303115, 61373039 and 61472290), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 2013014111002512).