情报学报 ISSN 1000-0135
第35卷第4期339-348，2016年4月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 4, 339-348 April 2016

doi:10.3772/j.issn.1000-0135.2016.004.001

基于迁移学习微博情绪分类研究——以H7N9微博为例[1)]

周清清[1] 章成志[1,2]

(1.南京理工大学信息管理系，南京 210094；
2.江苏省数据工程与知识服务重点实验室(南京大学)，南京 210093)

摘要 社交媒体的发展吸引大量用户，继而产生海量的用户生成内容。对用户生成内容的挖掘分析能够及时掌握用户的情绪动态，继而帮助事件处理、政策施行等。已有研究利用监督机器学习方法进行文本情绪分类，但是这类方法依赖于语料的标注、耗时耗力，并且存在领域适应性问题。迁移学习方法能够避免大量的语料标注、并且一定程度解决领域适应性问题。但是，目前迁移学习鲜有用于情绪分类任务。此外，情绪分类主要是针对博文等长文本，缺少针对微博短文本的相关实证研究。本文在主客观分类基础上，利用迁移学习方法对H7N9微博主观语料文本进行情感分类，并对结果进行情绪分类。实验结果表明，首先，设置形容词个数阈值为2时主客观分类效果最优；其次，利用迁移学习算法进行微博情感分类效果优于非迁移学习方法；最后，利用词频-相关频率作为特征权重计算方法时可以得到较好的情绪分类性能。

关键词 情感分类 情绪分类 迁移学习 微博挖掘

Microblog Emotion Classification Based on Transfer Learning ——A Case Study of Microblogs about H7N9

Zhou Qingqing [1] and Zhang Chengzhi [1,2]

(1. *Department of Information Management, Nanjing University of Science & Technology, Nanjing* 210094;
2. *Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing* 210093)

Abstract The development of social media has attracted lots of users, which generate mass user generated contents. The mining of the contents can grasp the emotional dynamics of users in time, and then provide assistance to event handling, policy implementation, etc. Existing researches use supervised learning methods to conduct emotion classification, which depends on the tagging of the corpus, are time-consuming and domain dependent. Transfer learning can avoid much corpus tagging, however, most of the existing transfer learning researches focus on emotion classification of long text. Therefore, this paper uses transfer learning method to conduct microblog sentiment classification based on the results of subjectivity classification, and then identify microblog emotions based on the sentiment polarities. The experimental results show that, the performance of subjectivity classification is optimal when the threshold of adjective is 2; Second, for sentiment classification, the performance of transfer learning algorithm is better than non-transfer learning method; Last, take Term frequency - Relevance frequency (TF-RF) as the feature weight calculation method can get best emotion classification performance.

Keywords sentiment classification, emotion classification, transfer learning, microblog mining

收稿日期:2015年6月3日

作者简介:**周清清**，女，1990年生，博士研究生，主要研究方向：文本挖掘与自然语言处理。**章成志**，男，1977年生，博士，教授，博士生导师，主要研究方向：信息组织、信息检索、数据挖掘及自然语言处理，E-Mail：zhangcz@njust.edu.cn。

1) 本文受国家社会科学基金重大项目“面向突发事件应急决策的快速响应情报体系研究”(13&ZD174)、国家社会科学基金项目“在线社交网络中基于用户的知识组织模式研究”(No. 14BTQ033)、杭州师范大学阿里巴巴复杂科学研究中心开放基金项目“基于多语言产品本体的电商评价情感分析研究”(No. PD12001003002003)资助。

1 前 言

互联网的飞速发展促进了社交媒体的兴盛,各大主流社交媒体用户数量激增。We Are Social 公司2014年全球社交网络数字移动报告显示,Facebook、YouTube 用户均已超过10亿人次,腾讯 QQ、新浪微博也已有超过5亿人次的用户①。用户数量的指数级增长,产生大量用户生成内容。对于用户生成内容的挖掘、分析和整理,有利于及时了解用户对于某一政策、事件的态度,能够有效疏导用户情绪,从而判断政策优劣以及确定事件处理方法。此外,还能够帮助企业掌握在线消费者对产品或服务的观点与态度,从而以此为基础改善企业产品质量或制定更好的营销方案。

情感分类即自动识别文本中用户表达的情感极性,从而快速了解大量用户对于某一事件、政策等的情感倾向,是持正面支持态度还是负面反对态度。情绪分类是指对用户表达的情感进行细分,确定其中包含的情绪,如负面情感可以细分为:愤怒、悲伤、恐惧、惊讶等。现有的情绪分类研究主要针对博客等长文本,如 Yang 等的工作[1],微博等短文本存在特征少、用词不规范、噪声干扰多等不利于情绪分类的问题。另一方面,情绪分类任务主要利用监督机器学习方法,这类方法依赖于大量的标注语料,但是实际应用中标注语料是非常欠缺的,人工标注成本高,且存在标注质量问题。基于迁移学习的分类能解决标注语料欠缺的问题[2],借助源领域的大量标注语料训练分类器从而对目标领域数据分类。

本文将迁移学习用于微博的情感分类,并对分类结果进行情绪分类。具体而言,我们首先利用基于规则的方法进行主客观分类,识别出主观微博;然后利用迁移学习方法对主观微博进行情感分类,得到微博的正负情感极性;随后对负面微博进行情绪分类;最终得到包括:高兴、愤怒、恐惧、悲伤以及惊讶五类情绪。本文实验结果表明,情感分类任务中迁移学习方法优于非迁移方法;情绪分类任务中以 TF-RF 计算特征权重,其分类性能最优。

2 相关研究概述

本文利用迁移学习进行微博情感分类,并将其结果用于情绪分类。情感分类是指判断用户表达的情感倾向。情绪分类则是分析用户的情绪,包括愤怒、悲伤、惊讶、恐惧、高兴等。

2.1 基于迁移学习的情感分类相关研究概述

文本的情感分类旨在自动预测文本的情感极性。现有的基于监督学习的方法通过标注的数据训练得到情感分类器,从而完成情感极性的自动分类。由于标注过程非常耗时耗力,且不同领域的同一词汇可能表述不同情感,因而情感分类性能的提高受到一定的限制。基于迁移学习的情感分类旨在将一个领域的标注语料用于另一领域的情感分类任务,可以在一定程度上解决情感分类中面临的语料标注问题。

2007年,Tan 和 Wu 等利用源领域训练生成的分类器标注部分目标领域的未标注样本,然后利用这些已标注样本再次训练分类器[3]。2009年,Li 和 Sindhwani 等提出一种情感迁移机制,该机制基于受约束非负矩阵来三角分解源领域与目标领域的术语文档矩阵[4]。2010年,Pan 和 Xi 等提出一种基于谱特征对齐的情感分类方法啊,他们通过领域无关词汇对齐不同领域的领域有关词汇,从而缩小两个领域的领域有关词汇之间的间隔,从而改善情感分类器性能[5]。2011年,Calais 和 Veloso 等利用迁移学习策略进行实时情感分析,他们首选计算社交媒体用户对于某个主题的偏好,然后将用户偏好迁移文本特征进行情感分析[6]。2014年,Fang 等提出了一种整合方法,即整合源领域标注数据的情感信息与一组预选情感词进行目标领域的情感分析[7]。

2.2 情绪分类相关研究概述

文本情绪分类是指对给定的文本预测其最有可能表达的情绪类别。2005年,Mishne 进行了博客文本情绪分类的初步试验,采用了多种特征,包括内容与非内容特征,以及一些在线文本的特有特征[8]。2005年,Alm 等利用监督机器学习和 SNoW 学习结构实证分析了文本情感预测问题,旨在对童话叙事结构域的句子情绪进行分类[9]。2006年,Jung 等提出了一种检测博客文本情绪的混合方法,包括高兴、悲伤、愤怒与恐惧[10]。此外,为捕捉博客文本的情绪转换,他们开发了基于情感流分析的段落级切分。

① http://wearesocial.net/blog/2014/01/social-digital-mobile-worldwide-2014/

2007 年，Yang 等利用支持向量机和条件随机场方法对网络博客语料的情绪分类，实验结果表明文本末句情绪对对于判断文本情绪尤其重要[1]。2007 年，Aman 和 Szpakowicz 介绍了识别情绪类别、情绪强度以及表达情绪的单词\词组的情绪标注任务[11]。2008 年，Strapparava 等介绍了六类基础情绪的大型数据集构造，包括：愤怒、厌恶、恐惧、喜欢、悲伤以及惊讶[12]。他们提出并评估了几种自动识别文本情感的基于知识方法和基于语料方法的性能。2008 年，Danisman 和 Alpkocak 利用向量空间模型，对 ISEAR（International Survey on Emotion Antecedents and Reactions）数据集进行自动文本情绪分类，包括：愤怒、厌恶、恐惧、喜欢以及悲伤[13]。2009 年，Keshtkar 等基于 Livejournal 博客语料的实验表明使用情感极性特征能够提高情绪分类的性能[14]。2014 年，Wen 与 Wan 介绍了一种基于类别序列规则的微博文本情绪分类方法，并在中文基准数据集上取得了优秀的性能[15]。2014 年，Chen 等利用过采样方法对小类别增加额外的句子向量，从而提高非均衡数据的情绪分类性能[16]。

现有研究表明，迁移学习与情绪分类已有较多研究见诸报道，但是鲜有利用迁移学习方法进行情绪分类任务；其次，目前情绪分类多在博客等长文本，在微博等短文本的情绪分析较少。因此，本文在主客观分类结果上，利用迁移学习方法对主观微博文本进行情感分类，并将分类结果进一步用于情绪分类，以识别 H7N9 微博文本情绪。

3 研究框架与关键技术描述

3.1 研究框架

本文以 H7N9 微博语料为例，进行基于迁移学习的情绪分类研究。总体的研究框架如图 1 所示，包括三个部分，即：微博语料的主客观分类、情感分类以及情绪分类。

在微博语料的主客观分类中，本文依据规则方法完成微博语料的主客观分类，即：通过文本中包含的形容词个数识别文本的主客观性[17]，认为包含 N 个以上形容词的文本为主观文本。在主观微博语料的情感分类中，本文分别利用迁移学习算法与非迁移学习算法进行情感分类，其中迁移学习包括朴素贝叶斯（NB）[18]，序列最小优化（SMO）[19]，K 近邻（KNN）[20]。最终得到微博的情感分类结果，即情感极性为正的微博（正面微博）和情感极性为负的微博（负面微博）。在情绪分类中，本文对包含负面情感的微博，依据 Ekman 的六大类情感分类体系进行情绪分类[21]，由于 H7N9 微博语料中愤怒与厌恶类别边界不明显，故将两类合并，最终得到愤怒、恐惧、悲伤以及惊讶等四种负面情绪分类结果。

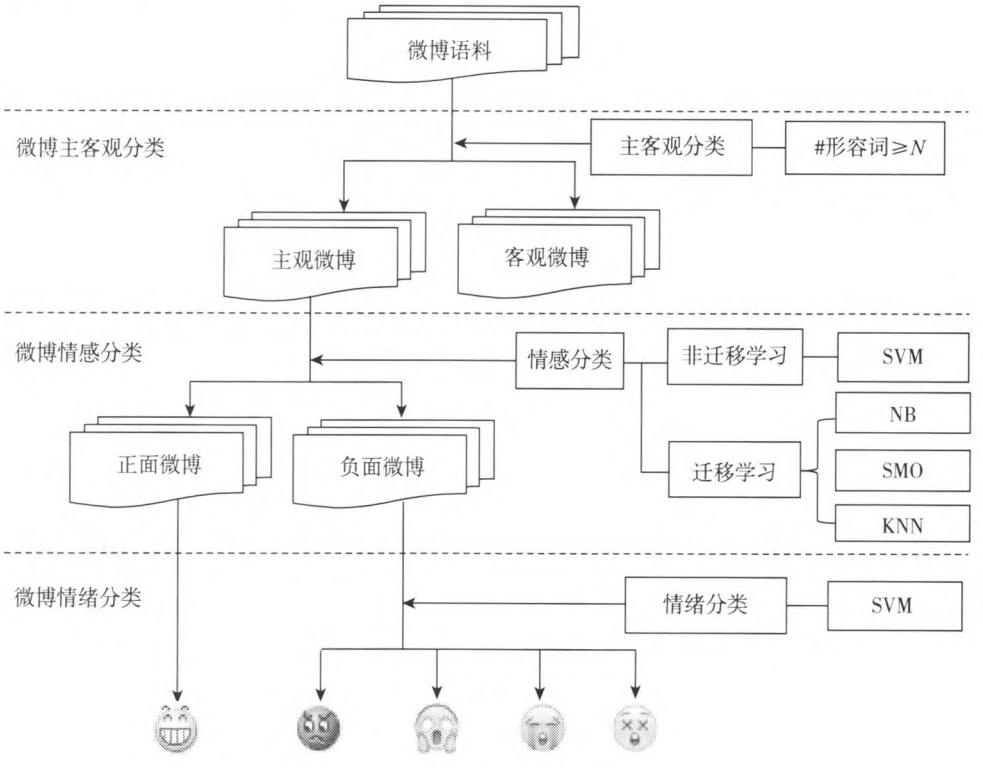

图 1 微博情绪分类流程图

3.2 关键技术描述

3.2.1 微博的主客观分类

表1 微博主客观分类示例

待分类微博文本	$N\geq1$	$N\geq2$	$N\geq3$
【预防 H7N9】尽量避免接触活禽，宰杀活禽尽快洗手，老年人尤其注意～扩散。	客观	客观	客观
据北京市卫生局消息，7月20日，一例河北来京就诊患者在北京朝阳医院确诊为人感染H7N9禽流感病例。患者来自河北廊坊，女，61岁，目前病情危重，正在呼吸重症监护病房进行全力抢救。	主观	客观	客观
#H7N9禽流感#清明节还是老实待家里别搞什么自驾游了，说不准遇到什么带毒禽类，真心觉得宅家里相对安全啊。	主观	主观	客观
#H7N9#港媒与外媒都吵翻天，国内一片详和，是因为我们吃地沟油、瘦肉精等不安全食品，百毒不侵，还有要维稳，政府怕引起恐慌。信息越不透明越恐慌。	主观	主观	主观

由于客观文本描述的是客观事实，故不具有情感、情绪分析的意义，而主观文本包含用户对产品、事件的观点、态度等，更具有分析意义，所以需要进行主客观分类。所谓主客观分类，即判断文本内容是主观的还是客观的。本文在进行主客观分类中之前，首先需要进行文本预处理。本文利用 Ansj① 进行分词与词性标注。

目前的主客观分类方法多数为监督学习方法[22,23]，但是本文缺少大量的标注数据可用于训练，故采用规则的方法进行主客观分类。通常，一句话中形容词起到修饰作用，这些修饰性的词，表达了用户的情感色彩，故本文利用形容词抽取主观句，认为包含 N 个以上情感词的文本为主观文本（其中，N 分别取值1,2,3），否则为客观文本[17]。

为了得到最优分类结果，本文比较了 N 等于不同数值对于分类结果的影响，部分示例如表1所示，具体分类结果见4.3节。

3.2.2 微博的情感分类

情感分类即判断主观文本中用户表达的情感倾向，是正面的还是负面的。迁移学习方法能够很好的解决目标数据缺乏标注语料的问题。迁移算法假设源领域与目标领域具有相似的边际分布与条件分布，但现实应用中，两个领域的边际分布与条件分布可能是显著不同的，因此这种情况下线性迁移算法将不适用。为确保假设合理，本文利用非线性迁移算法，即利用基于核映射的迁移学习算法进行情感分类[24]，首先利用自适应核方法映射源领域和目标领域数据的边际分布至同一个核空间，然后利用基于聚类的样本选择策略选择与目标领域分布相似样本用于训练分类器，其中：

（1）我们采用高斯内核（Gaussian kernel），并且根据 Zhu 等[25]提出的启发式策略设置内核距离，进行特征映射；

（2）我们采用二分 K 均值（Bisecting K-means）方法与 Ren 等[26]提出的自适应调整策略，进行样本选择。首先对源领域与目标领域的标注样本进行聚类，然后利用各个类簇中目标样本的类别标签确认类簇的类别标签，最后选择各个类簇中与类簇标签一致的源领域样本进行分类器训练。

为了减少单一映射空间的偏差，采用多次迭代训练分类器，最终整合分类器；为降低错分样本的影响，为每个样本赋予权重（初始权重为1.0），每次迭代中减少错分样本的权重，增加正确分类样本的权重（本文迭代次数为5），具体迁移学习算法[24]流程如图2所示。首先，对源领域语料与目标领域的标注语料进行初始权重设置，然后根据聚类抽样的方法抽选出源领域中与目标领域分布相似的样本，与目标领域合并作为训练语料。接着利用核映射方法将训练样本映射至同一个特征空间从而进行训练器学习。利用训练好的分类器对目标语料无标注样本进行情感分类，根据分类评估结果调整训练样本的权重；最后重复上述过程直至迭代结束。

权重更新计算公式[24]如下：

① http://www.ansj.org/

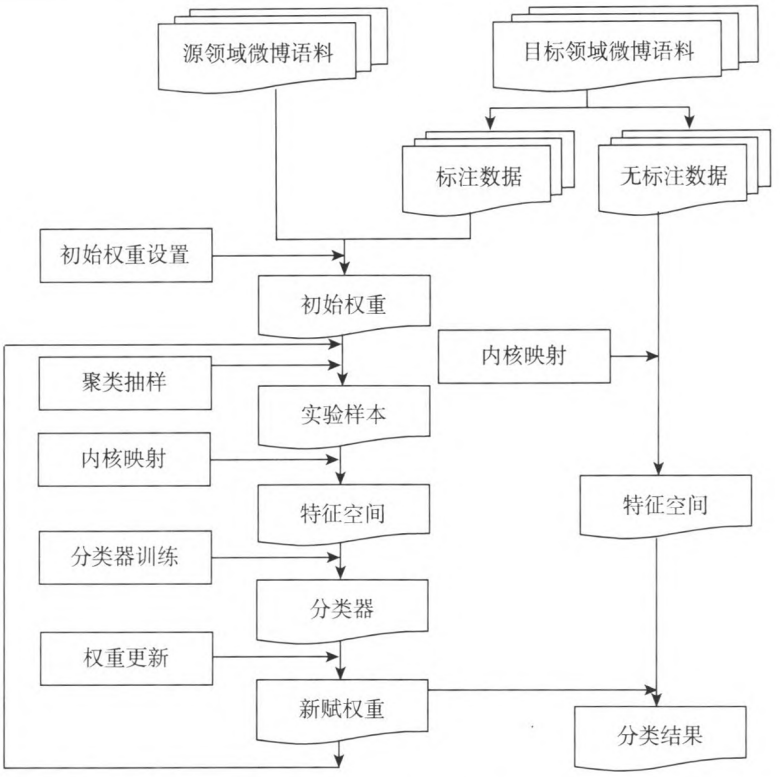

图 2　基于迁移学习的微博情感分类流程图

$$wi_j = \begin{cases} wi-1_j\gamma - |C_i(X_j) - y_j|_i, 1 \leqslant j \leqslant l \\ wi-1_j\gamma - |C_i(X_j) - y_j|, 1+l \leqslant j \leqslant l+o \end{cases} \tag{1}$$

$$\gamma_i = \varepsilon_i/(1-\varepsilon_i) \tag{2}$$

$$\varepsilon_i = \sum_{j=1}^{l} \frac{wi_j |C_i(X_j) - y_j|}{\sum_{j=1}^{l} wi_j} \tag{3}$$

$$\gamma = 1/\sqrt{(2ln\ (l+o)/N)} \tag{4}$$

其中，wi_j为第 i 次迭代样本 j 的权重，C_i为第 i 次迭代训练生成的分类器，y_i为第 i 次迭代样本 j 的类别标签，l 为目标领域有标注样本数量，o 为源领域样本数量，N 为迭代总次数。

最终分类结果计算如公式(5)所示：

$$TU_j = \begin{cases} 1, \prod_{i=N/2}^{N} \gamma - C_i(X_j)_i \geqslant \prod_{i=N/2}^{N} \gamma - 1/2_i \\ 0, otherwise \end{cases} \tag{5}$$

其中，TU_j为目标领域未标注样本 j 的分类结果。

本文共使用了三个分类器：NB、SMO、KNN。NB 是通过待分类样本的先验概率，利用贝叶斯公式计算出其后验概率，即该样本属于某一类的概率，选择具有最大后验概率的类作为该样本所属的类。SMO 算法提出后成为最快的二次规划优化算法，特别针对线性 Support Vector Machine (SVM)[27] 和数据稀疏时性能更优。KNN 算法即通过待分类样本周围的 k 个邻居来判断它的类别。如果待分类样本在特征空间中的 k 个最邻近的样本中的大多数属于某一个类别，则该样本也属于这个类别，在本文中设置 $K=3$。

与主客观类似，在进行情感分类之前，文本预处理过程必不可少。本文利用 Ansj 进行分词，利用 CHI[28] 进行特征选择，利用 Term Frequency-CHI (TF-IDF)进行特征权重计算，计算公式如下：

$$X^2(t,C_i) = \frac{N*(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{6}$$

其中，A 是特征 t 和类别C_i共同出现的次数，B 是特征 t 出现而类别C_i不出现的次数，C 是特征 t 不出现而类别C_i出现的次数，D 是特征 t 和类别C_i都不出现的次数，$N=A+B+C+D$。

$$TF-CHI = \frac{\#wordt}{\#word} * X^2(t,C_i) \tag{7}$$

其中，$\#wordt$ 是特征 t 在文本中出现的次数，$\#word$ 是文本的总词数。

3.2.3　微博的情绪分类

情绪分类即识别文本表达的情绪，本文涉及的情绪共包括五种，分别是：高兴、愤怒、恐惧、悲伤、惊讶，其中高兴属于正面情绪，后四种属于负面情绪。由于情感分类已经识别文本情感倾向，即已经识别正面情绪，故本文主要判断负面情绪。本文尝试利

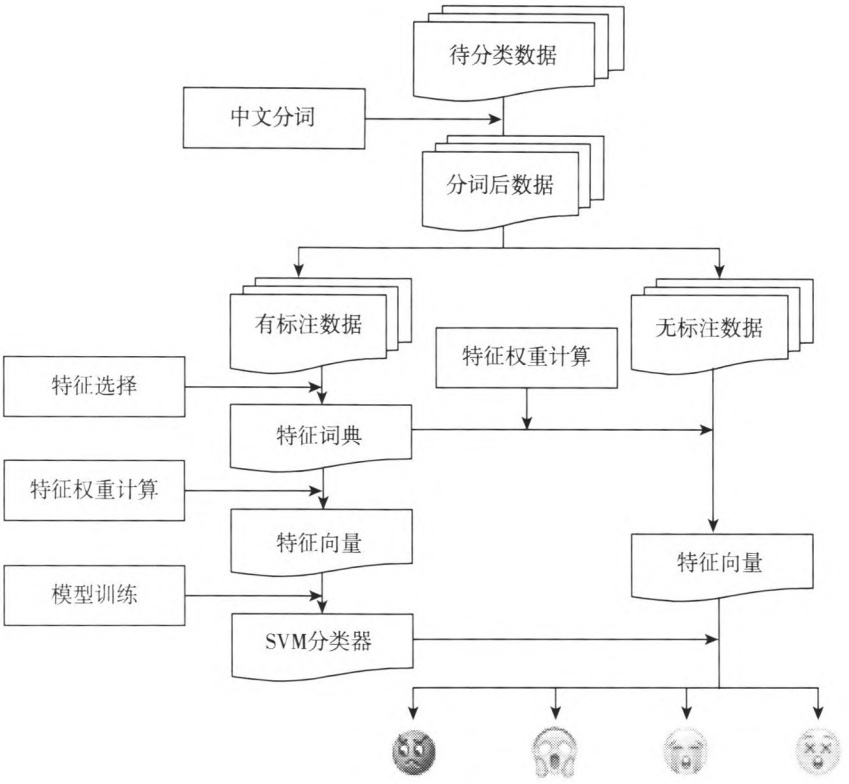

图3 微博负面情绪分类流程图

用迁移学习方法进行情绪分类,得到的分类效果较差,可能原因是微博等短文本存在特征少,噪声多等特点。故本文利用SVM分类器对H7N9中表达负面情感的微博进行情绪分类,具体情绪分类流程如图3所示。

本文利用LibSVM① 进行情绪分类。LibSVM是台湾大学林智仁(Lin Chih-Jen)教授等开发设计的一个简单,易于使用和快速有效的SVM模式识别与回归的软件包。在预处理过程中,本文仍然利用Ansj进行分词,利用CHI进行特征选择,并且比较了不同特征权重计算方法对于分类结果的影响,主要包括Term FrequencyInverse Document Frequency (TF-IDF)[29]、IF - CHI[30] 以及Term Frequency-relevance frequency (TF-RF)[31],计算公式如下:

$$TF-IDF=\frac{\#wordt}{\#word}*log\left(\frac{\#doc}{\#doct+1}\right) \quad (8)$$

其中,$\#wordt$ 是特征 t 在文本中出现的次数,$\#word$ 是文本的总词数,$\#doc$ 是文本集中的文本总数,$\#doct$ 是文本集中包含特征 t 的文本总数。

$$TF-RF=\frac{\#wordt}{\#word}*log\left(2+\frac{A}{C}\right) \quad (9)$$

其中,定量2是由于对数操作是以2为底的[31],A、C 的定义类似于CHI公式,即 A 是特征 t 和类别 C 共同出现的次数,C 是特征 t 不出现而类别 C 出现的次数。

4 实验与结果分析

4.1 实验数据概述

本文数据为来自新浪微博关于H7N9的真实数据。所采集的数据内容包括微博内容与发布时间,共采集微博20 027条。部分实验数据实例表2所示。

为了验证基于迁移学习的微博情感分类的效果,本文收集NLP&CC2013中文微博情绪分析评测任务②的已标注语料作为分类模型训练语料。该语料包括微博内容以及该条微博的情绪,共32 185条微博,除客观微博之外,共有7类情绪:高兴、喜好、愤怒、厌恶、悲伤、恐惧、惊讶。鉴于本文基于Ekman的六大类情感分类体系以及H7N9微博的特点,将高兴与喜好合并,将愤怒与厌恶合并,最终得到高兴、愤怒、恐惧、悲伤、惊讶五类情绪。为测试分类的准确性,本文标注了部分H7N9微博语料,具体如表

① http://www.csie.ntu.edu.tw/~cjlin/

② http://tcci.ccf.org.cn/conference/2013/pages/page04_eva.html#

3 所示。

表 2 H7N9 微博文本样例

微博文本	发布时间
【广东已确诊 H7N9 禽流感病例达 70 例】广东省卫生计生委昨天通报，佛山市新增 1 例人感染 H7N9 禽流感确诊病例。广东省的 H7N9 病例至此已达 70 例，其中死亡 14 例	2014.02.22
你妹的人传人!!!! //@ 微博位置：北京新确诊 H7N9 患者病危，病毒突变或可通过飞沫传播	2013.07.23
我觉得我也肯定感染了什么 H7N9 禽流感了。自从上礼拜五吃了烧烤以来难过了一个礼拜了……到现在又喉咙痛又感冒的	2013.04.01
广东竟然也出现 H7N9 了	2013.08.10

表 3 微博情绪分类训练及测试集

主客观	情感分类	情绪分类	NLP&CC2013 微博情绪数据	H7N9 微博情绪标注数据
客观	客观	客观	21 633	1 506
主观	负面情感	愤怒	3 298	271
		恐惧	167	42
		悲伤	1 578	130
		惊讶	461	56
	正面情感	高兴	5 048	477

4.2 实验结果评价方法

为评价分类结果的准确性，本文采用的评价指标包括：精确率（Precision）、召回率（Recall）、F_1 值、微平均精确率（MicroPre）、宏平均精确率（MacroPre）、宏平均召回率（MacroRec），计算方法公式分别如下：

$$\#Precision_i = \#corret_i / \#classsum_i \tag{10}$$

$$\#Recall_i = \#corret_i / \#sum_i \tag{11}$$

$$MicroPre = \frac{\sum_{i=1}^{n} \#corret_i}{\sum_{i=1}^{n} \#classsum_i} \tag{12}$$

$$MacroPre = \frac{1}{n}\sum_{i=1}^{n} \frac{\#corret_i}{\#classsum_i} \tag{13}$$

$$MacroRec = \frac{1}{n}\sum_{i=1}^{n} \frac{\#corret_i}{\#sum_i} \tag{14}$$

$$F_1 = \frac{2 * MacroPre * MacroRec}{MacroPre + MacroRec} \tag{15}$$

其中，$\#corret_i$ 为分类器正确分类为类别 i 的微博数目，$\#classsum_i$ 为分类器分为类别 i 的全部微博数目；n 表示类别数，$\#sum_i$ 为人工分类为类别 i 的微博数目。

4.3 H7N9 微博语料的实证结果分析

（1）H7N9 微博主客观分类结果分析

我们利用基于规则的方法对 H7N9 微博内容进行主客观分类。我们以 2482 条标注语料作为测试集（其中客观微博 1506 条，主观微博 976 条），分别测试了设定不同形容词个数阈值情况下，主客观分类的准确性，测试结果如表 4 所示。

表 4 H7N9 微博语料主客观分类正确率

形容词个数（N）	宏平均召回率	宏平均精确率	F_1 值
≥1	0.5392	0.5339	0.5365
≥2	0.5364	0.5379	0.5371
≥3	0.5233	0.5199	0.5216

从表 4 可以看出，$N \geq 1$ 时，宏平均召回率最高，其次为 $N \geq 2$ 时；$N \geq 2$ 时，宏平均精确率最高，其次为 $N \geq 1$ 时；$N \geq 2$ 时，F_1 值，其次为 $N \geq 1$ 时。综上所述，本文认为包含 2 个及以上形容词的微博为主观文本，否则为客观文本。分类结果如表 5 所示，可以看出，关于 H7N9 的微博大多数为客观文本，约占 56%，主观文本约占 44%。

表 5 H7N9 微博语料主客观分类结果

微博总数	客观微博	主观微博
20 027	11 201	8 826

（2）H7N9 微博情感分类结果分析

我们利用基于迁移学习的方法对微博进行情感分类，并比较迁移学习分类与非迁移学习分类结果。源领域数据为 NLP&CC2013 中文微博情绪已标注语料中的主观文本，共 10 552 条微博数据，其中正面微博 5048 条，负面微博 5504 条；目标领域数据为 H7N9 微博内容中的主观文本，共 8826 条微博，其中有标注样本 976 条（包括 477 条正面微博与 499 条负面微博），无标注微博样本 7850 条。我们比较了迁移学习与非迁移学习的分类准确率，结果如表 6 所示。其中迁移学习采用三种分类器，分别是

NB、SMO、KNN，非迁移学习使用 SVM 分类器。为了将各个分类的最优分类结果组合，我们将 H7N9 未标注的主观性数据平均分成三部分，分别得到三部分（Part1、Part2、Part3）的分类准确率。

表 6 H7N9 微博语料情感分类准确率

		迁移学习			非迁移学习
分类器		NB	SMO	KNN	SVM
整体准确率		88.6671	86.0265	83.2786	0.738266
部分准确率	Part 1	89.0433	88.3927	87.3211	
	Part 2	82.7669	87.1680	89.9617	
	Part 3	84.4529	89.2349	84.2617	

从表 6 可以看出，不管是整体准确率还是部分准确率，迁移学习的分类准确率均高于非迁移学习。其中整体准确率部分，NB 分类器得到的准确率最高；在各部分准确率中，Part 1，NB 分类器准确率最高，其次为 SMO 分类器；Part 2，KNN 分类器准确率最高，其次为 SMO 分类器；Part 3，SMO 分类器准确率最高，其次为 NB 分类器；此外，各部分最高准确率均高于各自分类器的整体准确率。故我们将 Part 1 的 NB 分类结果、Part 2 的 KNN 分类结果以及 Part 3 的 SMO 分类结果组合得到最终的分类结果，如表 7 所示。从表 7 可以看出，关于 H7N9 的主观微博，约有 63% 微博表达了负面情感，其数量远高于表达正面情感的微博数量。

表 7 H7N9 微博语料情感分类结果

主观微博总数	正面微博	负面微博
8826	3248	5578

（3）H7N9 微博情绪分类结果分析

本文利用基于监督学习的方法对表达负面情感的微博进行了情绪分类，其中以 H7N9 微博中有标注样本作为训练集训练 SVM 分类器，包括愤怒情绪微博 271 条、恐惧情绪微博 42 条、悲伤情绪微博 130 条以及惊讶情绪微博 56 条。我们比较了不同特征权重计算方法的五折交叉验证的分类准确率，结果如表 8 所示。其中特征权重计算方法包括三种，分别是 TF-CHI、TF-RF 以及 TF-DF。

从表 8 可以看出，利用 TF-RF 计算特征权重得到的分类精确率最高，故本文以该方法得到的分类结果作为最终结果，如图 4 所示。从图 4（a）可以看出，负面情绪中，愤怒为主要情绪，占全部负面微博的 68%，其次为悲伤情绪。恐惧情绪比例最低，仅有 2%。图 4（b）为关于 H7N9 的微博全部情绪分类结果，其中约有 43% 微博表达了愤怒情绪，其次约有 37% 微博在传达正面情绪高兴。微博中表达恐惧或惊讶情绪的比例较低。

表 8 H7N9 微博语料负面情绪分类精确率

特征权重	TF-CHI	TF-RF	TF-IDF
微平均精确率	0.943662	0.985915	0.983903
宏平均精确率	0.907299	0.980895	0.971583

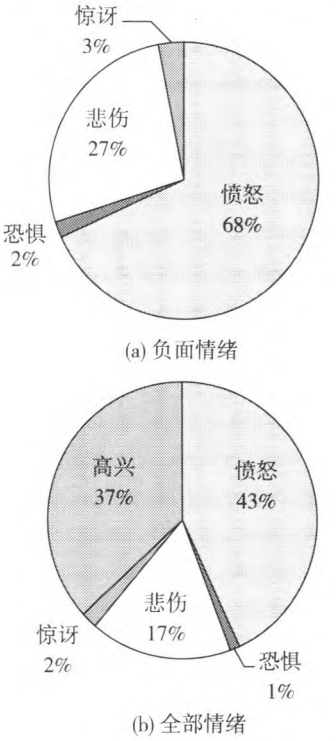

图 4 H7N9 微博语料情绪分类结果分布图

5 结论与展望

本文在主客观分类结果基础上，利用迁移学习方法对主观微博内容进行情感分类，并将分类结果进一步进行情绪分类，从而了解微博用户的情绪。实验结果表明：

（1）主客观分类任务：形容词个数阈值设置为 2 时，主客观分类效果最优；

（2）情感分类任务：利用迁移学习算法进行微博情感分类性能优于非迁移学习方法，而在情绪分类中的迁移学习效果差于非迁移学习。或许与短文

本特征少、噪声多等特性有关；

(3)情绪分类任务：利用 TF-RF 进行特征权重计算时可以得到较好的分类性能。

新浪微博中关于 H7N9 的微博数量众多，本文仅随机抽样了 20 000 余条微博进行情绪分类，今后需要扩大数据集以得到更全面的情绪分类结果，包括语料数量以及语料来源的扩展；其次，由于在前期实验中基于迁移学习的的情绪分类结果不佳，本文尚未对该问题做深入研究，今后我们将尝试更多的算法，从而提高迁移学习情绪分类的准确率。

参 考 文 献

[1] Yang C, Lin H Y, Chen H H. Chen. Emotion classification using web blog corpora[C]//Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007: 275-278.

[2] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(10): 1345-1359.

[3] Tan S, Wu G, Tang H, et al. A novel scheme for domain-transfer problem in the context of sentiment analysis [C]//Proceedings of the 16th ACM Conference on Information and Knowledge Management, 2007: 979-982.

[4] Li T, Sindhwani V, Ding C, et al. Knowledge transformation for cross-domain sentiment classification [C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009: 716-717.

[5] Pan Sinno Jialin, Ni Xiaochuan, Sun Jian-Tao, et al. Cross-domain sentiment classification via spectral feature alignment [C]//Proceedings of the 19th International Conference on World Wide Web, 2010: 751-760.

[6] Calais Guerra P H, Veloso A, Meira W, et al. From bias to opinion: a transfer-learning approach to real-time sentiment analysis [C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011: 150-158

[7] Fang F, Dutta K, Datta A. Domain Adaptation for Sentiment Classification in Light of Multiple Sources[J]. INFORMS Journal on Computing. 2014, 26(3): 586-598.

[8] Mishne G. Experiments with mood classification in blog posts[C]//Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access, 2005 (19): 321-327.

[9] Alm C O, Roth D, Sproat R. Emotions from text: machine learning for text-based emotion prediction [C]// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005: 579-586.

[10] Jung Y, Park H, Myaeng S H. A hybrid mood classification approach for blog text[C]//Proceedings of the PRICAI 2006: Trends in Artificial Intelligence, 2006: 1099-1103.

[11] Aman S, Szpakowicz S. Identifying Expressions of Emotion in Text[M]. Text, Speech and Dialogue. Springer Berlin Heidelberg, 2007: 196-205.

[12] Strapparava C, Mihalcea R. Learning to Identify Emotions in Text[J]. Unt Scholarly Works, 2008, 43(3): 254-255.

[13] Danisman T, Alpkocak A. Alpkocak. Feeler: Emotion classification of text using vector space model[C]// Proceedings of the AISB 2008 Convention Communication, Interaction and Social Intelligence, 2008(1): 53:59.

[14] Keshtkar F, Inkpen D. Using sentiment orientation features for mood classification in blogs[C]//Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2009: 1-6.

[15] Wen S, Wan X. Emotion Classification in Microblog Texts Using Class Sequential Rules[C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014: 187-193.

[16] Chen T, Xu R, Lu Q, et al. A Sentence Vector Based Over-Sampling Method for Imbalanced Emotion Classification [M]. Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2014: 62-72.

[17] Hatzivassiloglou V, Wiebe J M. Effects of adjective orientation and grad ability on sentence subjectivity [C]//Proceedings of the 18th Conference on Computational Linguistics-Volume 1, 2000: 299-305.

[18] Murphy K P. Naive bayes classifiers[M]. University of British Columbia. 2006.

[19] Platt J C. Sequential minimal optimization: A fast algorithm for training support vector machines[C]// Advances in Kernel Methods-support Vector Learning, 2015: 212-223.

[20] Larose D T. k-Nearest Neighbor Algorithm[M]. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc, 2005: 90-106.

[21] Ekman P. Basic emotions//Dalgleish T, Power M. Handbook of Cognition and Emotion. Sussex, U. K.: JohnWiley & Sons, Ltd., 1999.

[22] Wiebe J M, Bruce R F, O'Hara T P. Development and use of a gold-standard data set for subjectivity classifications[C]// Proceedings of the 37th annual

meeting of the Association for Computational Linguistics on Computational Linguistics, 1999: 246-253.

[23] Raaijmakers S, Kraaij W. A Shallow Approach to Subjectivity Classification[C]// Proceedings of the ICWSM, 2008: 216-217.

[24] Zhong E, Fan W, Peng J, et al. Turaga and O. Verscheure. Cross domain distribution adaptation via kernel mapping [C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009: 1027-1036.

[25] Zhu X, Ghahramani Z, Mit T J. Semi-supervised learning with graphs [C]// International Joint Conference on Natural Language Processing, 2005: 2465 - 2472.

[26] Ren J, Shi X, Fan W, et al. Type-Independent Correction of Sample Selection Bias via Structural Discovery and Re-balancing[C]// Proceedings of the SIAM, 2008: 565-576.

[27] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[28] Ng H T, Goh W B, Low K L. Feature selection, perceptron learning, and a usability case study for text categorization [C]// Proceedings of the ACM SIGIR Forum, 1997, 31(SI): 67-73.

[29] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval ☆[J]. Information Processing & Management An International Journal, 1988, 24(5): 513-523.

[30] Deng Z H, Tang S W, Yang D Q, et al. A Comparative Study on Feature Weight in Text Categorization [J]. Photochemistry and photobiology, 2004, 25(3): 326-336.

[31] Lan M, Tan C L, Low H B. Proposing a new term weighting scheme for text categorization [C]//Proceedings of the AAAI, 2006 (6): 763-768.

（责编编辑 王海燕）