基于分割和分类联合模型的中文微博情感分析*

陈 波[1,2]，姬东鸿[2]，吕晨[2]，柏云深[2]，魏小梅[2]

(1.湖北文理学院 文学院，湖北 襄阳 441053；2.武汉大学 计算机学院，湖北 武汉 430072)

摘要：研究微博情感分析具有重要的理论意义和现实价值。当前的中文情感分析研究性能较差，主要原因是已有的算法采用串行的模式对其进行研究，即先分词，然后根据分词的结果采取特征建模，进而来训练分类器。串行模式的缺点是分词的错误会进行传播，从而影响分类器的性能。本文提出一种创新的基于分割和分类的联合模型来实现中文微博情感分析。首先根据候选生成模型为句子生成多个候选，基于候选排序模型对该句子进行打分，选择分数最高的前 K 个候选，放入情感分类模型里面来训练分类器。根据分类器的性能高低来调整候选排序模型的参数权重。以此来进行迭代，直到算法收敛。本文所提算法不仅能修正一部分分词结果，而且可以生成一些情感描述短语，使得情感分析的性能获得提升。在 NLP&CC2013 中文微博数据集上的实验证实了本文所提算法的有效性。

关键词：微博；情感分析；联合模型；候选生成模型；候选排序模型

中图分类号：TP391　　**文献标识码：**A

A Joint Segmentation and Classification Model for Chinese Sentiment Analysis in Microblog

Bo Chen[1,2], Donghong Ji[2], ChenLv[2], Yunshen Bai[2], Wei Xiaomei[2]

(1.Department of Language & Literature, Hubei University of Art & Science, Xiangyang, Hubei 441053, China; 2.NLP Lab, Wuhan University, Wuhan, Hubei 430072, China)

Abstract: Research on microblogging sentiment analysis has great theoretical and practical value. Currently, the performance of Chinese sentiment analysis is not good enough, mainly due to the pipeline model. It extracts features based on segmented input, and then trains the classifier. Disadvantages of the pipeline model are that the errors of word segmentation will propagate to the performance of the classifier. This paper presents an novel Joint Segmentation and Classification Model for Chinese microblogging sentiment analysis. Firstly, we generate a set of candidates according to the candidate generating model, score the candidates based on the candidate ranking model, choose the top-K candidates, then use them to train the classifier. We adjust the parameters of the candidate ranking model according to the performance of the classifier, and then iterate until convergence. The model we proposed can enhance the performance, which can not only amend part of the segmentation results, and can generate some sentiment phrases. Experiment results on Chinese microblogging datasets in NLP&CC2013show the effectiveness of the model.

Keywords: Microblog; Sentiment Analysis; A Joint Model; Candidate Generation Model; Candidate Ranking Model

1 引言

面向微博的情感分析是当前 NLP 研究的一个重点与热点问题。随着网络技术的进步与推广，社交网络逐步流行起来，特别是微博（Microbolg）发展极为迅速。相对于传统媒体，微博具有着使用便捷、覆盖面广、实时性强、互动方便和传播速度非常快等优势[1-2]。截止 2014 年 12 月，中国微博用户规模为 2.49 亿，网民使用率为 38.4%[3]。由于微博的影响范围广，面向微博的情感分

*收稿日期：定稿日期：

基金项目：国家自然科学基金资助项目(61202193,61202304)；湖北省重点学科建设立项学科成果；中国博士后科学基金(2013M540593,2014T70722)；山东省语言资源开发与应用重点实验室开放基金资助。

作者简介：陈波（1976——），女，博士，研究方向：自然语言处理；姬东鸿（1967——），男，教授、博导，研究方向：自然语言处理；吕晨（1989——），男，博士研究生，研究方向：自然语言处理；柏云深(1991——)，男，硕士，主要研究方向为自然语言处理；魏小梅(1974——)，女，博士研究生，主要研究方向为自然语言处理。

析无论对个人或者社会团体都有着重要的研究意义，当前在电子商务、企业、政治事件、舆情监控等四个领域尤为突出。中国计算机学会(CCF)举办的自然语言处理与中文计算会议（NLP&CC），经常通过搜集、标注微博数据用于评测。

2 任务介绍及相关研究

情感分析的工作从粒度上分可以分为词级、句子级、篇章级；从语种来分，中文英文是两大重点领域；此外还分为主题相关与主题无关的情感分析。本文重点考虑主题无关的句子级别的情感分析。

目前主要有以下几类方法研究情感分析：

早期的**基于词典与规则**的分析方法，主要依靠情感词典与人工规则结合来判定极性，但是受到未登录词、中文词语的情感经常依赖语境、人工定义规则难以全面覆盖等种种原因影响，导致这种方法已经越来越少运用了。

有监督分类方法：Pang[4]提出将情感分析视为一个特殊的文本分类任务，利用机器学习训练情感分析分类器，利用标注语料让机器以某种监督学习的方式自动学习。目前常用的分类模型有朴素贝叶斯（Naive Bayes）[5]，最大熵（Maximum Entropy）[6]，k-近邻（k-Nearest Neighbour，KNN）和支持向量机（Support Vector Machine，SVM）。这种方法是目前最为常见的处理方法，实现难度不大，且随着特征的丰富与预处理的合理与完善，也能提升其准确率，性能上也不错。

无监督的分类方法：先找出基本的情感词作为种子词，然后计算待测文本中的情感短语和种子词之间的分值来判断倾向。如 Turney[7]提出 SO-PMI 算法，基于点互信息（Pointwise Mutual Information，PMI）来抽取文本中的关键词与种子词（如 excellent，poor）的相似度从而对文本情感倾向性进行判别。Zagibalov[8]、娄德成[9]针对 SO-PMI 算法进行改进。

深度学习：张铭[10]利用卷积神经网络算法来研究情感分析的任务。Richard[11]训练了 Recursive Neural Tensor Network 的递归神经网络模型。

联合模型：Tang[12]针对英文的 Twitter，提出一种分割与分类的联合框架（a joint segmentation and classification frameword, JSC），解决了类似“a great deal of”这一类的情感词包含问题。

上述方法都是在分词基础上执行的，然后以一种先分词后分类的串行的形式执行，因此存在缺陷，如果分词错误就会导致错误传播到情感分类，如例 1：

例 1：小王开心塞啊。

正确分词结果为：

{“小”，“王开”，“心塞”，“啊”}。

但是如果错误地分词为：

{“小王”，“开心”，“塞”，“啊”}

则很可能导致下一步情感分析的错误，即分词错误累积至情感分析阶段导致错误。再如例 2：

例 2：英雄难过美人关。

例 2 中，分词时必然不会将“难过美人关”作为一个整体，而是作为“难过”、“美人关”这两个词来进行分析，就很可能会导致分类的错误，这种情况下就需要划分为一个**情感单元**，也就是从情感短语要素的角度来解决情感分类的问题。

在上述国内外已有研究中，Tang[12]利用了一种分割的方法来处理英文的短语结构。我们将这种思路应用到中文里，但是中文与英文不同的是，首先会面临分词的问题。因此本文提出了分割与分类的联合模型，能够解决串行模型的分词错误传播到情感分类模型中，并且从情感短语要素的角度来解决情感分类的问题。

本文的大体思路是：

第一步，划分情感单元，挑选出候选。在利用现有分词结果，将一个句子以多种方式分割成不同短语组合，以短语作为情感分析的基本单元进行分类。对于每个句子而言，分割出的每一种短语组合方式称为一个候选。

第二步，训练分类器和训练排序模型。先挑选出部分候选训练分类器，再以分类的结果作为反馈，训练出一个排序模型，以被分类正确的候选排在错误的候选之前为排序依据，对每个句子的所有候选进行排序。重复上述过程，直到训练出一个最佳的排序模型，挑选出最合适的短语组合来训练分类器。即将以前分词，分类的无反馈的顺序模型修正为一个带有反馈的联合模型。

3 微博情感分析

3.1 任务描述

由于微博短文本的特点，常常只有一两个句子，篇章关系不是研究重点，本文便主要着眼于句子级的情感分析。即给出一些中文微博的句子，判断出其情感倾向，或者具体情感。如表 1 所示，表中[酷]等部分是微博的表情，显示在页面上是类似😎这样。由于现有方法大部分会或多或少借助情感词语来进行分析，所以示例中也标记出了情感词语[1]：

表 1 微博情感分析任务示例

序号	输入	情感极性	具体情感
1	天哪，简直是福音，不用做激光手术，也可以让大家摘掉眼镜，真的是**太棒了**，赶紧马住，分享给你身边需要的小伙伴吧**[酷]**	正面	高兴
2	都是一家人，得到的待遇差距咋就这么大呢[笑 cry]	负面	沮丧
3	即使是**最好的**朋友也应该保持一定距离，正如里尔克所说，友谊最高的境界是守护彼此的孤独。	中性	无
4	中国人看了那么多韩语中字的韩剧，**终于**也能让韩国人看我们的韩字中语的国产剧了。	正面	高兴
5	#唐嫣##何以笙箫默韩国#记得何以刚出来那会，有一条评论说唐嫣有这么一条特性，她演电视，一开始都被骂好惨，然后等放完，等好久后别人回味，会发现，噢，她演的其实还**不错**…现在又验证了这句话	正面	喜爱
6	呵呵……拒绝一次，然后关系急转直下	负面	厌恶
7	这孙子，真是好孙子	负面	厌恶

表 1 中 1、2 例是基本的分类情况，之后的例句则说明了一些常见的问题：

3 例的情感词是用来修饰“朋友”，对整句的极性没有影响；

4 例并没有明显的情感词语，但是整句却表达了一种欣慰、高兴的正面情感；

5 例中既有正面情感词也有负面情感词，这种情况实际是因为分句之间存在转折关系，其重点在后半句中；

6 例是褒义贬用的情况，“呵呵”这个本身正面的词语，现在已经越来越多用于贬义情况；

7 例则是反讽。

3.2 任务形式化

无论是情感极性判断问题还是细致情感分类问题，都可以形式化为一个文本分类问题。

分类的输入就是句子本身，输出则是给定的极性之一或者情感之一。这样，对于情感极性判断就可以形式化为一个二分类（不包括中性）或者三分类（包含中性）问题；而对于细致情感分类问题，则视为多分类问题了。

因此，对于句子级别的情感分析，其目标可以视为找到一个映射函数：

$$pol_i = f(s_i), s_i \in \{s_i | 1 \le i \le |T|\} \quad (1)$$

其中 T 是待分类句子的集合，s_i代表 T 中第 i 个句子，而pol_i代表这个句子的情感。pol的取值为给定的情感类别或者某一个具体的情感，如高兴，悲伤等。即对于情感类别判断问题，$pol \in \{Negative, Positive\}$。对于细致情感分类问题，$pol \in \{Happiness$，Fear，Disgust，Anger，…$\}$。

[1]加了下划线的词语为情感词，其中双划线表示正面极性，波浪线表示负面极性的词。

4 短语抽取与情感分析结合的联合模型

本文采用短语抽取与情感分析的联合模型来解决中文微博情感分析的任务。联合模型所联合的两个模型均可以归为最优化问题来解决：

对于候选排序模型可以视为一个回归分析，而最终利用对数线性模型转化为一个无约束的最优化问题；而对于情感分类模型，利用 SVM 求解时，视为了一个最大化间隔策略的有约束最优化问题。

4.1 整体框架

本文采用的方法基于以下假设将短语划分与情感分析联合起来：

假设一：短语划分的结果对于情感分析有着十分重要的影响。因为短语（将词语视为没有组合过的短语）是情感分类器的直接输入。正确的短语划分方式可以训练出效果最好的分类器。

假设二：短语划分的好坏可以利用分类器的效果，即能否正确的分类来评估。分类器分类效果越高，说明训练用的短语划分方式越正确。

基于以上假设，对于输入的某个句子，通过得到若干个不同方式组合出的短语构成的句子，称为候选（Candidate）。利用训练得到的候选排序模型（Candidate Ranking Model，CR），对所有候选，根据对情感分类的贡献度打分，令对分类有利的候选得到更高的分，然后对所有候选排序后选取 top-K。然后通过训练得到的情感分类模型（Sentiment Classification Model，SC）对于选取的 K 个结果分别进行分类预测，将 top-K 个分类结果按类别分组，然后求每一类的得分总和后，选取得分最高的类别作为句子类别。整体框架图详见本节图 1 与图 2.

4.1.1 模型训练

训练时，对于一个任务的输入的某个句子，先通过 CG 模型生成若干候选，然后利用初始化的候选排序模型 CR 从中挑出 top-K 中来初始化情感分类模型 SC。然后对所有候选进行预测，以预测结果是否正确作为输入，来重新训练 CR，然后重新打分排序后挑出 top-K 来更新 SC。

基于本节的假设，正确的短语划分训练出来的分类器后，利用其预测所有候选后，根据得分排序应该是稳定的，即满足预测正确的分会高于预测错误的。而反之，如果 CR 的参数不正确，根据训练出的分类器就是一个不准确的，也就会导致，正确的得分不一定会高于错误的。那么可以说，两次迭代间 CR 的参数变化越小，整个模型就越稳定，也就越接近于正确的 CR。

对于每次迭代，候选排序模型以分类正确与否作为输入，而分类模型则以排序模型的结果为训练样本。这样每一轮迭代，就是向着两个模型之间更协调的方向进行，也就是上一段所说的下一轮迭代比上一轮的 CR 参数变化更小的方向进行。

训练算法的伪代码如算法 1 所示：

算法 1：联合模型算法

输入：训练语料：TC = [s_i，pol_i^g],$1 \le \mathrm{i} \le |\mathrm{TC}|$

短语候选生成器：CG( )

排序特征抽取器：rfe( )

分类特征抽取器：cfe( )

输出：候选排序模型：CR

候选分类模型：SC

步骤：

[1]：对于 TC 中每一个句子s_i，利用短语候选生成器生成 n 个候选C_{ij}，$1 \le \mathrm{i} \le |\mathrm{TC}|, 1 \le \mathrm{j} \le \mathrm{n}$

[2]：利用排序特征抽取器抽取每个候选C_{ij}的排序特征，$\mathrm{rfe}(C_{ij})$

[3]：随机初始化候选排序模型CR^0

[4]：for r ← 1…最大迭代轮数 R do

[5]：　　利用上一轮的排序模型CR^{i-1}对每一个句子s_i选出 top-K 种候选C_{i*}，$1 \le \mathrm{i} \le |\mathrm{T}|$。

[6]：　　利用分类特征抽取器抽取C_{i*}的分类特征$\mathrm{cfe}(C_{i*}), 1 \le \mathrm{i} \le |\mathrm{T}|$来训练本轮候选分类模型，$SC^r$

[7]：　　利用SC^r预测每个候选C_{ij}的极性pol_{ij}

[8]：　　for i ← 1…|T| do

[9]：　　根据s_i的每个候选的极性pol_{ij}计算出s_i的极性pol_i

[10]：　　end for

[11]：利用每个句子的极性pol_i与每个候选的极性pol_{ij}输入至CR^{r-1}计算新的评估分数，并以此训练新的CR^r

[12]： for i ← 1…|T| do

[13]： 根据CR^r，通过$\mathrm{rfe}(C_{ij})$计算新的评估分数

[14]： end for

[15]：end for

[16]：SC ← SC^R

[17]：CR ← CR^R

模型预测流程如图 1 所示：

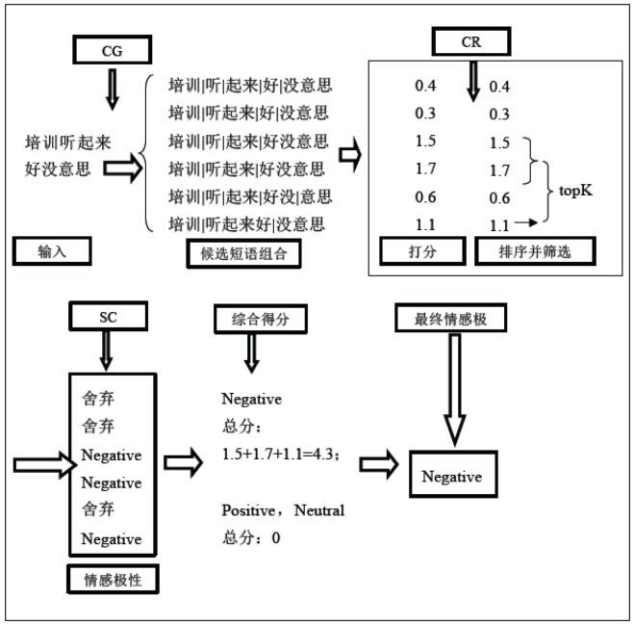


图 1 模型预测流程图

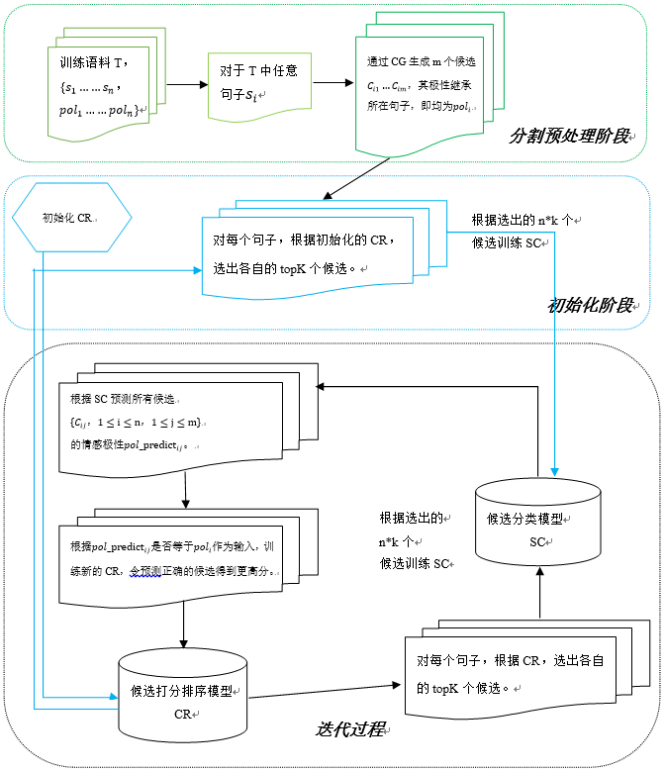


图 2 框架训练流程图

图 2 展示了迭代模型训练时的流程，迭代过程持续 n 次或者 CR 不再变化，其中 CR 随机初始化是默认策略，实验时发现 CR 以全 0 初始化，默认选择分词结果生成第一轮 SC 效果较好。

4.2 候选生成模型（Candidate Generation Model，CG）

在中文情感分析的时候，其情感单元有时候并不是单个词语，而是几个词语的组合，并且这个组合有时也并非是词法意义上的短语。但是这里生成候选的目的是为了找到对确定情感倾向有帮助的词的序列。所以这里暂且称为候选。以图 1 的句子为例，其生成了 6 种候选，如表 2 所示：

表 2 候选生成示例

句子	序号	候选内容
培训听起来好没意思	1	培训 | 听 | 起来 | 好 | 没意思
	2	培训 | 听起来 | 好 | 没意思
	3	培训 | 听 | 起来 | 好没意思
	4	培训 | 听起来 | 好没意思
	5	培训 | 听 | 起来 | 好没 | 意思
	6	培训 | 听起来好 | 没意思

由表 2 可以看到，生成的候选的结构有所差异（词语的个数、长度等都有不同），情感词比例也有所变化，因此，利用不同的候选进行训练，必然会影响到分类器的性能。

从对框架的简述中可以发现若第一轮迭代训练分类器的结果就很差，很有可能在迭代过程中偏离向错误的方向。因此这里的候选需要有着一定的准确率。

4.2.1 基于语言模型的短语组合

本文先利用 n 元模型与共现统计生成一个词汇表。然后利用 Beam Search 结合词汇表找出 n 种可能的短语组合。

这里利用 Mikolov 提出的方法[13]，利用一元模型(unigrams)和二元模型(bigrams)及共现频率，通过下述公式来生成一个基本短语表：

$$\mathrm{freq}(w_i, w_j) = \frac{\mathrm{freq}(w_i, w_j) - \delta}{\mathrm{freq}(w_i) * \mathrm{freq}(w_j)} \qquad (2)$$

其中δ是为了防止过多的低频词语出现影响最终的短语抽取。即防止短语中某个词频率$\mathrm{freq}(w_i)$过小导致该短语值增大。

在此基础上，对每一个句子，采用 Beam Search 的方法搜索出有限种的短语组合方式。

本文中，运用 Beam Search，从前往后扫描每个字，对于每个位置，若以其开始的文本片段在词汇表中存在就将其加入 beam。然后利用句中短语个数的倒数排列所有可能组合，并选出 top-N 种作为候选。即尽量选取句中短语总数更少的，这说明组合越充分。

4.2.2 基于 Ansj 的短语分割与组合

由于单纯使用语言模型在中文中对于词语的发现准确率并不高[2]，需采用另一种方式获得短语组合候选，即在 Ansj[3]分词的基础上进行切割与组合获得候选。

然后在其基础上通过定义两种操作，分割(split)与组合(merge)。以随机顺序进行这两种操作。而这两种操作依据语言模型，组成词的概率越低越可能被分割，两个词边界的两个字组成词概率越高越可能被组合。

4.3 候选排序模型（Candidate Ranking Model，CR）

候选排序模型是利用多种候选结果参与分类器训练，训练时必须挑出其中最为合适的 K 种候选进行训练，就需要将候选打分后排序。

打分的依据就是令排序在前面的获得更高的分，根据 4.1 节的假设二，可以得知这里的排序是希望将被正确分类的候选排在前面，优先选用。打分规则便可以转化为令分类正确的得分更高。即如下式，反之亦成立：

$$\begin{cases} \mathrm{pol}_i \text{正确} \\ \mathrm{pol}_j \text{错误} \end{cases} \Rightarrow \mathrm{score}(i) > \mathrm{score}(j) \qquad (3)$$

[2]此处召回率不需要很高，会通过组合获得长词语。

[3]Ansj 是用 Java 重写了中科院分词工具 ICTCLAS 的一个工具，其分词准确率据官网介绍可以达到 96%以上。

而为了评价该打分函数的好坏，则可以利用所有分类正确的候选的得分总和占所有候选得分总和的比重来判断。即正确的分值越重，说明这个评估函数性能越好。于是本文定义了一系列候选的分段特征用于打分，详见 4.5 节。在迭代模型中，通过一个评估函数，根据特征来计算出该候选的得分。然后利用对数似然函数设计损失函数。然后便可以将该问题转化为一个求损失函数最小值的无约束的最优化问题。

4.3.1 评估函数与损失函数

由于对候选的评分目的是为了让正确的分高于错误的，也就可以视为一个特殊的利用回归进行二分类的问题：高或者低。但是不同于一般情形，这里需要评估的是一个整体情况，同为正确的候选之间分数高低不需要关注，而是希望正确的整体能够得到更高的得分，这样能才能让更多的句子的 top-K 为正确的。

基于对数线性模型(log-linear model)，设计出一个评估函数，这里对于每个候选C_{ij}的分数，定义如下公式来计算：

$$\emptyset_{ij} = \exp(b + \sum_k sfe_{ijk} * w_k) \qquad (4)$$

其中$\emptyset_{ij}$代表第 i 个句子s_i的第 j 个候选C_{ij}的得分，k 代表该候选的第 k 个特征，sfe_{ijk}即为该特征的具体数值。其中w_k和 b 分别是特征的权重与偏移。也就是该排序模型所需要训练的参数。

根据评估函数给出的分数排序所有候选，并取每个句子的 top-K 种候选进行分类器的训练。然后利用分类器可以预测句子所有候选的极性。根据 4.1 节假设二，正确的顺序应该将预测正确的候选排在前面，而错误的排在后面。故而评估函数需要将极性预测正确的候选的分数最大化。

对数线性模型的损失函数一般根据其对数似然函数计算，因而有了下面的基于分类正确与否的损失函数如下：

$$\text{loss} = -\sum_{i=1}^{|T|} \log\left(\frac{\sum_{j\in H_i}\varphi_{ij}}{\sum_{j'\in A_i}\varphi_{ij'}}\right) + \lambda\|w\|_2^2 \qquad (5)$$

其中 T 是整个训练集，A_i表示句子s_i的所有候选的集合。H_i表示句子s_i的候选中命中的部分的集合。λ是参数向量 w 的 L2-正则化因子。

此时，当 loss 最小的评估函数，也就是最符合 4.1 节假设二的评估函数。于是将问题转化为求$\min_T loss$的无约束最优化问题。其中正则化是线性代数中的一个概念，是为了解决反问题的不适定性，这个理论作为一个纯数学问题这里不做讨论。用于损失函数是为了避免出现过拟合(over-fitting)。简单的来说，就是如果模型过于复杂且精确，训练数据中如果出现了个别错误或者说有部分数据规律与之后的预测数据不同，就会导致模型的预测能力的下降，也就是所谓的泛化能力不强，而加上正则化项后，可以提高模型的泛化能力。具体证明不在本文讨论范围内，实际运用中对于 BFGS 以及 LBFGS 一般采用 L2 正则化，也就是$\|w\|_2^2$。

对于这个模型，本文利用 LBGFS 方法来求解。

4.4 情感分类模型（Sentiment Classification Model，SC）

分类模型常用的有朴素贝叶斯（Naive Bayes），最大熵（Maximum Entropy）与支持向量机（SVM）。这些模型利用现有的标注数据，从句子中提取出特征，然后训练模型的参数后，利用模型来预测句子的情感。

根据以往工作可知，SVM 一般情况下可以取得更好的结果，因此本文直接采用 SVM 作为分类模型。

4.5 特征选取

4.5.1 分类特征选取

对于微博，其分类特征选择与预处理上，已有很多工作。比如提取标签(#**推荐好文**#)、表情提取、情感词等方式。由于除了情感词部分在提取情感单元后无法使用，其余部分均可直接迁移至以情感单元为输入的分类模型中。本文采用基本的词袋模型（Bag of Word，BoW）作为分类特征，进行对照实验。由于本文使用的是情感单元，所以对应的更换为 Bag of Unit。

4.5.2 排序特征选取

本文基于如下三个假设选取了 5 个特征作为排序打分时的特征，详见表 3。

假设一：中文的情感单元以词语为单位组成。

假设二：中文正确单元划分方式的字，词，单元的统计特征符合某种分布。

假设三：Ansj 分词结果与单元之间符合某种分布。

表 3 候选评估特征

序号	特征	特征描述
1	单元数	每个候选中被分割出的短语单元数
2	单元数与字数比例	每个候选中单元数除以候选中总字数
3	字数与单元数差值	每个候选中字数减去单元数
4	拆分次数	对标准分词结果的拆分操作次数
5	组合次数	对标准分词结果的组合操作次数

特征 1~3 用以描述一个句子中单元与字数的联合分布，而特征 4~5 用以描述与标准分词结果的偏离程度。利用上述特征结合模型评估一个候选的好坏。用以上五个特征，再加上词袋信息，共同构成排序用特征。

5 实验与评估

本文主要实验分为两组对照实验：一组用于验证对于候选生成模型的修改的有效性；另外一组包含 3 个实验，分别为传统的 SVM 分类，迭代模型与分布式遗传算法，用以验证 3 种模型的情感分析效果。

5.1 数据集处理

训练数据利用第二届自然语言处理与中文计算会议（NLP&CC 2013）技术评测中文微博情绪识别样例。由于本文针对句子级别处理，所以将数据集中的句子抽取出来，组成实验用数据集。实验所用情感采用数据集本身标注的 7 类(原有八种，去除 none)进行[4]。

表 4 数据集的句子 8 类情感类型分布[5]

emotion-type	数量	比例	去除 none 后
ANGER	718	5.42%	14.55%
DISGUST	1004	7.58%	20.34%
SADNESS	838	6.32%	16.98%
HAPPINESS	728	5.49%	14.75%
SURPRISE	310	2.34%	6.28%
LIKE	1223	9.23%	24.78%
NONE	8314	62.75%	——
FEAR	115	0.87%	2.33%
总数	13250	——	4936 条

由于细致情感分类的效果过低，为了更好的证明模型的有效性，将 emotion-type 中 ANGER，DISGUST，SADNESS，FEAR 作为 Negative；HAPPINESS，LIKE 作为 Positive，SURPRISE，None 作为 Neutral。处理后情感类型分布如表 5 所示：

表 5 数据集的句子正负面分布

emotion-type	数量	比例	去除 none 后
NEUTRAL	8624	65.09%	——
NEGATIVE	2675	20.19%	57.83%
POSITIVE	1951	14.72%	42.17%
总数	13250	——	4626 条

5.2 实验设置

本文采用将 5 折交叉验证（K-fold cross-validation）的方式来评估模型的效果。

[4]由于 none 的比例过大，超过 60%，排除此类进行测试。

[5]由于各类别数据分布亦不平均，特别是七分类情况，数量最多的类和最少的类数量之比接近于 10:1。于是这里从原始数据中随机选取一个小量（总数约为 2000 条）的各类别均衡的数据作为最终实验所用数据。

评估标准方面，本文采用 3 个标准对效果进行评估：对于每组交叉测试，统计结果中每一类，即每一个情感极性的准确率，召回率与 F1-Measure。计算所有类别总的准确率，召回率与 F1-Measure。利用宏平均来评估。计算所有类别总的微平均值，作为总体微平均准确率。此外，参照大多数工作，以 F1 的宏平均值 macro-F1 作为最主要评估指标。

5.3 候选生成模型对照实验

引言中描述了两种候选生成的模型。其中一种相当于直接将中文每个字视为英文一个单词，采用类似处理英文的方式来处理，另外一种针对中文特性基于中文分词结果进行调整得到。这里需要做一个对照实验来验证第二种修改后的方式是否能得到比第一种更优的结果。本实验针对两种不同候选生成方式，分别调整 cost 参数。n 元模型（cost 为 0.5）与基于分词的候选生成模块的实验结果如表 6 所示。

表 6　N 元模型与基于分词的候选生成模型实验结果对比

极性	准确率		召回率		F1-Measure	
	N 元模型	基于分词	N 元模型	候选生成模型	N 元模型	候选生成模型
HAPPINESS	62.621%	23.488%	25.656%	53.979%	36.399%	32.733%
FEAR	40.278%	58.366%	9.359%	18.345%	15.189%	27.916%
SURPRISE	23.397%	24.334%	11.874%	35.185%	15.753%	28.770%
ANGER	31.012%	28.777%	7.291%	8.747%	11.806%	13.417%
SADNESS	15.738%	22.337%	72.952%	15.392%	25.890%	18.225%
LIKE	32.540%	26.799%	6.196%	11.942%	10.410%	16.521%
DISGUST	27.778%	59.444%	5.589%	7.416%	9.306%	13.187%
宏平均值	33.338%	34.792%	19.845%	21.572%	24.880%	26.632%

可以看到，利用中文分词为基础进行拆分与组合，无论在最终 F1-Macro 上还是稳定性上均高于直接利用 n 元模型的结果。

5.4 情感分析性能对比试验

对于情感分析，由于情感极性（即二分类任务）更为常见，此后实验将数据进行了处理转化为情感极性的判别。本文直接使用 SVM 对所选的分类特征进行分类作为 baseline。训练特征选用词袋。在选用同样核函数，并选用同样的特征的情况下，用单纯使用训练分类器。

参数说明：本文中除了 Libsvm 参数以外，其余参数为每个句子生成候选数：10，每个句子选取 top-K 候选数为 3，LBFGS 的正则化系数为 0.003。实验结果数据见表 7：

表 7 情感分析性能对比试验结果

极性	准确率		召回率		F1-measure	
	Baseline	联合模型	Baseline	联合模型	Baseline	联合模型
POSITIVE	71.06%	71.91%	73.70%	76.70%	72.36%	74.23%
NEGATIVE	72.79%	75.12%	70.00%	70.00%	71.37%	72.47%
宏平均值	71.93%	73.51%	71.85%	73.35%	71.89%	73.43%

由表 7 可以看到，联合模型能获得比 baseline 更好性能，证明了本文改进的有效性。

5.5 错误分析

将所有分类错误的句子候选导出，进行分析，可以得到以下几类错误：其中示例部分是导出的结果，第一列是该句的正确情感，第二列为预测的情感，第三列为候选的划分方式，详见表 8。

表 8　错误分析

序号	描述	示例	错误分析
1	句子过短，无法形成有效分割	FEAR SADNESS　{拜托}	句子过短，只能分出一个词，且训练语料中没有“拜托”这个词

2	分割后不能产生有利结果，即所有候选都是错误极性	SURPRISE SADNESS {他, 说了} SURPRISE SADNESS {他说了}	分割出的短语对于情感分析没有作用
3	排序错误	SURPRISE SURPRISE{我, 天啊, 时间, 转瞬即逝} SURPRISE SADNESS {我天, 啊, 时间, 转瞬即逝}	正确的分词结果应该是“天啊”作为一个词，但是第二句却将“我天”作为了一个词，但是此时这两个句子的几个排序用特征的值基本相同。所以排序所用参数过少，可以看到无法充分体现排序依据。

6 结论及未来工作

本文提出了一种创新的基于分割和分类的联合模型来实现微博情感分析，采用了情感单元分割与情感分析结合的模型来处理中文领域的微博句子级别情感分析。该方法解决了先前研究的串行模式中因为分词错误传播到情感分类的问题。根据候选生成模型为句子生成多个候选，基于候选排序模型对该句子进行打分，选择分数最高的前 K 个候选，放入情感分类模型里面来训练分类器。然后根据分类器的性能高低来调整候选排序模型的参数权重。本文所提的算法不仅能修正部分分词结果，而且可以生成一些情感描述短语，使得情感分析的性能获得提升。在 NLP&CC2013 测评数据集的实验结果表明，F1-measure 值方面，联合模型的方法比 baseline 在 POSITIVE、NEGATIVE 和宏平均值上分别提高了 1.87%、1.1%、1.54%，证实了本文所提算法的有效性。

参考文献:

[1] 周胜臣, 瞿文婷, 石英子等. 中文微博情感分析研究综述[J]. 计算机应用与软件, 2013(3): 161-164.

[2] 吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014(4): 735-752.

[3] 三川. CNNIC 发布第 35 次《中国互联网络发展状况统计报告》[J]. 中国远程教育, 2015(2): 31.

[4] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques[C]//Proceedings of EMNLP. 2002:79--86.

[5] Tan S, Cheng X, Wang Y, et al. Adapting naive Bayes to domain adaptation for sentiment analysis[M]//Advances in Information Retrieval. Springer Berlin Heidelberg, 2009: 337-349.

[6] Batista F, Ribeiro R. Sentiment Analysis and Topic Classification based on Binary Maximum Entropy Classifiers[J]. Procesamiento Del Lenguaje Natural, 2012, 50.

[7] Turney P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[J]. Proceedings of ACL, 2002:417--424.

[8] Zagibalov T, Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese text[J]. Coling ’08 Proceedings of International Conference on Computational Linguistics, 2008:1073-1080.

[9] 娄德成, 姚天昉. 汉语句子语义极性分析和观点抽取方法的研究[J]. 计算机应用, 2006(11): 2622-2625.

[10] 张铭. 基于 CRFs 的微博评论情感分类的研究[D]. 东北师范大学, 2014.

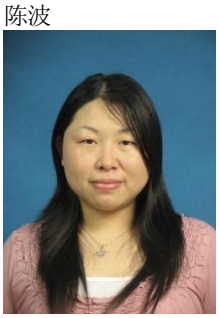
[11] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment Treebank[C]//Proceedings of EMNLP. 2013, 1631-1642.

[12] Tang D, Wei F, Qin B, et al. A Joint Segmentation and Classification Framework for Sentiment Analysis[C]//Proceedings of EMNLP. 2014:477-487.

[13] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. 2013: 3111-3119.

●陈波，联系方式：湖北省襄阳市隆中路 296 号，湖北文理学院文学院， 邮编 441053
手机：18995633440
Email：cb9928@gmail.com
● 通讯作者：
姬东鸿，男，武汉大学计算机学院 教授、博导，研究方向：自然语言处理。
联系方式：湖北省武汉市武昌区武汉大学计算机学院，邮编 430072
手机：15927260117
Email：dhji@whu.edu.cn

陈波



姬东鸿

吕晨