文章编号：1674-8085(2019)04-0046-06

# 基于 LDA 及语义相似度的商品评论情感分类研究

*曾 寰[1]，龙小建[2]，刘 华[1]

（1. 井冈山大学电子与信息工程学院，江西，吉安 343009；2. 井冈山大学继续教育与培训学院，江西，吉安 343009）

**摘 要**：提出一种结合 LDA 及语义相似度的商品评论情感分类方法。该方法首先使用 LDA 对商品语料库建模，获取文档-主题矩阵；人工选择 *k* 对褒义词、贬义词，基于 HowNet 语义相似度计算主题(评价对象+观点内容)与各个褒义词和贬义词的相似度，达到对观点词极性判断，计算文本观点词情感极性的加权和作为文本的情感极性。实验表明，与基于向量空间的 SVM 分类方法相比，该情感分类方法在分类指标上表现更好。

**关键词**：情感分类；LDA；语义相似度；观点词

**中图分类号**：TP391 **文献标识码**：A DOI:10.3969/j.issn.1674-8085.2019.04.009

# SENTIMENT CLASSIFICATION RESEARCH IN PRODUCT REVIEWS BASED ON LDA AND SEMANTIC SIMILARITY

*ZENG Huan[1]，LONG Xiao-jian[2]，LIU Hua[1]

(1. School of Electronics and Information engineering; Jinggangshan University, Ji'an, Jiangxi 343009, China;

2. School of Continuing Education, Jinggangshan University, Ji'an, Jiangxi 343009, China)

**Abstract:** A product reviews classification method based on LDA and semantic similarity is proposed, The method firstly uses LDA to model the reviews corpus and get the review-topic matrix. Artificially, it selects k positive and negative words and use "HowNet" semantic similarity to calculate the similarity of opinion content words with each positive and negative words to get the polarity of opinion word. The weighted polarity sum of words will be the polarity of the review. The experiment result shows, compared with SVM classification method based on vector space text representation, the proposed method works better in classification index.

**Key words:** sentiment classification; LDA;semantic similarity; opinion word

近年来，随着 web3.0 的快速发展，电商网站及影视音乐等娱乐推荐网站给用户带来了很大便利，同时让很多用户留下了大量的评论信息，这些信息汇聚了用户对评价对象（产品与服务、新闻、图书、影视音乐等）的偏好信息。这些评论可以辅助消费者做出购物选择，运营商通过这些评论信息能够及时得到用户反馈，从而提高产品质量。但是，当网络评论信息过多时，用户很难从大量评论中发现有效信息。因此，如何自动地分析评论并从评论中抽取出有效信息成为亟待解决的问题。评论(opinion)是评论者对评论对象持有的态度和观点的阐述，网络评论的特点是数据海量、复杂多样、可用价值高与非结构化。依据评论对象的不同，最主要的几类可以划分成：影视音乐评论，产品与服务评论，图书评论，政治言论及新闻评论。影视音乐、图书评论由感而发，随意性强，形式松散，情感倾向不明显，评论对象相对容易判断；政治言论和新闻评论形式规范，情感倾向明显；产品和服务评论内容简短，评论对象和情感倾向都较为明显[1]。Kim 和 Hovy 认为评论(opinion)可以定义成一个四元组：评

收稿日期：2019-01-18；修改日期：2019-05-10
基金项目：吉安市社会科学研究项目（18GH113）
作者简介：*曾 寰(1990-)，男，江西吉安人，助理实验师，硕士，主要从事数据挖掘研究(E-mail:584251395@qq.com);
龙小建(1971-)，男，江西吉安人，工程师，主要从事软件开发、数据库研究(E-mail:113789072@qq.com);
刘 华(1981-)，男，江西吉安人，实验师，硕士，主要从事大数据研究(E-mail: 19249658@qq.com).

价对象(opinion target)，持有者 (Holder)，观点内容(Claim)，情感倾向(Sentiment)]，它们之间的关系为：评论的持有者针对某一事物对象发表了具有一定情感倾向的观点内容[2]。

文本所采取的分类步骤为：首先将获取的语料库进行分词，根据停用词表去除停用词，之后将分词转化成到向量空间上，然后选取合适的特征和分类模型进行训练，最后将得到的模型对新的文本进行预测。基于向量空间的分类方法并没有涉及概率主题模型[3]，TF-IDF(Term Freqency-Inverse Document Frequency)方法基于统计学计算策略，综合考虑了特征词在文档中的分布和特征词在整个语料中的分布，但没有将相关词汇的语义融合进来。基于信息增益(Information Gain)的方法将信息熵与基于特征的条件熵考虑进来，在文本分类问题中效果好[4]，但是该方法用于分类中只是考虑了特征在整个语料库中的影响，并没有考虑在类别上的影响程度。

目前，针对商品评论的评论对象和评论内容的提取中，Qiu et al[5]根据句子间的语法依赖关系方法泛化为一种半监督的双向传播方法。首先利用少量的观点词(种子)，根据与评价对象间的依赖关系抽取评价对象，抽取到的评价对象又反向抽取观点词和评价对象，一直持续到没有新的观点词和评价对象抽取出来为止。Zhang 等人[6]引入“part-whole”和“no”模式并根据评论对象的重要程度进行排序来改进双向传播算法。总之，利用评价对象和观点词之间语法依赖关系进行评价对象和观点词的抽取可以取得较好的性能。但基于这种关系的提取模板不能泛化到所有情况，同时这种抽取模板得到的结果可能包含错误，进而导致评价对象抽取错误。

对于评论的情感分类，我们知道评论文本的长短不一，句式不定，形式多样，根据评论的粗粒度不同，分为词语级、语句级和文档级。语句级和文档集情感判定都是基于词语级的情感判定。Hu 和 Liu[7]通过人工选取一些褒义词和贬义词做为种子集，并从 WordNet 通过同义词和反义词扩展生成情感字典。词语的情感倾向还可以通过分析该词语在词典或其他资源中的注释来做出判断[8]。词语的情感除了利用情感极性字典来判断情感倾向外，还可以通过分析相关语料来获取词语情感倾向。比如通过语料中的连接词来获取相同的极性（例如连接词“又”、“和”、“并且”可以获取相同极性）或相反极性（例如连接词“但是”）[9]；通过在语料中统计词语的共现来判定词的情感倾向，与词语出现较多的为褒义词，则该词语为褒义词，否则为贬义词[10]。对于语句级、文档级的情感倾向判定，Hu 和 Liu[7]在提取和判断完词语的情感倾向后，根据句子中占优势的情感词汇的语义倾向来判断该句子的情感倾向；Wiebe 等人[11-12]通过将句子转换成二元特征构建朴素贝叶斯分类器来对句子的情感倾向进行分类。唐慧丰等[13]使用 n-gram、动词、副词、形容词、名词作为文本特征，分别使用信息增益、互信息、文档频率和 CHI 统计量作为选择特征的方法，分别使用 KNN(K 最近邻)、WinNow、朴素贝叶斯、支持向量机、中心向量法作为文本分类方法，进行文档情感倾向判定。实验结果表明，使用 BiGrams 特征表示法，信息增益作为特征选择法，支持向量机作为分类方法，在足够大的训练集和适量特征情况下，分类效果优于其他分类算法。

对于商品评论的情感分类，本文提出一种结合 LDA 及语义相似度的商品评论情感分类方法。该方法首先使用 LDA 对商品语料库建模，获取文档-主题矩阵；人工选择 k 对褒义词、贬义词，基于 HowNet 语义相似度计算观点词与各个褒义词和贬义词的相似度，达到对观点词极性判断，计算文本观点词情感极性的加权和作为文本的情感极性。

## 1 LDA 模型

LDA(Latent Dirichlet allocation)是用在类似文本离散型数据的三层贝叶斯生成概率模型，它在语料库建模，表示成随机混合的隐含主题(topic)的概率分布；主题又表示成词汇的概率分布。LDA 基于词袋模型(bag-of-words)，它不考虑语料库中的文档集之间以及文档中的词汇集之间的顺序。LDA 生成文本数学建模过程如下：

首先定义符号含义：

1) 词是语料数据最基本的单元，词库中包含的词汇量是 V，那么该词可以表示成一个 V 维向量，这样第 $v$ 个词出现时，它表示成向量 $w$ 第 $v$ 个分量 $w^v=1$，其他分量 $w^u=0$，$(\mathrm{v}\neq\mathrm{u})$。

2) 文档是由 $N$ 个词组成的序列，表示为

$d=(w_1,w_2,\cdots,w_N)$，其中 $w_n$ 是序列中的第 $n$ 个词。

3) 文档集是 M 个文档组成的集合，表示成 $D=(d_1,d_2,\cdots,d_M)$

生成过程：

1) 选择 $N\sim Possion(\xi)$ 和 $\theta\sim Dir(\alpha)$

2) 在文档 $d$ 中生成第 $n$ 个词 $w_n$:

(a) 依据多项式分布 $z_n\sim Multinomial(\theta)$ 抽样得到 $w_n$ 所属的主题 $z_n$；

(b) 依据概率 $p(w_n|z_n)$ 抽样得到具体的词 $w_n$。

给定参数 $\alpha$ 和 $\beta$，LDA 生成文档 $d$，$N$ 个主题 $Z$，$N$ 个词汇 $w$ 的联合概率具体表示为公式(1)所示：

$$p(d,z,w|\alpha,\beta)=p(d|\alpha)\prod_{n=1}^{N}p(z_n|d)p(w_n|z_n,\beta) \quad (1)$$

通过期望最大化算法 EM 求最大似然函数公式(2)估计 $\alpha,\beta$ 的参数值，从而确定 LDA 模型。

$$l(\alpha,\beta)=\sum_{i=1}^{M}\log p(d_i|\alpha,\beta) \quad (2)$$

其中文档 $d$ 发生的条件概率分布如公式(3)所示：

$$p(d|\alpha,\beta)=\frac{\Gamma(\sum_i\alpha_i)}{\prod_i\Gamma(\alpha_i)}\int\left(\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j}\right)\mathrm{d}\theta \quad (3)$$

上式无法直接求出 $\theta$ 和 $\beta$，需要采用 Gibbs 抽样。

## 2 HowNet 词语相似度

知网(HowNet)是以英语和汉语词汇为描述对象，构建概念之间及概念与其属性之间的知识库。该知识库包含了丰富的词汇语义知识及世界知识，不同于其他语义词典，它使用“义原”来对概念进行描述，知网采用了 1500 个义原用于描述词汇的语义特征，词性及与其他词汇之间的关系。因此知网对词汇相似度的计算最后归结于用义原相似度的计算。所有义原由其从属关系构成一个树状层次结构体系[14]。因此据此根据层次路径长度 $d$，可以计算两个义原之间的语义距离如公式(4)所示。

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \quad (4)$$

其中 $\alpha$ 是一个可以调控的变量。

假设对于两个词语 Word1 和 Word2，如果 Word1 包含 $n$ 个义原：$(P_{11},P_{12},\cdots,P_{1n})$，Word2 包含 $m$ 个义原 $(P_{21},P_{22},\cdots,P_{2m})$，那么 Word1 和 Word2 的相似度为各个义原相似度的最大值如公式(5)所示。

$$Sim(Word1,Word2)=\max_{i=1\ldots n,j=1\ldots m}\left(Sim(P_{1i},P_{2j})\right) \quad (5)$$

## 3 研究方法

基于 LDA 和语义相似度的商品评论的分类主要包括文本语料预处理、使用 LDA 对语料建模获取主题(评价对象+评价内容)、确定最优主题、情感极性（正向、负向、中立）分类。图 1 为商品评论分类过程。

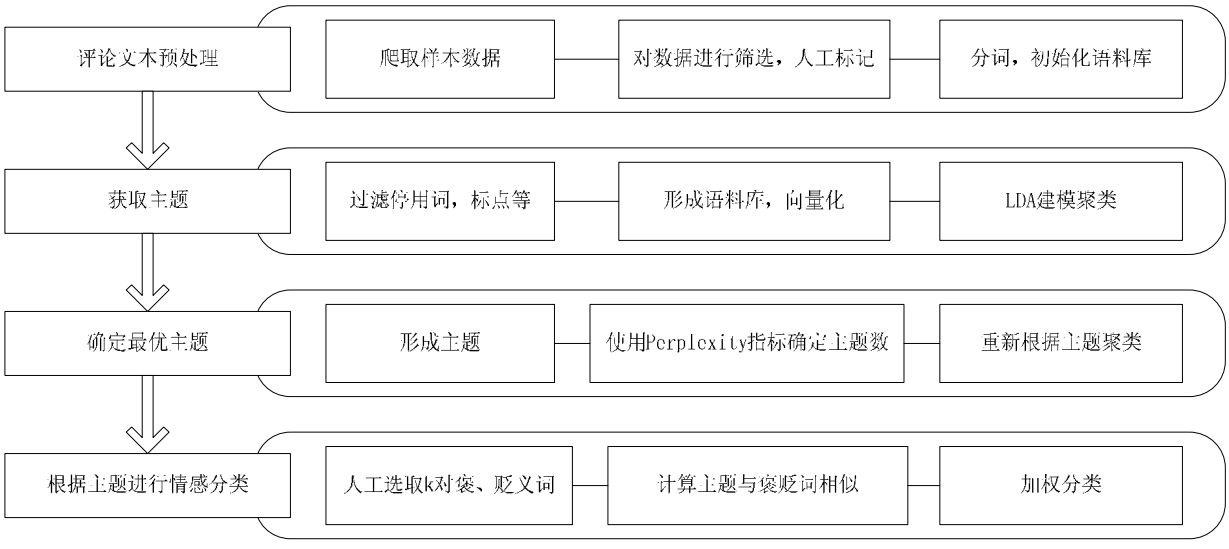


图 1 基于 LDA 和语义相似度的商品评论分类流程图

Fig.1 Flow chart of product reviews classification method based on LDA and semantic similarity

### 3.1 文本数据预处理

商品评论文本内容简短，同时包含表情符号，因此在预处理过程需要导入表情转换表，预处理过程包含以下步骤。

1) 用 Python 爬取收集关于商品相关评论文本；

2) 对评论文本进行人工标记和筛选；

3) 导入表情转换表，停用词表；

4) 用分词工具 THULAC[15]对评论文本进行分词，词性标注；

5) 对步骤 4 获取的语料过滤停用词、标点符号，同时根据表情转换表转换表情；

6) 初始化语料库。

文本预处理过程与后期运用 LDA 主题模型主题聚类有很大关系，因此分词、语料过滤有极大作用。

### 3.2 确定主题数

文本预处理获取商品评论的语料库，使用 LDA 对其建模，采用 Gibbs 抽样确定生成的 LDA 模型参数。虽然建立好了 LDA 模型，但主题数量值无法由模型确定，而主题数多少对抽取的主题影响很大。当主题数量过多时，会产生很多不具明显分类语义信息的主题；当主题数量过少时，会产生比较粗粒度的主题，这样对分类影响也很大。因此，如何科学地确定主题数量非常重要。本文采用 perplexity(混乱度)来确定最优主题数量值[16]。

perplexity(混乱度)在对语料建模中特别常用，它关于测试语料概率单调递减，在代数上等价于所有词概率的几何平均值倒数。perplexity(混乱度)主要衡量模型的预测能力，值越小，预测力越强，同时推广性也越高。计算如公式(6)，其中 $D_{test}$ 为测试语料数，$w_d$ 为文本 $d$ 词汇序列，$N_d$ 为文档 $d$ 的词汇数量。

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \quad (6)$$

### 3.3 文本情感判定

通过如上步骤获取到评论文本主题后，接着对文本情感判定，评论文本中主要由评价对象。评论内容构成。使用 LDA 模型获取的主题中包含了评价对象和包含了情感极性的观点词。本文通过人工选择 k 对褒义词和贬义词，结合 HowNet 语义相似度，最终判定主题的情感极性，计算如公式(7)。其中 $key\text{-}p_i$ 为褒义词，$key\text{-}n_i$ 为贬义词。

$$Oritation(w) = \sum_{i}^{k} Sim(key - p_i, w) - \sum_{i}^{k} Sim(key - n_i, w) \quad (7)$$

由以上公式可以看出，当主题词 $w$ 的极性值大于 0 时，为正向；当值小于 0 时，为负向；当为 0 时，为中立。

对于评论文本情感极性通过计算文本包含的主题情感词加权和来计算，如公式(8)。

$$Oritation(Opinion) = \sum_{i=1}^{n} Oritation(Word_i) \quad (8)$$

### 3.4 分类指标

文本的分类性能评价主要为 F-measure 值，该指标综合度量了模型的精准率(Precision)和召回率(Recall)。混淆矩阵(confusion matrix)是分类问题常用的评价方法。本文情感分类涉及三类即：中立(0)、正向(1)、负向(2)，因此三分类的混淆矩阵如表 1 所示。

**表 1 三分类混淆矩阵**

Table1 Confusion matrix of Three-classification

| 混淆矩阵 | | 预测值(Predict) | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| 真值(Real) | 0 | a | b | c |
| | 1 | d | e | f |
| | 2 | g | h | i |

为了衡量模型性能，本文使用微平均精准率(Micro_Precision)和微平均召回率(Micro_Recall)和微平均 F-指标(Micro_F_measure)来评估，计算如公式(9)、(10)、(11)所示。

$$Micro\_Precision = \frac{1}{3}\left(\frac{a}{a+d+g} + \frac{e}{b+e+h} + \frac{i}{c+f+i}\right) \quad (9)$$

$$Micro\_Recall = \frac{1}{3}\left(\frac{a}{a+b+c} + \frac{e}{d+e+f} + \frac{i}{g+h+i}\right) \quad (10)$$

$$Micro\_F\_measure = \frac{2\times Micro\_Precision\times Micro\_Recall}{Micro\_Precision + Micro\_Recall} \quad (11)$$

# 4 实验及结果分析

## 4.1 实验工具使用及实验过程

本实验使用 Python 在淘宝上爬取某手机相关商品评论文本 4435 条，实验目的是通过使用 LDA 主题模型对这些文本聚类提取主题（评价对象+评论内容），之后对这些文本进行情感分类。具体实验过程如下。

1) 使用 Python 爬取某手机商品评论 4435 条，导入表情转换表、停用词表，人工标记情感极性；

2) 使用 THULAC[15]对文本进行分词、词性标注；

3) 使用表情转换表、停用词表转换文本表情，过滤停用词、标点等操作；

4) 使用 gensim[17]将文本转化为向量，并使用该工具中包含 LDA 主题模型工具对文本向量建模，获取主题；

5) 对获取的文本根据步骤 4 获取的主题，人工选取 k 对褒、贬义词，结合 HowNet 语义相似度，对文本进行情感分类。

## 4.2 LDA 主题数量值设置

使用 LDA 主题模型建模过程中，主题数量的最优值采用混乱度(Perplexity)函数来确定，采用 Gibbs 抽样，抽样迭代参数值设为 1000。其中主题数量值依次为 5、10、20、40、80、160、200，直到 300，通过设置不同的主题对 Perplexity 指标分析，获取最小 Perplexity 的最优主题数。

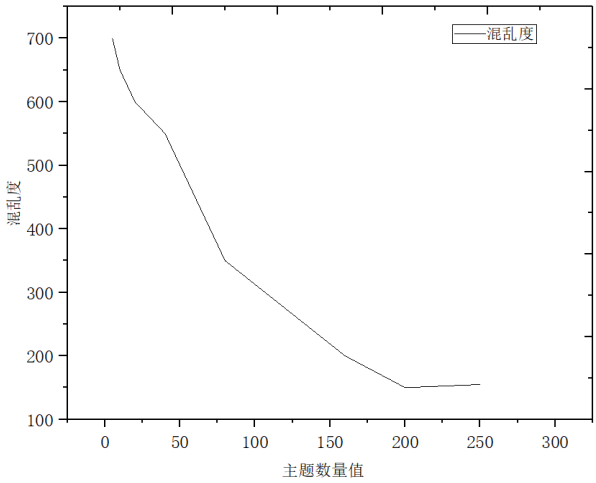


图 2 LDA 主题模型混乱度随主题数值变化趋势

Fig.2 The trend of perplexity of LDA model varies with the number of topics

由图 2 可知，当设置主题数值为 200 时，训练得到的 LDA 主题模型的 Perplexity 最低，之后趋于平稳。

经 LDA 主题模型建模得到 200 个主题及其分布情况。为了展示建模效果，这里只展示前 4 个主题，每个主题只显示前 9 个词汇及其分布情况，如表 2 所示。

**表 2 主题挖掘结果**

Table 2 Result of topic mining

| 主题 1 | | 主题 2 | | 主题 3 | | 主题 4 | |
|---|---|---|---|---|---|---|---|
| 好 | 0.10213509 | 好用 | 0.2721742 | 效果 | 0.05532863 | 价 | 0.10538627 |
| 送 | 0.10009696 | 好 | 0.0535067 | 错 | 0.053619135 | 喜欢 | 0.08745869 |
| 开心 | 0.09410068 | 错 | 0.04176 | 给 | 0.03158043 | 比 | 0.08306422 |
| 能 | 0.085789956 | 是 | 0.037092 | 用 | 0.03132137 | 买 | 0.074961804 |
| 钢化膜 | 0.06483134 | 手机 | 0.0359714 | 满意 | 0.02732855 | 手机 | 0.06861099 |
| 价 | 0.06329198 | 同事 | 0.0310096 | 手机 | 0.026769832 | 推荐 | 0.058247488 |
| 软件 | 0.04193385 | 反应 | 0.0308515 | 大 | 0.024761377 | 朋友 | 0.047342844 |
| 用 | 0.0410363 | 买 | 0.0301933 | 力 | 0.024263378 | 最高 | 0.042465713 |
| 老爸 | 0.03686542 | 小米 | 0.0299974 | 强大 | 0.023631848 | 高 | 0.033502903 |

## 4.3 情感分类实验结果分析

本文提出的基于 LDA 主题模型、结合 HowNet 语义相似度的商品评论文本情感极性分类方法，实验过程为，对爬取的数据进行文本预处理、将语料库分为 80%的训练样本，20%的测试样本；运用 LDA 方法对训练样本语料库建模，通过在测试检测测试样本 Perplexity 指标确定最优主题数，并提取出各主题；对语料库中的主题情感极性判断采用人工标记 *k*(本文选取 10)对褒、贬义词并使用相似度加权差来最终确定主题情感极性，评论文本的最终情感极性由整个文本主题词汇的文本情感极性加权和确定，大于 0 为正向，小于 0 为负向，等于 0 为中立。

参照实验，将获取的语料库转换为向量空间表示模型，运用 TF-IDF(Term Frequency-Inverse -Document Frequency)计算特征值权重，最后采用 SVM 训练进行分类。

本文方法与参照实验方法在各项分类评估指

标上的性能见图 3。

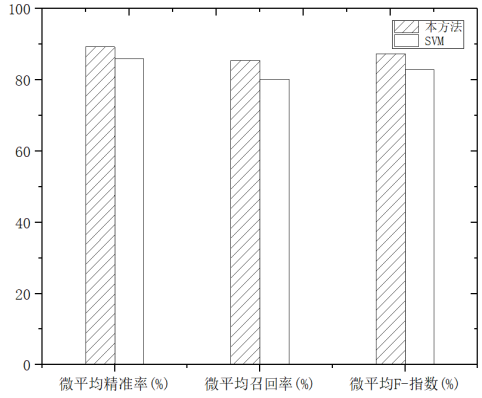


图 3 本文方法与 SVM 方法的实验结果

Fig.3 Experimental result of method proposed and method based on SVM

由实验结果可知，本文提出的方法与基于向量空间、使用 TF-IDF 的 SVM 方法相比，在分类评估指标上要更好。因此在商品评论情感极性分类上，本文方法表现更好。

## 5 结语

本文方法主要通过 LDA 主题模型结合 HowNet 语义相似度来对商品评论进行情感极性分类。将 LDA 主题模型来对商品评论文本进行主题聚类，获取商品评论的评论对象和包含情感极性的观点词，接着人工选取 K 对褒、贬义词，结合 HowNet 语义相似度，计算主题与褒、贬义词的相似度，来对主题词进行情感极性判断，最后对评论文本的情感进行分类，通过对主题情感极性加权差来获取文本极性。在与参照算法基于向量空间的文本表示 SVM 分类算法比较，本文算法在性能指标上表现更优。本文在情感分类时，在选取褒、贬义词的时，数量和选取的词对评论文本极性判断很重要，当 k 值选择过大时，增加了相似度的计算时间。若褒贬词选择过少，由于相似度的计算方法原因，可能无法判断主题的极性。下一步的工作将围绕如何更好选择褒贬义词来优化情感分类算法。

**参考文献：**

[1] 韩忠明,李梦琪,刘雯,等. 网络评论方面级观点挖掘方法研究综述[J]. Journal of Software, 2018, 29(2).

[2] Kim S M, Hovy E. Determining the sentiment of opinions[C]//Proceedings of the 20th international

[3] 廖列法,勒孚刚,朱亚兰. LDA 模型在专利文本分类中的应用[J]. 现代情报, 2017, 37(3): 35-39.

[4] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//Icml. 1997, 97: 412-420.

[5] Qiu G, Liu B, Bu J, et al. Opinion word expansion and target extraction through double propagation[J]. Computational linguistics, 2011, 37(1): 9-27.

[6] Zhang L, Liu B, Lim S H, et al. Extracting and ranking product features in opinion documents[C]//Proceedings of the 23rd international conference on computational linguistics: Posters. Association for Computational Linguistics, 2010: 1462-1470.

[7] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.

[8] Esuli A, Sebastiani F. Determining the semantic orientation of terms through gloss classification[C]//Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005: 617-624.

[9] Kanayama H, Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis[C]//Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006: 355-363.

[10] Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4): 315-346.

[11] Wiebe J. Learning subjective adjectives from corpora[J]. Aaai/iaai, 2000, 20(0): 0.

[12] Wiebe J, Wilson T, Bruce R, et al. Learning subjective language[J]. Computational linguistics, 2004, 30(3): 277-308.

[13] 唐慧丰,谭松波,程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6): 88-94.

[14] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学, 2002, 7(2): 59-76.

[15] THULAC [EB/OL] . http://thulac.thunlp.org/.

[16] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[17] gensim[EB/OL] . https://radimrehurek.com/gensim/.