基于主题情感句的汉语评论文倾向性分析

杨江 侯敏 王宁

中国传媒大学 北京 100024

E-mail: yangjiang@cuc.edu.cn;houminxx@263.net; ningzi72@163.com

摘 要：提出一种基于主题情感句的汉语评论文倾向性分析方法。根据评论文的特点，采用基于词语 n-gram 匹配的方法识别主题，通过对比与主题的语义相似度和进行主客观分类抽取出候选主题情感句，计算其中相似度最高的若干个句子的情感倾向，将其平均值作为评论文的整体倾向。基于主题情感句的评论文倾向性分析方法避免了进行篇章结构分析，排除了与主题无关的主客观信息。实验结果表明，该方法准确率较高，切实可行。

关键词：主题情感句；评论文；倾向性分析；情感

Recognizing Sentiment Polarity in Chinese Reviews Based on Topic Sentiment Sentences

Yang Jiang, Hou Min, Wang Ning

Communication University of China, Beijing 100024

E-mail: yangjiang@cuc.edu.cn;houminxx@263.net; ningzi72@163.com

Abstract: We present a n approach to recognizing sentiment polarity in Chinese Reviews based on topic sentiment sentences. Considering the features of Chinese reviews, we identify the topic of a review using an n-gram approach. To extract topic sentiment sentences, we compute the semantic similarity of a candidate sentence and the ascertained topic and meanwhile determine whether the sentence is subjective. A certain number of these sentences are selected as representatives according to their semantic similarity value with relation to the topic. The average value of the representative topic sentiment sentences is calculated and regarded as the sentiment polarity of a review. Experiment result shows that the proposed method is feasible and can achieve relatively high precision.

Key words: topic sentiment sentence; review; orientation analysis; sentiment

1 引言

随着我国互联网事业的迅速发展，网络作为一种新型媒体不但成为各种社会思潮、利益诉求和意识形态较量的场所，而且也成为民众评议时政、谈是论非、交流观点的集散地。有关网络舆情监测和分析的研究由此引起研究人员的重视。

网络舆情分析中一个重要的研究内容是倾向性分析，这是一项用计算机来分析和处理文本中的观点、情感、态度、倾向等主观性信息的研究，又称“观点挖掘”、“情感分析”。近年来，由于在观点搜索、舆情分析、产品推荐、问答系统等诸多领域有着重要的应用，倾向性分析越来越受到人们的关注。

倾向性分析以主观性文本为处理对象。主观性文本“主要描述了作者对事物、人物、事件等的个人（或群体、组织等）想法或看法。”[1]其中，评论文是一类典型且常见的主观性文本，它针对具体的人、物、事件，就其有关方面做出批评议论。评论文，尤其是新闻评论，是社会舆论的集中反映。因此，评论文的倾向性分析对网络舆情分析具有重要的价值和意义。

2 相关工作

目前倾向性分析通常在词语、句子和篇章3个语言层级上展开，所采用的技术主要有基于语义的方法和基于机器学习的方法。专门研究篇章的整体倾向性的工作以Turney（2002）[2]、Pang等（2002）[3]、Yi等（2003）[4]为代表。Turney采用无指导的学习算法对评论文进行褒贬分类，首先通过计算给定词或短语与“excellent”和“poor”的互信息差来度量其语义倾向，然后将文本中词和短语的平均语义倾向作为给定评论文的整体倾向。Pang等分别使用朴素贝叶斯（Naïve Bayes）、最大熵（Maximum Entropy）和支持向量机（SVM）三种分类模型对电影评论文本的倾向性分类进行了研究，选取的特征包括词语的一元组、二元组、词性、位置以及特征的频数和特征出现与否等。Yi 等首先使用语法分析器对句子进行语法分析，然后参照情感词汇表和情感模式库对句子进行倾向性分类，并将其运用于文本的倾向性分类中。

由于语言是具有层级体系的符号系统，因此篇章的整体倾向性分析要以句子和词语的倾向性为基础。Wiebe 等[5]的研究表明，形容词可以作为判别句子主客观性的依据。Kim 和 Hovy[6]、Wiebe 和 Riloff[7]探讨了主客观句子的分类，Yu Hong 等[8]提出了面向自动问答系统的观点句抽取方法，再对抽取的观点句进行情感分类，判断其极性。Hu 和 Liu[9]通过 WordNet 的同义词—反义词关系，得到情感词汇及其情感倾向，然后由句子中占优势的情感词汇的语义倾向决定该句子的极性。Wang 等[10]选取形容词和副词作为特征，提出了基于启发式规则与贝叶斯分类技术相融合的评论句子语义倾向分类方法。王根、赵军[11]提出了一种基于多重冗余标记的 CRFs 句子情感分析方法，刘康、赵军[12]进行了基于层叠 CRFs 模型的句子褒贬度分析的研究。

与以往研究不同，本文提出一种基于主题情感句的评论文整体倾向分析方法。文章余下部分组织如下：第3节对本文研究的问题进行了分析，描述了提出的方法；第4节介绍了评论文的主题识别和主题情感句的抽取；第5节阐述了基于主题情感句的评论文倾向性分析；第6节给出实验结果及其分析；最后是结论。

3 问题分析和方法描述

3.1 文章的篇章结构

篇章的整体倾向性是其组成部分倾向性的总和，但各组成部分在特定篇章中的重要程度却有不同。这是因为不同的文章体裁有不同的篇章结构，而篇章结构体现了组成部分的重要程度。

篇章结构即篇章内部大于句子的意义单位的组织和构造规律，主要包括开头和结尾、过渡和照应、段落层次关系及谋篇布局的手段和方法。篇章结构在形式上标志了篇章内容的层次性，在意义上保证了篇章内容的完整性，在逻辑上体现了篇章内容的连贯性。篇章结构是篇章形式、篇章意义和篇章逻辑的统一体。从形式上看，篇章内部大于句子的意义单位中，自然段是人们可以自然地察觉到的基本单位，节、章等则是建立在自然段基础上的更大意义单位。从意义上看，篇章由若干个意义段组成，篇章的中心意义是各意义段的中心意义按一定逻辑关系的组合。

意义段是篇章内表达相对完整意义的自然段的集合，小到一个自然段，大到一个章节。不同文体划分意义段的依据不尽相同。就议论文而言，一篇典型的议论文依据其结构模式可以分为“引论”（或“总论”）、“分论”和“结论”等意义段。划分意义段对理解文章的篇章结构、把握中心思想具有重要意义。

3.2 评论文的特点

评论文是议论文的一种，也称作“评论”，从内容上看，包括人物评论、时事评论、经济评论、政治评论、军事评论、文学评论、艺术评论、商品评论、服务评论等。评论文具有以下特点：（1）主题明确。评论文与一般的议论文不同，它总是针对具体的人、物、事件的有关方面做出评议，议论的对象明确。（2）一篇评论文通常只有一个主题，评论者对主题有明确的倾向性。有的评论文会对主题的下位主题展开议论，但不影响其对该主题的基本立场。对下位主题的评论同样具有上述2个特点。（3）评论文的主题与其标题有着密切的关系。评论文为了让读者看到标题即了解主旨，通常会用精炼的语言道出文章的主题，有时甚至概括出主题和主旨。因此，一般来说，总可以在标题中找到文章的主题。（4）评论文的结构通常遵循一定的“范式”。概括起来，评论文的结构有4种基本类型：归纳型、演绎型、演绎归纳结合型和分论型，并分别对应4种不同的表达模式：“分—总”式、“总—分”式、“总—分—总”式和“分—分—分”式。评论者对主题的情感表达一般会出现在“总论”和“结论”部分，而“分论”部分的情感不影响其基本的倾向。在有的评论文中，对下位主题的情感表达会出现在“分论”部分。

对560篇评论文的考察结果印证了上述特点。以下为部分统计数据：

表1：评论文各项特点统计结果

评论文特点		所占比例（%）
主题明确		100
一个主题		100
标题反映主题		99.64
表达模式	引—分—总	40.17
	引—总—分—总	33.9
	引—总—分	18.75
	其他	7.18

3.3 主题情感句

由上述分析知，评论文的倾向性通过若干意义段按照特定的表达模式反映出来，其整体倾向一般出现在“总说”部分。因此，一个自然的想法是，对评论文进行意义段的划分，判定其表达模式，对包含“总说”的意义段进行倾向性分析，即可获得评论文的整体倾向。然而，对篇章结构进行自动分析本身是一件困难的工作，这个过程中损失的精度直接影响着篇章倾向性分析的准确率。为了避免完全的篇章结构分析，同时在一定程度上反映文章的篇章结构，我们引入主题情感句的概念，利用主题情感句潜在地表示了评论文的篇章结构这一特点，对评论文进行倾向性分析。

主题情感句是主观性文本中包含主题概念及与之相关的情感倾向的句子，它既包含着文章的主题，又表达了针对该主题的主观态度。就评论文而言，主题情感句是表达文章中心思想（主题+情感）的最典型、最直接、最有力的手段。主题情感句对于主题情感的表达具有鲜明的特点。首先，主题情感句在主题上是“同质”的。也就是说，主题情感句针对相同的主题发表意见。这就使得每个主题情感句中的情感可以计算。以往的研究（Turney，2002；Pang 等，2002）没有考虑主题及与之相关的情感应该相互对应这一问题，导致有可能把不同主题情感或不相关情感混合在一起计算，影响了结论的可信度。其次，主题情感句与文章主题在主题上的语义相似度大小

潜在地反映了主题情感句与不同意义段的相关度。主题情感句与文章主题的语义相似度越大，它出现在“总说”部分的可能性就越大。反之，出现在“分说”部分的可能性则越大。再次，主题情感句的分布情况，包括分布的密度和广度，不但潜在地表示了评论文的篇章结构是“总—分”，“分—总”抑或是其他类型，而且还或多或少地体现了作者对所讨论主题的情感强度，对更深层次上的情感分析有所帮助。

3.4 方法描述

综合以上分析，提出一种基于主题情感句的评论文倾向性分析方法。基本的思路是，在确定评论文主题的基础上，抽取出主题句；然后对主题句进行主客观分类，抽取出主题情感句；计算主题情感句与评论文主题的语义相似度，选取相似度最高的若干个句子计算情感倾向，将其平均值作为评论文的整体倾向。下面分别对各环节进行阐述。

4 评论文主题识别和主题情感句抽取

4.1 评论文主题识别

评论文的主题概念表示为词语串集合 $T=\{Wn_1, Wn_2, \ldots, Wn_i\}$，其中，$Wn_i$ 是一个词语或多个词语组成的词语串。评估 Wn_i 是否属于 T，依据的指标是其位置和频率信息。Wn_i 的位置信息表明了其分布度 $D(Wn_i)$：Wn_i 在评论文中的分布越广，其与主题相关的可能性越大。Wn_i 的频率信息表明了其重要度 $I(Wn_i)$：Wn_i 在评论文中的出现次数越多，其重要性就越大，与主题相关的可能性也就越大。由此，将 Wn_i 隶属于 T 的程度称为 Wn_i 的隶属度，Wn_i 的隶属度 $C(Wn_i)$ 定义为：

$$C(Wn_i)=\alpha \cdot D(Wn_i)+\beta \cdot I(Wn_i) \quad (1)$$

其中，α 和 β 是加权系数，用以调节 $D(Wn_i)$ 和 $I(Wn_i)$ 的权重。

为了快速有效地获取评论文的主题，采用一种基于词语 n-gram 匹配的方法进行识别。按照下述算法获取 T：

（1）对评论文标题和正文进行分词标注，分词标注结果分别存入队列 T_q 和 B_q 中。

（2）当 $n \leq m$ 时（其中，$1 \leq m \leq T_q$ 中词语的个数，n 初始值为 1 并自增），循环执行以下操作：连续地从 T_q 中取出一个 n-gram 词语串 Wn_i，并在 B_q 中进行查找；如果 B_q 中存在 Wn_i，则将其插入索引表 $G=\{Wn_i, position, frequency\}$ 中。令当 n=1 时，$W1_i$ 必须为实词。

（3）根据公式（1）计算每个 Wn_i 的隶属度，将隶属度大于预设阈值 L_c 的 Wn_i 加入 T 中。

4.2 评论文主题情感句抽取

主题情感句是主观性文本中包含主题概念及与之相关的情感倾向的句子，它既是主题句，又是情感句。主题情感句决定评论文的情感极性，是判别评论文整体倾向的关键。基于主题情感句的评论文倾向性分析方法将与主题无关的情感要素排除在外，使所分析的情感具有“主题同质性”，从而获得可计算性。主题情感句的抽取分为 2 个步骤。

（1）从评论文中抽取主题句。在已确定主题 T 的前提下，抽取主题句即选取与 T 在语义上相似度较高的句子，其相似度大小主要取决于二者等同词串（即形式完全相同的词或词串）的数量、等同词串的长度（即词串中所含词语的数量）、长度为 1 的非等同词的语义相似度、候

选主题句的位置等因素。根据索引表 G 中每个 Wn_i 的位置信息，可以确定一部分主题句。由于这些句子中含有一个或多个等同词串 Wn_i，按照 Wn_i 的数量及长度赋予一个相应较高的权值，表示这些句子与主题 T 的相似度很高。对于其他句子，根据刘群、李素建（2002）[13]提出的基于《知网》的词汇语义相似度计算方法，依次计算其所含词语与 T 中长度为 1 的 Wn_i 的语义相似度。考虑句子在文本和段落中的位置，将所有相似度大于预定阈值 L_s 的句子确定为主题句。为了获得较高的召回率，L_s 的值通常设置得较小。

（2）从主题句中抽取主题情感句。从主题句中抽取情感句，其实质是进行主客观分类。采用一种基于词典匹配的方法，使用预先编制好的情感词典来判别一个句子是否带有情感倾向。

通过以上步骤抽取了评论文中的主题情感句，每个句子均带有一个表示其与主题语义距离的权值，将其称为候选主题情感句集。

5 基于主题情感句的评论文倾向性分析

基于 2.3 节的认识，在评论文中，与主题相似度越高的主题情感句，越有可能成为作者表达基本倾向的关键句子。同时，为了避免过度依赖于少数的候选主题情感句，又要求对更多的句子进行分析。因此，从候选主题情感句集中选取的用于最后分析和计算的句子数量，是一个值得考虑的问题。评论文主题情感句的数量是不定的，这受多种因素影响。根据我们对 560 篇评论文的考察发现，一般而言，一篇评论文所包含的主题情感句不多于 7 个，而平均的主题情感句数量约为 4 个。此外，篇幅较长的评论文，其所包含的主题情感句也通常较多。由此，定义一个可调节的参数 γ（依据所分析的评论文篇幅与参考篇幅确定），则对于任一评论文，其所需分析的主题情感句数量 N(tss) 为：

$$N(tss) = 4 \pm \gamma \tag{2}$$

从候选主题情感句集中提取 N(tss) 个权值最大的句子，将所有句子的情感极性（SP）的平均值作为评论文的整体倾向 O(r)，即：

$$O(r) = \frac{1}{N(tss)} \sum_{i=1}^{N(tss)} SP(tss_i) \tag{3}$$

对于句子的倾向性分析，采用基于词典的语义方法进行。对于每一个待分析的句子，首先使用依存句法分析器对句子成分做依存分析，然后依据一个预先编制好的情感词典计算句子中情感表达式的情感倾向，并以此作为句子的情感倾向。分析过程中主要考虑了以下句法和上下文因素：（1）情感表达式与主题的关系；（2）情感表达式与其修饰成分的关系，包括否定词、强调成分等；（3）连接词语；（4）话语标记；（5）标点符号。

6 实验及结果

6.1 数据

实验中使用的语料为汉语时事评论，原始语料采集自人民网观点频道（http://opinion.people.com.cn），均经过了清洗和基本整理，使必要的文本结构信息可用。从中随机挑选出 400 篇文本，训练和指导 3 名标注人员独立地标注其情感主题句和整体情感倾向。以下是部分标注结果。

表 2：测试语料部分标注结果

标注者	正向情感文本	负向情感文本	其他
1 号标注者	87	302	11
2 号标注者	93	298	9
3 号标注者	88	288	14

对标注结果进行了一致性检查，最终得到 370 篇（其中，正向情感文本 86 篇，负向情感文本 284 篇）标注结果完全一致的评论文，将其作为测试数据。

6.2 资源和工具

实验使用了以下资源和工具：(1)情感词典。我们手工建设了正向情感词典(CUCPosSentDic)和负向情感词典（CUCNegSentDic），分别收集词条 9701 和 11681 例。每个词条均包含词性、正向情感值和负向情感值。不同于其他情感词典，我们由专家对词语的情感倾向进行 5 级赋值。所收词条部分来源于“知网”情感分析用词语集（beta 版）和 NTUSD（国立台湾大学情感词典），也收录了《学生褒贬义词典》、《褒义词词典》、《贬义词词典》等词典条目。（2）影响倾向性分析的上下文词典。包含否定词、连接词、话语标记等词典。（3）知网（2000 版）。使用了免费的知网（2000 版）用于词语相似度计算。（4）语言技术平台 LTP。使用了其中的依存句法分析器用于句法分析。（5）中国传媒大学分词标注软件（CUCSeg）。使用 CUCseg 进行词语切分和标注。

6.3 实验结果

主题情感句的抽取是本文工作中的至关重要的环节，我们对此进行了实验。采用传统的准确率（precision）、召回率（recall）以及 F_1 值（F-measure）等评价指标对性能进行衡量。实验结果如下：

表 3：主题情感句实验结果

隶属度阈值	准确率（%）	召回率（%）	F_1 值
L_s=0.64	89.9	82.3	86.1
L_s=0.55	86.1	90.6	88.3
L_s=0.37	77.8	98.4	88.1

可见，当隶属度阈值 L_s 为 0.55 时，可以获得较好的准确率和召回率。

对于评论文整体倾向性分析实验，采用准确率这一指标衡量本文方法的性能。测试数据的实验结果及与 Turney[2]、Pang 等[3]报告的结果比较如下：

表 4：评论文倾向性分析实验结果

方法	准确率（%）
Turney 的方法	74.39
Pang 的 SVM 方法	82.9
本文方法	86.8

显然，本文的方法在准确率上有较大提高。

我们对 49 个错误结果进行了分析，查阅了各个环节的中间分析结果。分析显示，约有 35%的错误来自主题识别阶段，大约 49%的错误是由于对主题情感句分析错误所导致，此外还有约

4%的错误由情感词典造成。其余错误原因有待进一步查明。因此，提高主题识别的准确率，加强对句子级的情感倾向的研究以及编制更好的情感词典，将有助于提高基于主题情感句的评论文倾向性分析结果。

7 结论

本文提出了一种基于主题情感句的汉语评论文倾向性分析方法。根据评论文的特点，采用基于词语 n-gram 匹配的方法识别主题，通过对比与主题的语义相似度和进行主客观分类抽取出候选主题情感句，计算其中相似度最高的若干个句子的情感倾向，将其平均值作为评论文的整体倾向。基于主题情感句的评论文倾向性分析方法避免了进行篇章结构分析，排除了与主题无关的主客观信息，实验结果表明，该方法准确率较高，切实可行。

参 考 文 献

[1]姚天昉, 程希文, 等. 文本意见挖掘综述[J]. 中文信息学报, 2008, 22(3): 71-80.

[2]P. D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews [A]. In: Proceedings of A CL-02,40th Annual Meeting of the Association for Computational Linguistics [C]. USA: 2002, 417-424.

[3]B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques [A]. In: Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing [C]. Philadelphia, USA: 2002, 79-86.

[4]J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques [A]. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003) [C].Melbourne, Florida: 2003, 427-434.

[5]J. Wiebe. Learning subjective adjectives from corpora[A]. In: Proceeding of the 17th National Conference on Artificial intelligence. Menlo Park[C]. Calif. AAAI Press, 2000:735-740.

[6]S.-M. Kim and E. Hovy. Determining the Sentiment of Opinions[A]. In: Proceedings of COLING-04, the Conference on Computational Linguistics (COLING-2004) [C]. Geneva, Switzerland: 2004, 1367-1373.

[7]J. Wiebe, E. Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Text[A]. In: Proceedings of CICLING[C], Mexico City, Mexico: 2005, 486-497.

[8]H. Yu and V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences[A]. In: M. Collins and M. Steedman (eds): Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing [C]. Sapporo, Japan: 2003, 129-136.

[9]M. Hu, B. Liu. Mining and summarizing customer reviews[A]. In Proceedings of the 10th ACM SIGKDD[C]. Seattle, USA :2004,168- 177.

[10] C. Wang, J. Lu, G. Zhang. A semantic classification approach for online Product reviews[A]. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on web intelligence[C]. Hongkong, China: 2005, 276- 279.

[11]王根, 赵军. 基于多重冗余标记 CRFs 的句子情感分析研究[J]. 中文信息学报, 2007, 21(5): 51-55.

[12]刘康, 赵军. 基于层叠 CRFs 模型的句子褒贬度分析研究[J]. 中文信息学报, 2008, 22(1): 123-128.

[13]刘群, 李素建. 基于《知网》的词汇语义相似度计算[A]. 第三届汉语词汇语义学研讨会[C], 台北, 2002:4-7.