**学校编码：10384 分类号 密级**

**学号：31520121153288 UDC**



**硕 士 学 位 论 文**

**一种用于GIF动画情感识别的时间-空间中层特征体系**

**A Spatial-Temporal Visual Mid-Level Ontology for GIF Sentiment Analysis**

**指导教师姓名：曹冬林 助理教授**

**专 业 名 称：计算机应用技术**

**论文提交日期：2016 年 4 月**

**论文答辩时间：2016 年 5 月**

**学位授予日期：2016 年 月**

**答辩委员会主席**：

**评 阅 人：**

**2016年5月**

**厦门大学学位论文原创性声明**

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（ ）课题（组）的研究成果，获得（ ）课题（组）经费或实验室的资助，在（ ）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年 月 日

**厦门大学学位论文著作权使用声明**

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（ ）1.经厦门大学保密委员会审查核定的保密学位论文，于　　 年　 月 　日解密，解密后适用上述授权。

（ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年 月 日

# 摘 要

近年来社交媒体快速发展，人们日益倾向于在社交网络中表达自己对社会热点事件的态度、网购产品的评价、热门影视作品的观感。在日益增多的社交媒体内容中，GIF动画正在占据日益重要的地位。据我们从新浪微博中获取到的数据，有15.1%的微博含有GIF动画等动态多媒体内容。在这些多媒体内容中，只有不到10%的内容是描述客观存在的物体或事件（例如堵车，雾霾），其余的都是用来表达博主主观情感的。（如同Brendan Jou在文64中写到的一样，GIF动画对预测用户情感很有帮助）

对GIF动画进行情感倾向性分析，可以帮助政府或公司了解网民对舆论热点的态度，为公司和政府提供决策支持。

目前为止，大部分微博情感分析研究都只关注于如何对文本信息进行分析，但是微博用户情感表达方式正逐渐转变，从以往的文本为主到现在的图文结合，基于文本信息的情感分析方法已经不足以很好的获取微博消息的情感极性。因此在文本信息的基础上加入图像信息来对微博情感进行分析，变得尤为迫切和重要。本文主要研究微博环境下图像情感语义特征表示以及文本与图像相融合的跨媒体情感分类问题，主要内容和创新点如下：

1. 针对现有基于SentiBank的图像情感特征ANP的检测噪声问题、区分度问题和情感话题相关性问题，本文提出了基于话题相关性模型的图像情感语义特征表示。话题相关性模型充分利用微博中同一话题下图像间的相关性和不同话题下图像间的差异性信息，对原始SentiBank特征进行选择和优化，一定程度上解决了ANP特征的检测噪声问题和ANP区分度问题，最终实验验证了基于话题相关性模型的图像情感特征表示的有效性。该图像情感语义特征的有效性使结合文本和图像的跨媒体微博情感倾向性分析成为可能。
2. 针对微博中短文本信息不足和用户表达方式转变的问题，本文提出了基于文本和图像信息融合的情感倾向性分类模型。该模型在文本词袋特征和图像情感语义特征表示的基础上，采用了特征层融合和决策层融合两种方法对文本特征和图像特征进行结合。通过图像和文本两部分信息的融合解决微博中表达方式从单文本到图文结合的转变问题，同时图像信息在一定程度上弥补了文本信息的不足。实验表明本文提出的融合模型在微博情感倾向性分类任务中，相比传统的单文本模型，分类准确率有所提升。本文进一步对比了文本和图像的特征层融合和决策层融合两种方法，实验表明目前采用决策层融合方法可以获得更高的准确率。

**关键词**：微博情感分类；话题相关性模型；跨媒体；特征融合

# Abstract

Sentiment analysis on Microblog has attracted ever increasing attention in recent years as the rapid development of social network. Large amount of Weibo users like to post their opinions on hot social events, comment on films or other technology products on Microblog site. Mining sentiment knowledge from such massive amount of data has a wide variety of applications. For example, it can help knowing the attitude of users to hot topics, providing decision support for the government or company.

To this end, most exiting methods mainly focus on analyzing the textual information. This has faced huge difficulty as nowadays users are more likely to express their feelings in a hybrid manner not only with texts, but also with images. It is therefore essential to take images into account in addition to texts. In this paper , we mainly focus on sentiment sematic feature representation of Microblog images and sentiment classification combing texts and images. The main contents and innovations of this paper are as follows:

1. To solve the problems of exiting sentiment ANP feature of image based on SentiBank, such as noise issues, low discrimination issues and topic relevance problems, we propose a sentiment sematic feature representation for image based on topic relevant model. This model take advantage of the correlation information between images in the same topic and the dissimilarities between images in different topics to select and optimize the original SentiBank feature. Our model solves the noise issues and low discrimination issues of the original SentiBank feature to some extent. Experiments demonstrate the effectiveness of the image sentiment feature based on our topic model, which makes the sentiment analysis combining texts and images possible.
2. To solve the information insufficiency of Microblog short texts and user expression manner change issues, we propose a cross-media sentiment classification model based on text information and image information fusion. This model employs feature-level fusion and decision-level fusion methods to combine texts and images based on bag-of-words text features and image sentiment sematic features. Our model solves the expression manner change issues in some degree by fusing text information and image information. In the same time, image information compensate for the shortage of text information to some extent. Experiments validate that the cross-media model performs better than the traditional text-based model. This paper further compares the feature-level fusion method and decision-level fusion method on image feature and text feature combination. Experiments show that the decision-level fusion method achieves higher accuracy than the feature-level fusion method.

**Key Words:** Microblog sentiment classification; Topic relevant model; Cross-media； Feature fusion

目录

[第一章 绪论 1](#_Toc417422794)

[1.1 研究背景及意义 1](#_Toc417422795)

[1.2 研究现状 2](#_Toc417422796)

[1.2.1 文本情感分析研究现状 2](#_Toc417422797)

[1.2.2 图像情感分析研究现状 4](#_Toc417422798)

[1.3 本文的研究内容及创新点 6](#_Toc417422799)

[1.4 本文的组织结构 7](#_Toc417422800)

[第二章 微博相关特性及情感分析技术 9](#_Toc417422801)

[2.1 微博相关特性 9](#_Toc417422802)

[2.2 文本情感分析相关技术 10](#_Toc417422803)

[2.2.1文本预处理 10](#_Toc417422804)

[2.2.2 基于情感词典的无监督情感分类 12](#_Toc417422805)

[2.2.3 基于机器学习的有监督情感分类 13](#_Toc417422806)

[2.3 图像情感分析相关技术 18](#_Toc417422807)

[2.3.1 颜色特征 18](#_Toc417422808)

[2.3.2 纹理特征 19](#_Toc417422809)

[2.3.3视觉词袋特征 20](#_Toc417422810)

[2.3.4 SentiBank语义特征 20](#_Toc417422811)

[第三章 基于话题相关性模型的图像情感语义特征表示 23](#_Toc417422812)

[3.1 引言 23](#_Toc417422813)

[3.2基于话题相关性模型的图像情感语义特征表示 26](#_Toc417422814)

[3.2.1 话题相关性模型构建流程 26](#_Toc417422815)

[3.2.2 基于话题模型和背景模型的相关性模型 27](#_Toc417422816)

[3.2.3 基于话题相关性模型的图像情感分类 29](#_Toc417422817)

[3.3 实验结果及分析 29](#_Toc417422818)

[3.3.1 数据集构建 29](#_Toc417422819)

[3.3.2 评测方法及评价指标 30](#_Toc417422820)

[3.3.3 图像情感分类效果 31](#_Toc417422821)

[3.3.4 实验案例分析 35](#_Toc417422822)

[3.4 本章小结 37](#_Toc417422823)

[第四章 基于文本和图像信息融合的情感倾向性分类模型 39](#_Toc417422824)

[4.1 引言 39](#_Toc417422825)

[4.2 文本特征抽取 41](#_Toc417422826)

[4.2.1 多信息结合的文本词袋特征 41](#_Toc417422827)

[4.2.2 特征选择 43](#_Toc417422828)

[4.3 图像特征抽取 45](#_Toc417422829)

[4.4 结合文本和图像的跨媒体特征融合 45](#_Toc417422830)

[4.4.1 特征层融合模型 45](#_Toc417422831)

[4.2.2 决策层融合模型 46](#_Toc417422832)

[4.5 实验结果及分析 52](#_Toc417422833)

[4.5.1 实验数据集构建 52](#_Toc417422834)

[4.5.2 实验设计 53](#_Toc417422835)

[4.5.3 实验结果展示 53](#_Toc417422836)

[4.5.4 实验案例分析 57](#_Toc417422837)

[4.6 本章小结 59](#_Toc417422838)

[第五章 总结与展望 61](#_Toc417422839)

[5.1 工作总结 61](#_Toc417422840)

[5.2 工作展望 62](#_Toc417422841)

[参考文献 64](#_Toc417422842)

硕士期间发表论文及参与项目情况 70

致谢 71

Table of Contents

[Chapter 1 Introduction 1](#_Toc417304941)

[1.1 Background and Significance 1](#_Toc417304942)

[1.2 Research Status 2](#_Toc417304943)

[1.2.1 Research Status of Text Sentiment Analysis 2](#_Toc417304944)

[1.2.2 Research Status of Image Sentiment Analysis 4](#_Toc417304945)

[1.3 Main Research Contents 6](#_Toc417304946)

[1.4 Outline 7](#_Toc417304947)

[Chapter 2 Microblog Characteristics and Sentiment Analysis Techniques 9](#_Toc417304948)

[2.1 Microblog Characteristics 9](#_Toc417304949)

[2.2 Techinques of Text Sentiment Analysis 10](#_Toc417304950)

[2.2.1 Text Preprocessing 10](#_Toc417304951)

[2.2.2 Unsupervised Sentiment Classification Based on Sentiment Lexion 12](#_Toc417304952)

[2.2.3 Supervised Sentiment Classification Based on Machine Learning 13](#_Toc417304953)

[2.3 Techinques of Image Sentiment Analysis 18](#_Toc417304954)

[2.3.1 Color Feature 18](#_Toc417304955)

[2.3.2 Texture Feature 19](#_Toc417304956)

[2.3.3 Bag of Visual Words 20](#_Toc417304957)

[2.3.4 Semantic Feature based on SentiBank 20](#_Toc417304958)

[Chapter 3 Sentiment Semantic Feature for Image Based on Topic Relevant Model 23](#_Toc417304959)

[3.1 Introduction 23](#_Toc417304960)

[3.2 Sentiment Semantic Feature for Image Based on Topic Relevant Model 26](#_Toc417304961)

[3.2.1 Construction Process of Topic Relevant Model 26](#_Toc417304962)

[3.2.2 Relevant Model Based on Topic Model and Background Model 27](#_Toc417304963)

[3.2.3 Image Sentiment Classification Based on Topic Relevant Model 29](#_Toc417304964)

[3.3 Experiment Result and Analysis 29](#_Toc417304965)

[3.3.1 Dataset Construction 29](#_Toc417304966)

[3.3.2 Performance Evaluation 30](#_Toc417304967)

[3.3.3 Image Sentiment Classification Results 31](#_Toc417304968)

[3.3.4 Case Study 35](#_Toc417304969)

[3.4 Conclusion 37](#_Toc417304970)

[Chapter 4 Sentiment Classification Model Based on Text and Image Information Fusion 39](#_Toc417304971)

[4.1 Introduction 39](#_Toc417304972)

[4.2 Text Feature Extraction 41](#_Toc417304973)

[4.2.1 Bag of Text Words Feature Based on Multi-Information Combination 41](#_Toc417304974)

[4.2.2 Feature Selection 43](#_Toc417304975)

[4.3 Image Feature Extraction 45](#_Toc417304976)

[4.4 Cross-media Feature Fusion Combining Texts and Images 45](#_Toc417304977)

[4.4.1 Feature-level Fusion Model 45](#_Toc417304978)

[4.2.2 Decision-level Fusion Model 46](#_Toc417304979)

[4.5 Experiment Result and Analysis 52](#_Toc417304980)

[4.5.1 Dataset Construction 52](#_Toc417304981)

[4.5.2 Experiment Design 53](#_Toc417304982)

[4.5.3 Experiment Comparison 53](#_Toc417304983)

[4.5.4 Case Study 57](#_Toc417304984)

[4.6 Conclusion 59](#_Toc417304985)

[Chapter 5 Summary and Prospect 61](#_Toc417304986)

[5.1 Summary 61](#_Toc417304987)

[5.2 Prospect 62](#_Toc417304988)

[References 64](#_Toc417304989)

[Publications and Research Programs](#_Toc417304989) 70

[Acknowledgements](#_Toc417304989) 71

# 第一章 绪论

## 1.1 研究背景及意义

近年来社交媒体迅速发展，国外如Twitter、Facebook，国内如新浪腾讯微博、微信。微博是社交媒体中极具优势和代表性的网站，中国的新浪微博从2009年成立以来，发展迅速，截至2013年初，其注册用户数已超过五亿，每日微博发帖量超1亿。大量用户在微博上发表评论或表达自我情感，比如对社会热点新闻的态度，对科技产品的喜恶，对电影的观感或心情等。如何对这些评论内容进行情感分析处理，满足各种应用需求，比如政府舆情监控，公司产品反馈分析等，非常具有应用意义，所以面向微博的情感分析成为研究者们日益关注的焦点。

由于微博发布字数的限制以及移动互联网的发展，在微博中一个引人关注的趋势是：除了文本信息之外，图像内容占据了越来越大的比例，用户的表达方式从以往的文本为主逐渐转化为图文结合。对新浪微博进行抽样数据统计，结果显示，一万条微博中附带图片的消息占比高达70%以上，而每条消息的平均字数仅有38个字。越来越多的用户在发布消息时附带了图片，这种情况在使用手机平台发布微博信息时更加普遍。在拍照及其方便，而文字信息相对繁琐且有长度限制的时候，用户更加愿意用文本与图片结合的方式来表达自己的心情或观点。比如图1.1中的两条包含文本和图像的微博消息，第一条消息为正极性，第二条为负极性，但是单独看文本并不能明显的看出其表达的情感，而结合图像之后便能轻易的判断出极性是正是负。微博只是社会生活中的一个缩影，在互联网不断发展变化的今天，生活的各个层面已经渐渐趋向于高度的视觉化，比如现在流行的微信朋友圈，发布的信息几乎都包含了图像。据观察，用户发布的信息中如果只包含文字信息，则其关注度明显不如文图结合的消息。相比于文字，大部分用户更喜欢浏览丰富多彩的图片。图像的特点是直观和形象，且没有文字的语言差异和文化差异，A picture is worth a thousand words(一张图胜过一千个字) ， 不同文化背景和不同知识结构的人可以很容易的理解图像的内容。

#星巴克# 解压解压



#雾霾# 这让我怎么去上学？



1. (b)

图1.1 包含文本和图像的微博消息示例

所以在微博情感倾向性分析中，考虑到文本信息不足和图像数据越来越多的现状，相比于传统单独基于文本信息的情感倾向性分析，加入图像信息更有助于准确把握用户的观点和情感。本文的研究内容是如何获得图像的情感语义特征表示，并在此基础上结合文本信息和图像信息提升微博情感倾向性分析的分类准确率，以解决微博中文本信息不足和表达方式从文本为主到图文结合转变的问题。

## 1.2 研究现状

传统的情感分析主要集中在文本分析上，吸引了很多研究者关注，且取得了一定的研究进展。而图像情感分析的研究相对而言要少很多，随着互联网图片数量的爆发式增长，近两年图片情感分析引起了一些研究者的关注。目前结合文本和图像的情感分析相关研究还很少，但是由于微博消息的内容已经从以往的文本为主体内容变成现在的文图结合，所以有必要结合文本和图像进行情感分析，于是本文做出了这方面的创新性尝试。下面分别对文本情感分析及图像情感分析的研究现状进行介绍。

### 1.2.1 文本情感分析研究现状

情感分析又称意见挖掘或倾向性分析，其任务是根据文本内容判断用户的正面、负面或中性情感倾向性。对以往的研究进行总结，发现文本情感分析方法大致分为两类：基于情感知识的无监督方法以及基于机器学习的有监督方法。

1. 基于情感知识的无监督方法

基于情感知识的无监督方法主要是依靠已有的情感词典、领域词典以及文本中带有情感极性的组合评价单元计算文本的情感极性[1,56]。该方法的优点是无需标注训练数据，但是对情感词典依赖性大。情感词典的构建，有基于语料、基于词典等方法，比如早期的Riloff等人[2]基于语料数据库用统计的方法构建了语义词典，Hatzivassiloglou等人[3]从大量语料中提取所有同现的形容词对，采用对数线性回归分类器和最后的聚类方法将形容词分为褒贬两个类别。基于词典的方法主要是根据现有的单词网络或通用词典，比如国外的WordNet，国内的HowNet或同义词林来构建[4-5]。在基于情感知识的分类方法中，一个典型的方法是Turney[6]提出的，他们利用预先定义词性组合从而在数据中提取出情感短语，并通过点互信息方法计算情感短语的极性值，最终对句子中的情感短语极性值进行整合，对句子进行情感分类。Yu等人还构建了特殊领域情感词典，从而对股票指数进行预测[7]。

1. 基于机器学习的有监督方法

有监督方法解决情感分类任务，需要有包含一定数量的标注样本数据集作为基础，利用特征抽取和机器学习分类方法训练分类器，从而对文本进行分类。Pang[8,9]是首次在情感极性分类中引入机器学习的，他选取了n-gram(n元文法，n一般取值1，2或3)和词性标注的组合作为特征，使用支持向量机(SVM)，朴素贝叶斯(Naïve Bayes)和最大熵(Maximum Entropy)三种分类算法对电影评论进行极性分类，并对比各种特征和分类器的效果，最后发现基于一元文法的特征表示结合SVM分类算法可以取得最佳的分类效果。此后，便有很多人在Pang的基础上做了很多特征表示、特征选取和分类算法的研究，以提高情感分类的准确率。在特征表示和特征选择的研究中，文献[10-12]采用了不同的特征，比如上下文、依存关系以及文档子成分信息等。Deng等人[13]对特征的加权策略进行了研究。唐慧丰等人[14]对特征选择方法及机器学习分类技术进行了比较研究，选择了名词、动词以及形容词等作为文本特征，对比了文档频率、信息增益等特征选择方法以及支持向量机、贝叶斯等机器学习分类算法。Zheng[15]采用了N元文法特征和常用特征选择算法。陈俊杰等人[16]在基于SVM的情感识别中加入了模糊理论。

1. 微博情感分析现状

以往本文情感分析主要集中在长博客或者较为规范的商品评论和电影评论上。自从提供英文微博服务的社交网站Twitter迅速发展起来后，研究者们发现对其中的海量数据进行情感分析非常具有现实意义，所以面向微博的情感分析在近几年来成为了研究热点。由于Twitter相对国内的微博发展较早，且针对中文的情感分析比英文有更多的困难，所以相比来说基于Twitter的情感分析研究比基于中文微博的多，分类效果较好。Davidiv等人[17]使用Tweets(微博消息)中的hashtag(标签)和表情符号作为类别标注，利用K近邻算法训练出分类模型，从而对微博消息进行情感极性分类。Barbosa[18]等人利用已有的面向Twitter的情感分析网站和工具(Twitter Sentiment，TweetFeel和Twendz)对Tweets消息进行极性分类，对不同工具的分类结果进行统计预处理，过滤掉预测结果不一致的消息，最终根据结果对数据进行标注得到训练集；然后选取情感词、表情、动词等作为特征，采用支持向量机进行分类。Go等人[19]通过设定的词语检索Twitter中含有笑脸和哭脸表情的Tweets消息，根据表情对消息进行正负极性标注，选取1-gram(一元文法)、2-gram(二元文法)以及POS(词性标注)的组合作为特征，利用常用的支持向量机、贝叶斯和最大熵分类算法训练分类器，最终在正负极性2分类任务上取得了80%多的准确率。Bravo-Marquez等人[20]将流行的情感分类系统进行结合，加入表情等其他特征，对Tweets进行极性分类。

关于中文微博的研究在这几年迅速增加。较为典型的研究是谢丽星等人[21]提出的基于SVM的层次结构多策略方法，采用了基于情感词典的主题无关和主题相关统计值特征，对文本进行主客观分类和正负极性主观分类，并将提出的方法与基于表情符号和基于情感词典的方法进行了对比。最终实验结果表明基于主题相关的特征和SVM的层次结构多策略方法效果较好，准确率为67.28%。除了微博本身的消息外，也有研究者将评论信息加入进行分析[22]。Lu等人在进行情感分类时还考虑了微博用户间的关系[23]。

### 1.2.2 图像情感分析研究现状

以往情感分析研究主要集中在文本上，图像或者视频相关的情感分析工作相对较少，近两年开始慢慢受到一些研究者的关注。一个不可忽视的问题是相对传统的单文本内容，网络上的图像和视频内容占据了越来越大的比例，能够对图像的情感分类进行准确预测是非常有必要且有意义的，比如可以应用到图像情感检索任务中，或者与文本结合更好的判断社交网络中用户发布信息的情感极性。图像情感分析问题的研究存在着很多挑战，首先，它比图像目标识别更难，因为后者可以很好的定义各个类别，而图像情感具有更高层的抽象语义，人在判别的过程中也会有很多的主观性，比如要对一张图像进行情感极性判别，首先要看到图像中包含的对象、场景、动作和事件等，然后在这些内容的基础上升级为情感识别。其次，要构建一个有一定规模且覆盖性广的图像情感分类的数据集是非常困难的，除了数据源选择，数据抓取和标注也非易事。另外，训练的模型需要有领域自适应性，因为微博的图像没有特定领域，可能包含任何一个领域的内容。

1. 图像情感分析相关研究

较早的图像情感分析相关研究工作是图像美学质量研究[24-26]和基于情感的图像检索[27-29]。在图像美学质量研究中，Marchesotti等人[26]通过选取合适的特征对图像的美学质量进行分类判别，他们选取了较为通用的图像特征，包括GIST[30], 视觉词袋特征(Bag-Of-Visual Words，BOV)[31]以及Fisher Vector(FV)特征[32]，利用SVM分类器训练分类模型，最终取得了不错的分类效果。Zhao等人[29]提出了通用底层特征、美学特征及中间层特征结合的特征表示，底层特征抽取了HOG、GIST、颜色直方图和纹理特征等，美学特征采用艺术启发的构图特征，中间层属性特征是102个由SUN数据库设计的属性特征[33]。基于情感的图像检索研究中，Wang [28]等人进行了较为全面的调研，对视觉情感语义特征表示、特征提取(多为颜色纹理以及美学启发的构图等底层特征)，分类算法(回归分析，神经网络，支持向量机等)进行了全面的分析介绍。

近两年与图像情感分析相关工作的有图像受喜爱度分类(affection) [34-37]或情绪(emotions)分类[38-39]。Machajdik[34]等人根据心理学和美学理论设计出较为细致的底层特征组合，这些特征更能代表图像的情感内容，具体颜色特征，纹理特征，复杂度、景深、三分构图法、动力学等图像结构特征，还有人脸和肤色的图像内容特征。最终在International Affective Picture System (IAPS)[40]数据集上训练分类器，将图像分为快乐(Amusement)、生气(Anger)、敬畏(Awe)、满足(Contentment)、恶心(Disgust)、兴奋(Excitement)、恐惧(Fear)、悲伤(Sad)八个情绪类别，最终取得了平均约60%的准确率。Joshi等人对图像美学和情感进行了全面介绍[47]。

1. 图像情感特征表示概括

目前为止，关于图像情感分析的特征表示主要有颜色、纹理等底层特征表示[41-42]和中层特征表示[43-45]以及深度学习特征表示[46]等方法。在底层特征表示研究中，Siersdorfer等人定义了全局及局部的RGB颜色直方图，同时还抽取了基于SIFT的视觉词袋特征(Bag of Visual Terms)，用贝叶斯和支持向量机将Flickr上获取的图像分类为正负两类，最终实验表明颜色直方图与词袋特征的融合取得了最好的效果。

中间层特征表示主要有面部表情特征[43]、属性(Sentribute)特征[44]或者ANP(Adjective Noun Pairs，形容词名词对)特征[45]。其中D Borth等人[45]提出的ANP特征是目前分类效果比较好的特征表示。该特征是D Borth在2013年提出的图像语义特征表示方法，并构建了较大规模的视觉情感本体数据集和相应ANP特征(Adjective Noun Pair, ANP)检测子。他们首先根据Plutchnik的情绪轮模型[53]中的24个基本情绪词在Flickr和YouTube上搜索还有情绪词的标签，然后对标签进行处理得到有明显情感倾向性的形容词名词对，并对每个形容词名词对构建对应的图像数据集，最终形成了3000多个ANP集合构建出视觉情感本体库(比如happy baby, dark clouds)，每个ANP都有一个在-2~2之间的极性值和对应的图片集。根据ANP对应的图像数据集最终训练出了ANP检测子，并选取出准确率较高的2089个ANP检测子构建出SentiBank。这些ANP可以构成图像的中间层语义特征表示，最终训练分类器对图像进行情感极性分类，实验证明了该特征相比底层特征具有更高的分类准确率。该文章提供的数据集也让面向社交网络的图像情感分析有了新的进展。

## 1.3 本文的研究内容及创新点

本文的研究内容：

本文的研究内容围绕微博情感倾向性分析展开。针对目前微博中文本信息不足、图像数据越来越多、用户更倾向于使用图文结合的方式表达观点和情感这三方面的问题，本文研究微博环境下图像情感倾向性特征表示以及文本与图像相融合的跨媒体情感倾向性分类问题。在该研究思想的指导下，首先为解决现有基于SentiBank的图像ANP特征的噪声问题和区分度问题，本文利用微博话题下图像的相关性信息构建话题相关性模型，得到图像的情感语义特征表示，形成跨媒体情感倾向性分析的特征基础。其次，针对文本和图像特征的跨媒体信息融合问题，选取合适的特征融合方法对文本信息和图像信息进行结合，更好的提升微博情感倾向性分类性能。

本文的主要创新点：

1. 针对现有基于SentiBank的图像情感特征ANP的检测噪声问题、区分度问题和情感话题相关性问题，本文提出了基于话题相关性模型的图像情感语义特征表示。话题相关性模型充分利用微博中同一话题下图像间的相关性和不同话题下图像间的差异性信息，对原始SentiBank特征进行选择和优化，一定程度上解决了ANP特征的检测噪声问题和ANP区分度问题，最终实验验证了基于话题相关性模型的图像情感特征表示的有效性。该图像情感语义特征的有效性使结合文本和图像的跨媒体微博情感倾向性分析成为可能。
2. 针对微博中短文本信息不足和用户表达方式转变的问题，本文提出了基于文本和图像信息融合的情感倾向性分类模型。该模型在文本词袋特征和图像情感语义特征表示的基础上，采用了特征层融合和决策层融合两种方法对文本特征和图像特征进行结合。通过图像和文本两部分信息的融合解决微博中表达方式从单文本到图文结合的转变问题，同时图像信息在一定程度上弥补了文本信息的不足。实验表明本文提出的融合模型在微博情感倾向性分类任务中，相比传统的单文本模型，分类准确率有所提升。本文进一步对比了文本和图像的特征层融合和决策层融合两种方法，实验表明目前采用决策层融合方法可以获得更高的准确率。

## 1.4 本文的组织结构

1. 绪论

介绍了本课题的研究背景和意义，对本文情感分析和图像情感分析的现状进行了总结，在此基础上提出了本文的主要研究工作和组织结构。

1. 微博相关特性及情感分析技术

对微博的相关特性进行概括，说明研究存在的挑战和机遇；对文本情感分析的相关技术进行介绍，包括文本预处理，常用分类方法；对图像情感分析的特征表示进行总结，为文本和图像结合的微博情感倾向性分析提供理论基础。

1. 基于话题相关性模型的图像情感语义特征表示

在本章中，首先指出了基于SentiBank的图像情感特征存在的问题以及微博话题下图像间可利用的相关性信息；然后针对问题提出了基于话题相关性模型，详细介绍了话题相关性模型的构建过程以及基于该模型得到的图像语义特征表示；最后通过实验验证模型的有效性，同时对实验结果进行较为详细的分析。

1. 基于文本和图像特征融合的跨媒体情感倾向性分类

本章首先介绍文本和图像结合的跨媒体模型中文本情感特征的构建流程；然后在文本情感特征和图像情感特征的基础上采用了特征层融合方法和决策层融合方法，对两种特征融合方法进行了详细介绍；最终设计实验对比单独文本情感倾向性分类模型和跨媒体情感倾向性分类模型，对本文提出模型的有效性进行说明，并对实验结果进行分析。

1. 总结和展望

总结全文的工作，分析存在的不足，对未来的研究提出建议和设想。

# 第二章 微博相关特性及情感分析技术

## 2.1 微博相关特性

微博与传统的长篇博客、新闻、商品评论等文本相比，具有很多的独有特性，这些特性使得微博文本情感分析与传统文本情感分析不同，具有更多的挑战。微博的相关特性具体可分为如下五个方面：

1. 长度限制：微博消息最多只能书写140个字。从我们包含9千多条消息的训练集中统计得出每条微博的平均字数约为38个字，且通常只包含一到三个短句，这就导致了微博消息的数据稀疏性问题。由于字数太少，比较难抽取到明显的情感要素，特征空间模型表示遭遇严重数据稀疏，这与以往字数常常为几十到几百的的文本情感分析有着较为巨大的区别，挑战性更大。
2. 文本不规范性：由于微博中的内容是上亿网民自由创作，表达方式十分口语化，频繁使用形象生动的表情符号、新近流行的网络词汇及网络链接等，且常常出现缩写词汇，拼写错误的情况时有发生。由于文本分析常常需要分词处理，语言的不规范性和网络新词的出现，大大增加了分词的难度，从而对最终情感分类效果造成不良影响。
3. 领域多样性：传统的评论性文本大多针对某特定领域，比如影评、购物评价等。但是微博内容涉及的话题多种多样，可以说是无领域性的。从微博的热点话题我们就可以看到，有针对社会、体育、综艺、娱乐、电视电影、旅游、美食、科技、财经等。用户发表的内容可能是对某社会热点事件的看法，对电影或者科技产品的评论，抑或某刻的感悟或心情等。
4. 跨媒体多模态：微博内容已经不再是单独的文本形态，表情、图像、音频及视频等其他模态的信息已经占据了越来越多的领地，其中图片信息是文本信息以外分布最多的媒体数据。据抽样调查的统计结果显示，10000条微博中有7081条消息附带了图片。越来越多的用户在发布信息时，更加倾向于用图片、或者文本结合图片的方式来表达自己的观点和心情，比如图2.1中的微博消息。文本图像结合的表达方式更加造成了文本信息的不足，同时也使得加入图像的微博倾向性分析成为可探索的新方向。
5. 数据可用性：微博中庞大的数据量蕴藏了很多有用的信息，能获取微博数据进行研究可以有广泛的应用前景。微博提供了获取数据的接口，同时我们也可以自己编写代码有针对性的采集所需要的实验数据。



图2.1 包含图片的微博消息示例

综合以上特性可以看出，微博的情感极性研究中，存在着文字信息不足，而图像数据越来越多的问题，所以文本创新性的采用了文本和图像结合的方式进行微博情感倾向性分析，并从微博的可用数据中采集了文本和图像数据构建出实验数据集。下面我们针对文本情感分析和图像情感分析的相关技术进行介绍。

## 2.2 文本情感分析相关技术

### 2.2.1文本预处理

(1) 标签符号过滤

微博文本的多元化表达使得语言中有较多的特征，比如标签符号、网址链接、表情符号等，其中标签符号对于后续的文本分类不起作用，需要进行过滤。而网址链接和表情符号则可保留作为情感特征的一部分。

1. 四种常用标签符号：

：该符号代表某用户，后跟着被用户的昵称。如果转发某条微博，系统则会在转发内容前加上“用户昵称”。

//：出现第二次及以上转发已转发的微博时，微博系统会为消息自动添加符号//，起分割作用。转发后的文本的内容主要是对原始微博的评论、补充、或与之相关的信息内容。

#：微博可以由用户发起话题讨论，且会根据讨论帖子数和阅读数对话题进行排行。话题标签就是有两个#之间的文字。热点话题包含了社会、娱乐、科技、体育、综艺影视等分类，比如#雾霾#，#云南鲁甸地震#，#科比#，#iPhone6#等。从话题排行榜上，我们可以了解到当下广大用户所关心的热点问题，以及社会发生的热点事件。从热门话题中抽取微博消息并进行情感倾向性分析，不仅可以了解公众对某个事件或产品的观点和倾向性，还可以对社会整体的一个舆论情况进行了解，同时也可以获取地域信息，从而掌握舆情的地域分布情况。本文的实验数据便是从微博热门话题中进行采集的。

1. 网页链接：微博服务的功能中保护了分享图片、网页、视频等功能，用户在分享内容时，系统会自动加上所分享网页的链接，如：http://t.cn/z8AMj4k。
2. 表情符号：微博平台提供了很多表情供用户使用，用户也很喜欢通过表情来表达自己的心情。对198570条微博热门话题数据进行统计，发现含有表情的微博消息有58776条，占29.6%，可见表情符号在微博情感倾向性分析中有着非常重要的作用。

对于上述@ ，//，#这几种标签符号，对情感极性分析无作用，所以预处理时要去除这些标签符号，同时去除用户昵称、话题标签。

(2) 微博分词

中文句子书写中词汇之间没有分隔符，但是通常词才是表达语义的最小元素，单个字无法完整获知其所要表达的意思。在文本情感识别中，通常情感词是构成情感句子的最基本单位，通过对情感词所表示的情感进行识别，基本就可以确定句子所表达的情感，所以对中文文本进行分词处理是情感分析的基础。目前自动分词技术较为成熟的算法有：基于隐马尔可夫模型的分词算法、基于n-最短路径的分词算法和最大匹配分词算法等[57]。在这些算法基础上实现的分词系统有很多，效果较好的系统有NLPIR、IKAnalyzer、庖丁解牛分词等。

目前分词系统中最广为人知、被广泛应用的是中国科学院计算技术研究所的NLPIR分词系统，也称ICTCLAS [48]。该分词系统使用的分词算法是隐马尔可夫模型。系统主要功能包括中文分词、词性标注、命名实体识别、用户词典等功能。目前的分词精度达 98.45%，分词速度为单机 500KB/s。本文的中文文本预处理便采用了NLPIR分词系统。

### 2.2.2 基于情感词典的无监督情感分类

基于情感词典的无监督方法通过对句子中的情感语素进行识别，进而获得整个句子的情感极性[56]。具体判别思路为：基于现有的情感词库，利用大规模的语料数据，构建出正向情感词典和负向情感典，从而得到待预测文本中词汇的情感极性值，最后根据下面的公式计算待预测文本的情感极性强度：

 (2-1)

其中和分别表示文本d中包含的正极性情感词和负极性情感词的情感权值。文本最终的极性类别判断公式为：



情感词典构建主要有人工标注和统计学习两种方法。人工标注通过语料收集、词语分割和人工标注词语情感值，从而得到情感词典，比如现有的HowNet情感词典、NTUSD、大连理工大学情感词汇本体库都是通过人工标注得到的。统计学习方法构建情感词典的典型思路是：选取一定数量的正负情感极性种子词，计算候选词与种子词的语义相似度，根据该相似度对候选词进行情感倾向性判别，将具有明显情感倾向性的词语加入情感词典。文献[5]便是利用HowNet计算汉语常用词与情感种子词的语义距离，从而得到词语的褒贬倾向性。另外一种典型的方法是Turney等人[6]提出的点互信息(PMI)方法，根据互联网搜索结果计算情感短语与预定义种子词之间的PMI值，从而获得情感短语的倾向性值。基于人工标注的情感词典构建方法优点是精确率高，缺点是耗时耗力；基于统计学习方法构建词典的优点是耗费资源少，效率高，但是种子词的选取和词语相似性计算的准确性都直接影响着所构建词典的优劣。

基于情感词典的情感极性分类方法优点是简单高效，无需标注训练语料，但是其分类效果受限于情感词典的准确性和全面性。

### 2.2.3 基于机器学习的有监督情感分类

基于机器学习的文本情感分类流程如图2.2所示。其中文本特征抽取、文本特征选择和分类器三部分都是影响最终分类效果的关键步骤。

训练集

文本特征抽取

文本特征选择

训练分类器

文本预处理

测试样例

文本预处理

分类模型

文本特征向量

分类结果

**训练过程**

**测试过程**

图2.2 基于机器学习的文本情感分类流程

1. **文本特征抽取**

在文本倾向性分析中，最为常用的文本特征是N元文法特征(N-gram)，其中一元文法特征也称为词袋特征(Bag-of-Words)。Pang B等人最早证实了N元文法特征在情感倾向性分类任务中的有效性[8]。N元文法是由N个连续的词(字)组成的观察序列，在文本特征表示中通常采用一元文法和二元文法。

通过对微博预料进行预处理分词，得到的N元文法词汇集合为，特征的权值计算方法一般分为三种：

布尔权值：文档的特征向量表示为。该计算方法中，特征向量的权值只能取值为1或0，1表示词汇在文档中出现，0表示未出现。布尔模型不考虑文档中词的出现频率，表示简单，由于微博文本中包含的词语较少，出现频率也多为1次，所以布尔模型在微博短文本的特征表示中性能较好；

词频权值：在词频权值计算方法中，特征向量权值为词汇在文档中的出现次数；

TF-IDF(Term Frequency-Inverse Document Frequency)：TF-IDF用于估测特征词对于文档的重要程度，其计算公式为：

 (2-2)

其中表示特征词在文档中出现频率的归一化值， (逆向文件频率)是语料库文档总数除以包含词汇的文档数。

刘志明等人[49]采用了一元文法(unigram)和二元文法(bigram)作为文本特征对微博进行情感倾向性分析。实验表明采用布尔权值计算的一元文法特征取得了更好的分类性能。

1. **文本特征选择方法**

在基于N元文法的文本特征表示中，一方面由于不同的词语数量巨大，导致特征向量维度很高，另一方面由于微博的短文本特性，特征会非常稀疏。如果将所有特征直接用来分类，不仅计算复杂度高，还会有很多噪声影响分类效果。所以采用合适的特征选择方法从原始特征中筛选出类别区分能力强的特征，是文本处理中的重要步骤，在降低特征维度的同时提升分类性能。

目前常用的特征选择方法有：信息增益(Information Gain, IG)、点互信息(Point-Wise Mutual Information，PMI)、文档频率(Document Frequency，DF)、卡方统计(Chi-Square correlation，CHI)等[49]。

1. 信息增益

信息增益是文本处理中典型的特征选择方法，它通过评估文档中特征项出现与否对分类性能的影响，来对特征的区分度进行排序。信息增益计算公式如下：



 (2-3)

其中，和分别表示类别和特征项*t*在语料库中出现的概率。表示包含特征项*t*的文档属于类别的条件概率，表示语料库中文档不含特征项*t*的概率，表示不包含特征项*t*的文档属于类别的条件概率。特征项的信息增益值越大，则表明该特征越重要。

1. 点互信息

点互信息在统计语言模型中被用来度量两个随机变量间的相关性。在分类任务中，它可以用来衡量特征项和类别之间的依赖程度。如果依赖程度越大，则特征越重要。互信息计算公式为：

 (2-4)

其中，是特征t出现的概率。表示特征t在类别中出现的概率。如果=0，则表示特征t与类别相互独立，没有相关性。如果PMI值越大，则特征与类别的关联性越大，该特征越重要。

1. 文档频率

文档频率方法是计算整个数据集中包含某特征项的文档数量与数据集总文档数量的比值 ，然后与根据设置的阈值相比来进行特征选择，除去文档频率小于某个阈值或者大于某个阈值的特征。该方法的优点在实际应用中简单有效，适合超大规模的数据集，但是没有考虑特征项与类别直接的相关性，去除了一些可能对分类起重要作用的低频词或者高频词。

1. 卡方统计

卡方统计通过估测特征与类别间的依赖性来进行特征选择 [1]。CHI值越大，特征与类别的依赖性越强。值计算公式如下所示：

 (2-5)

特征t的CHI值计算公式为：

 (2-6)

公式(2-5)中，N表示数据集中的样本总数。表示不包含特征且不属于类别的消息数，表示不包含特征t且属于类别的消息数，表示包含特征t且不属于类别的消息数，表示即包含特征t又属于类别的消息数。

1. **分类算法**

在文本情感倾向性分类算法，朴素贝叶斯和支持向量机是两种非常常用且分类效果较好的分类算法。

1. **基于朴素贝叶斯的文本情感分类**

朴素贝叶斯是贝叶斯方法中实用性很高的是一种分类算法，它的优点是分类速度快、准确性高，可处理大规模数据，在文本分析中有不错的分类性能。贝叶斯方法的理论基础是贝叶斯定理，其公式表达为：

 (2-7)

朴素贝叶斯算法最大的特点是它基于条件独立性假设，也就是给定类别时文本特征词之间相互条件独立。该算法的分类思路是根据类别的概率和给定类别下特征项出现的条件概率，计算给定特征值时各个类别的后验概率，将文本分到后验概率最大的类别中。下面是该算法解决文本情感分类问题的具体流程描述：

* 设文本训练样本集分为k个类别，类别集合记为，在文本情感极性中一般有正向、负向和中性三个类别；
* 设T为数据集中的一个待分类文本，根据上一小节中文本特征抽取和特征选择后得到T的m维特征表示为；
* 朴素贝叶斯的目标是在给定文本的特征属性下，根据给定特征值下的类别后验概率，得到最可能的类别：



根据贝叶斯公式上述表达式可重写为：



 (2-8)

其中根据朴素贝叶斯的条件独立性假定可以改写为：



将其代入公式(2-8)中，即可得文本T最终的目标分类：

 (2-9)

1. **基于支持向量机的文本情感极性分类**

支持向量机(SVM)是V.Vapnik等人在1995年提出的一种机器学习算法[50]，目前为止，SVM已经应用到各个领域的分类和回归问题，比如图像分类、文本分类、行人检测、语音识别等等，且都取得了不错的分类效果。

SVM方法是建立在统计学习理论中的VC理论和结构风险最小化(Structural Risk Minimization，SRM准则)基础上的，其分类思想为： 基于结构风险最小化理论，在特征空间中构建最优分割超平面，通过间隔最大化的学习策略，将问题转化为一个凸二次规划问题进行求解，保证学习器全局最优化。SVM针对线性可分情况进行分析，如果线性不可分，则通过核函数方法将低维空间特征映射到高维特征空间，使得数据线性可分。

将支持向量机应用到文本情感极性分类任务中，可以归结为在文本特征空间中求解最优分割超平面，该超平面使得文本正负极性类别支持向量之间的距离最大，分类图解如图2.3所示。图2.3中，H表示最优分割超平面，Gap为最大分类间隔。假设有n个正负2类文档组成的训练数据，存在一个超平面将文档完全分开，即找到判别函数，如果，则样本判别为正类，如果，则样本判别为负类。如果数据集中的样本是线性可分的，则直接通过求解最大化分类间隔来求得最终的分类函数；如果样本线性不可分，则需要通过核函数将数据映射到高维空间进行求解。常用的核函数有多项式核函数、径向基核函数(RBF)和Sigmoid核函数。

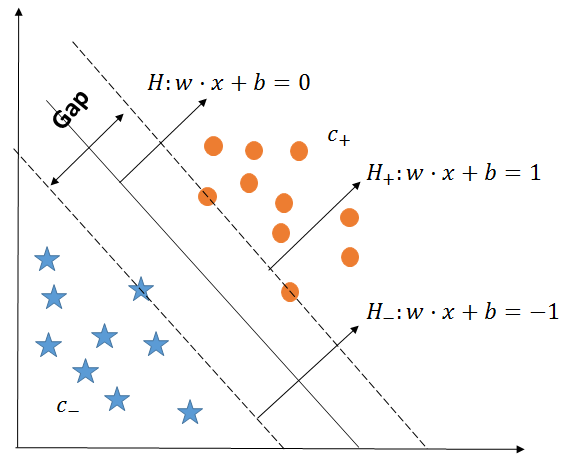


图2.3 SVM线性分类最优分割超平面

## 2.3 图像情感分析相关技术

图像情感倾向性分类的流程是图像情感特征抽取和分类器训练，其中最为关键的步骤便是图像情感特征的表示，分类器采用目前分类效果较好的支持向量机，朴素贝叶斯、逻辑回归等作为分类器，其分类原理与上述文本情感倾向性分类相似。而图像情感特征表示方面，目前典型的图像情感特征表示可以分为颜色、纹理等底层特征表示、Bag-of-Words中间层特征表示和SentiBank语义特征表示。

### 2.3.1 颜色特征

颜色是图像情感极性判断的一个重要元素，比如一幅蓝色天空图和一幅灰色天空图，主要根据颜色判断出其情感极性。描述图像颜色信息的特征主要分为下面三种：

饱和度(Saturation)、亮度(Brightness)和色调(Hue)统计信息：文献[51]的研究表明，饱和度和亮度可以直接影响用户对图像情感空间中开心(Pleasure)、兴奋(Arousal)、主导(Dominance)三个维度的主观判断，而色调也会影响用户对图像情感的反应，所以图像的饱和度、亮度、色调信息是图像情感底层特征表示中的重要部分；

颜色多样性(Colorfulness)：根据文献[51]中的理论，通过计算图像颜色直方图和统一颜色分布直方图的EMD距离(Earth Mover's Distance)来评估图像的颜色多样性。图像中颜色的多样性也会影响用户对其情感语义的主观反应。

颜色名称(Color Names)：在情感表达中，每一种颜色都有其对应的特殊意义。通过Weijer等人提出的方法[52]统计图像中11种基本颜色：黑色，蓝色，棕色，绿色，灰色，橙色，粉色，紫色，红色，白色，黄色包含的像素数。

### 2.3.2 纹理特征

纹理特征同样是图像情感表达的一种重要特征。(Local Binary Pattern，局部二值模式)是纹理特征中非常有效和常用的图像局部纹理特征描述算子[53]。

通过计算每个像素值来获得图像最终的特征表示，计算流程如下图所示：



图2.4 LBP特征值计算示例

从图中可以看到，对于每个像素点P，将其像素值作为阈值，算子在该像素的邻接区域内每个像素的灰度值与像素点P的阈值进行比较。若邻域像素值大于中心像素值，则该像素点的位置被标记为1，否则为0。邻域的标记值获得一个八位二进制编码，该二进制编码的十进制值即为该像素点的LBP特征值值，这个值反映了该区域的纹理信息。

LBP纹理特征向量具体计算步骤如下：

(1) 图像切分。将图像分割为的子块，计算每个子块中每个像素点的LBP值；

(2) 直方图统计。对每个子块进行直方图统计，得到每个子块的直方图；

(3) 特征表示。个子块直方图构成图像的LBP纹理特征。

### 2.3.3视觉词袋特征

视觉词袋特征(Bag of Visual Words)是在文本词袋模型的思想上衍生出的一种图像特征表示方法，该方法在图像特征表示中尤为重要。作为一种较为通用的特征表示方法，最初被应用到图像分类任务中，在图像情感倾向性分析中也可以采用。词袋模型的原理是将文档看做相互独立的单词集合，应用到图像特征表示中，其构建步骤可以概括为：特征提取，视觉词典生成。

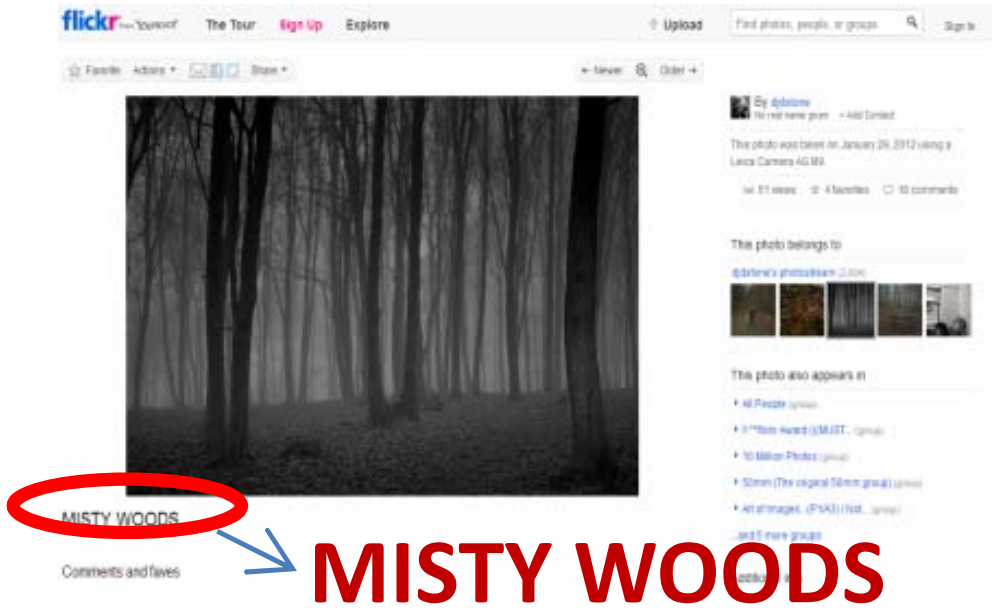
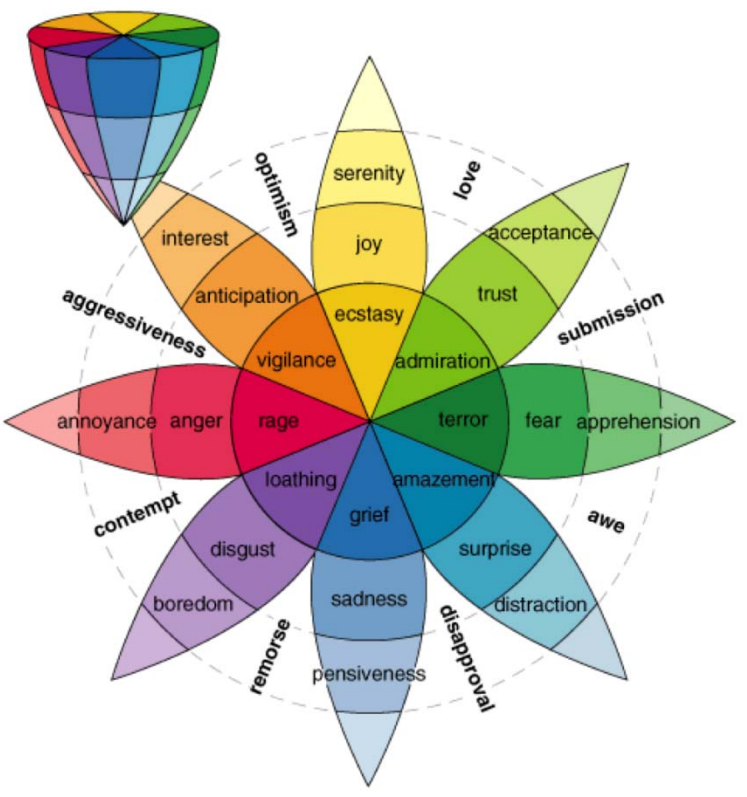
特征提取：特征提取是从图像块中抽取出具有区分性的全局或局部特征向量，对图像进行描述。主流的的特征提取算子为SIFT描述子[55]；

视觉词典生成：采用K-means算法对上一步中产生的单词(图像块特征向量)进行聚类，将每个聚类中心定义为一个视觉单词，所有视觉代词组成构成视觉词典，聚类中心数即为词典的大小。将图像表示成基于视觉单词的统计直方图。

### 2.3.4 SentiBank语义特征

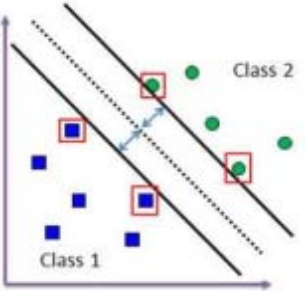
图像底层特征与图像高层情感语义之间存在着语义鸿沟，底层特征无法很好的表达图像的情感语义，因此分类效果不尽如人意。针对该问题，D Borth等人构建了一个较大规模视觉情感本体(Visual Sentiment Ontology，VSO)，在此基础上提出了一种视觉情感语义特征表示方法[45]。该方法相比于以往的底层特征，一定程度上解决了图像底层特征与高层语义直接的语义鸿沟问题，在图像情感分类上取得了不错的效果，在本篇文章中我们选用了该特征表示方法作为基础，在此方法上根据话题相关性得到图像最终的语义特征表示。

D Borth等人提出图像包含两层语义的观点，一层是客观存在的事物，比如图像中所包含的车辆、人、动物、花草等视觉对象，另外一层是主观存在的情感语义，比如图像所表现出来的漂亮的、开心的、阴沉的、悲伤的等情感，基于这两方面的语义，提出了能在语义层面很好的表达图像情感的特征表示：形容词名词对(Adjective Noun Pairs, ANP)，名词是图像所包含的客观对象语义，而形容词是图像的主观情感语义，两者的结合能够更加准确的表示图像的情感内容。在这个思想的引导下，作者通过Flickr的图像及其对应的标签资源构建了一个由形容词名词对(Adjective Noun Pairs, ANP)组成的视觉情感本体(Visual Sentiment Ontology, VSO)，在视觉情感本体的基础上训练出对应的ANP检测子构建成SentiBank，从而形成图像的情感语义中间层特征表示。视觉情感本体VSO和SentiBank检测子的构建流程如图2.5所示。

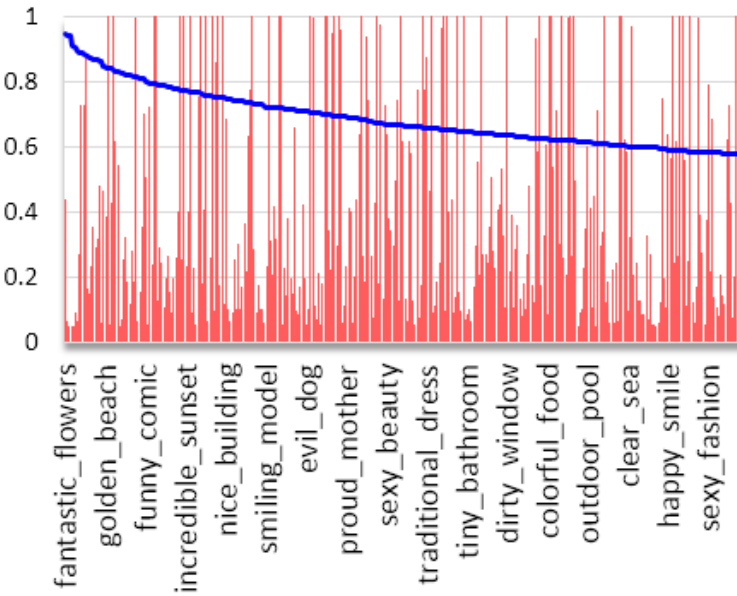


**Adj-Noun Pair construction**

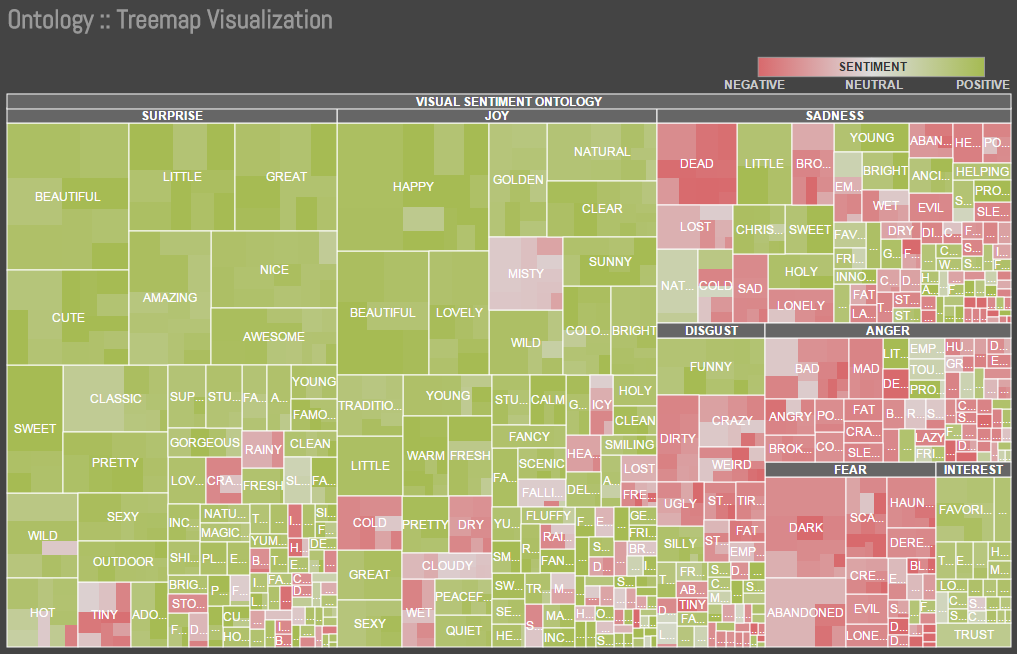
**Visual Sentiment Ontology**



**Train ANP Detectors**



**ANP Detectors Filtering**



**Colorful clouds**

**Broken tree**

**…**

**SentiBank (2089 ANP Detectors) Filtering**

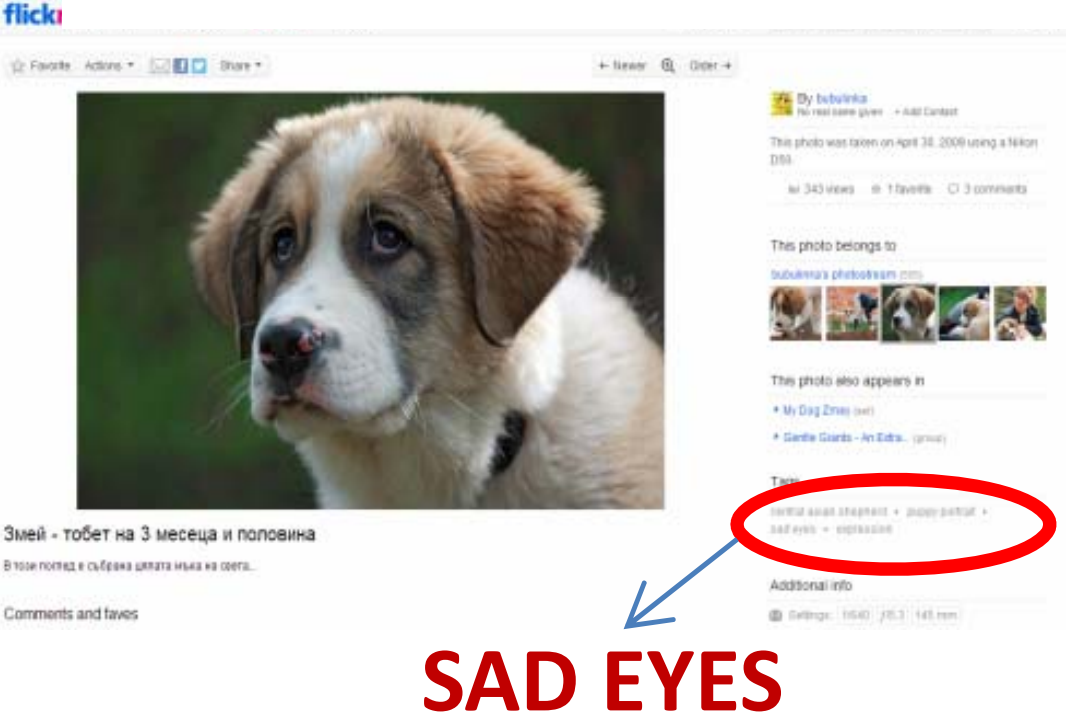


图2.5视觉情感本体(VSO)和SentiBank检测子构建流程

视觉情感本体VSO和SentiBank检测子构建的具体步骤如下：

* 第一步：形容词和名词的抽取。利用Plutchik心理学模型[54]中的24个基本情感词在Flickr和YouTube中检索标签中含有该情感词的图像和视频，抽取标签中与24个基本情感词共现的形容词和名词，并对词语进行去重和筛选，最终保留268个形容词(比如beautiful，sad)和1187个名词(比如people, flower)；
* 第二步：形容词名词对构建(ANP Construction)。将上一步获得的形容词和名词进行搭配得到候选ANP，并根据其出现频率等规则进行筛选；
* 第三步：视觉情感本体构建(Visual Sentiment Ontology Construction)。对上一步获得的每个ANP，在Flickr中检索对应的图片，并根据图片检索结果数量对ANP进行排序，选取出Top 3000个ANP构成Visual Sentiment Ontology(VSO，视觉情感本体)。其中每个ANP都有对应的极性值，且至少包含125张对应的图像。VSO所包含的ANP无论是情感覆盖性还是目标覆盖性都较为全面，比如happy smile(开心的笑容)、colorful clouds(多彩的云)、handsome guy(帅气的小伙)、 sad eyes(悲伤的眼神)、crying baby(哭泣的孩子)、dusty sky(阴霾的天空)；
* 第四步：训练ANP检测子。对每个ANP对应的图像集，抽取图像底层特征，包括颜色、LBP、Bag-of-Visual Words等特征，用LibSVM训练ANP检测子；
* 第五步：SentiBank构建。将ANP检测子根据其检测准确率进行排序，选取准确率在一定阈值上的ANP检测子构成情感检测库SentiBank，共包含2089个ANP检测子。

对于需要进行情感倾向性分析的图像，通过底层特征提取和SentiBank检测子检测，便可以获得2089维的图像ANP特征表示。

# 第三章 基于话题相关性模型的图像情感语义特征表示

## 3.1 引言

基于Sentibank检测子的图像ANP特征表示虽然取得了较好的图像分类效果，但是还是存在一些问题：

首先是ANP噪声问题：由SentiBank检测子得到的图像ANP特征存在一定的错误检测，这些错误检测会成为影响图像情感分类效果的噪声。SentiBank中的检测子无法做到百分之百检测正确，所以图像的中间层特征表示中会存在错误的检测响应值，这些错误检测就造成了特征中影响分类效果的噪声。比如表3.1和表3.2是微博上的两幅图像，其ANP检测结果按置信度降序排列后的Top 10个ANP如表4.1(a)和(b)所示，其中带‘？’的ANP表示检测结果是有争议的，带‘\*’的ANP属于错误的检测结果。可以看到特征检测结果中，大部分ANP是正确的，但是还是会有一部分检测错误，所以如何弱化这些错误检测造成的分类噪声，是提升分类效果的一个关键因素。

表3.1 微博图像ANP检测结果及其置信度（一）

|  |  |  |
| --- | --- | --- |
|  | ANP | 置信度 |
| Innocent angel | 0.985 |
| Cheerful face | 0.966 |
| Chubby face | 0.966 |
| Hungry baby？ | 0.965 |
| Yummy cake | 0.961 |
| Grumpy baby？ | 0.960 |
| Handsome smile | 0.957 |
| Excited face | 0.953 |
| Beautiful smile | 0.953 |
| Cute baby | 0.945 |

表3.2 微博图像ANP检测结果及其置信度（二）

|  |  |  |
| --- | --- | --- |
|  | ANP | 置信度 |
| Heavy clouds | 0.986 |
| Rainy city | 0.980 |
| Tranquil morning\* | 0.980 |
| Amazing city\* | 0.977 |
| Cloudy view | 0.976 |
| Stormy rain | 0.974 |
| Awesome view\* | 0.967 |
| Rainy bridge | 0.965 |
| Famous beach\* | 0.961 |
| Lonely train | 0.957 |

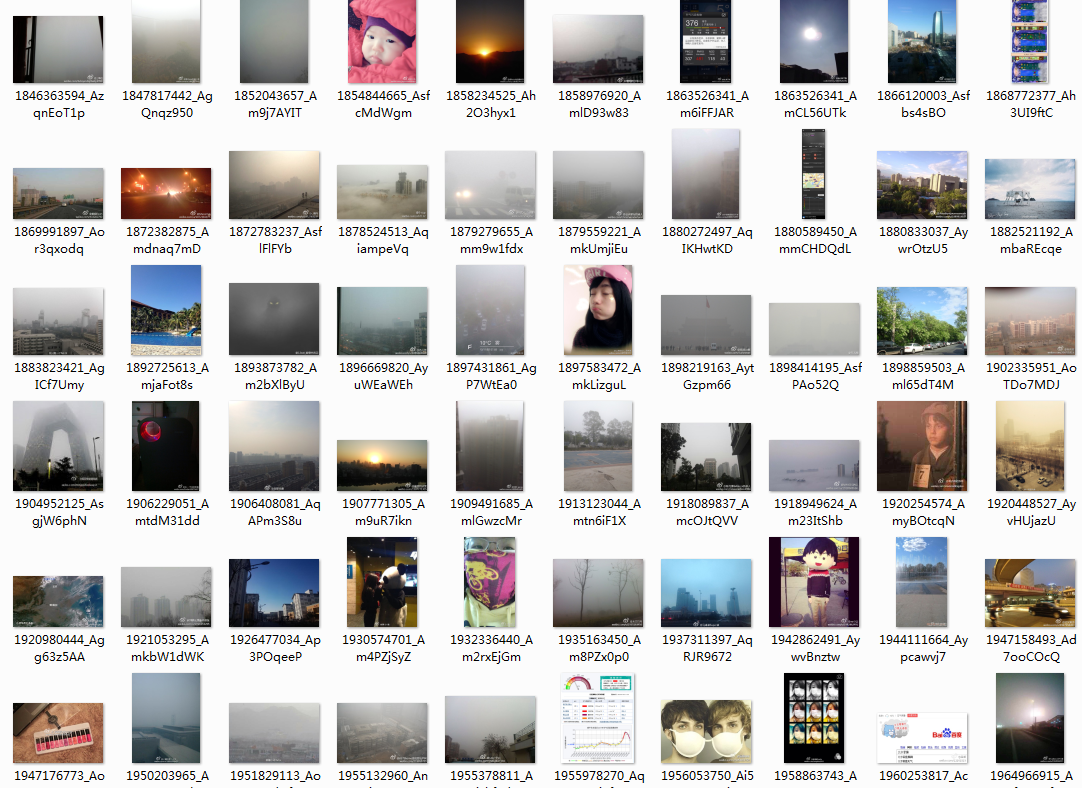
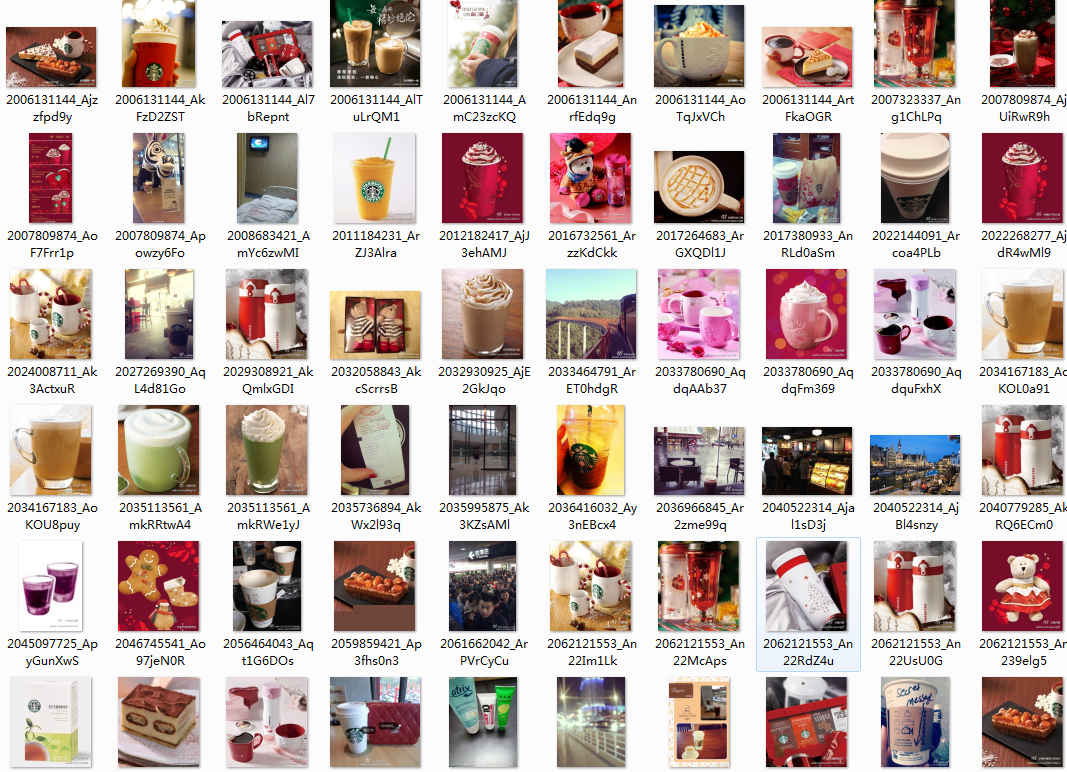
其次是ANP的区分度问题：SentiBank检测得到的众多ANP中，分类器无法判断哪个ANP表达了图像最主要的情感语义。如果按照话题分类，那么图像中与话题相关的视觉内容才包含了用户想要表达的情感语义，其他不相关的视觉内容则会对分类结果产生干扰。比如图3.1 (a)-(c) 所属话题为#雾霾#，用户主要想展示的内容是图像中的灰色天空，排放的烟尘等，这些视觉内容与话题相关，表达了用户的情感。而其他图像的内容，比如图中的建筑物，车辆，行人等便只是背景噪声。所以如何根据话题信息对图像中主要的视觉情感语义特征进行判别，减少背景噪声带来的干扰，是需要解决的问题。

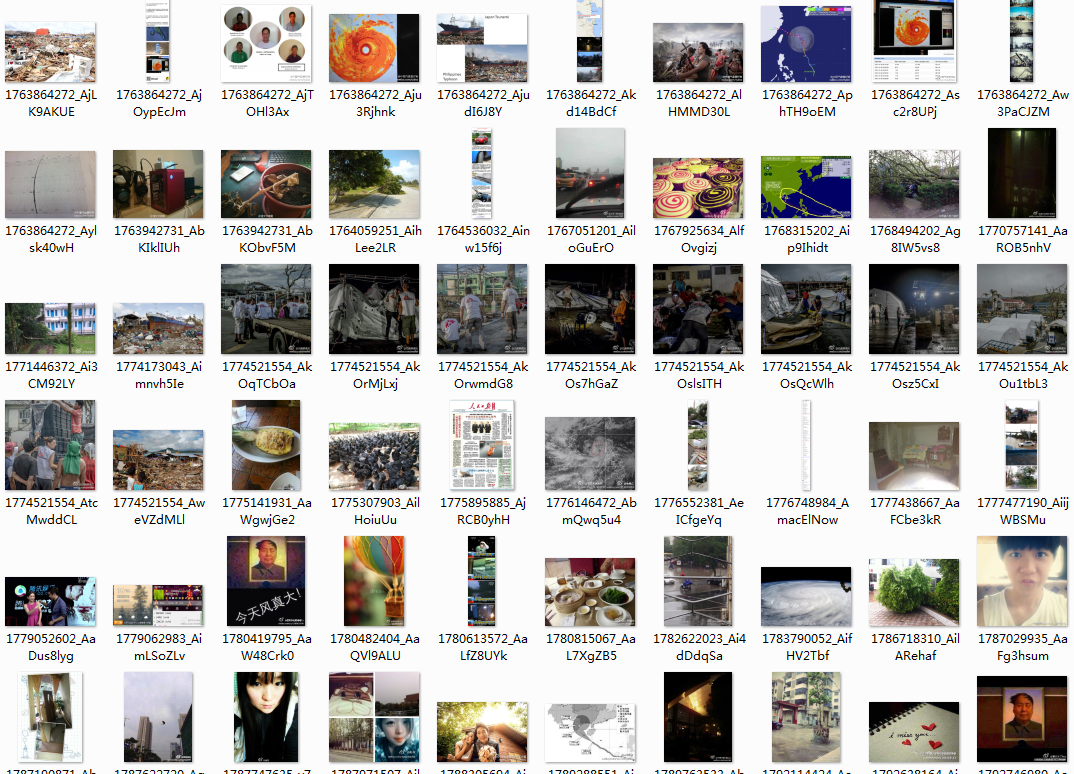
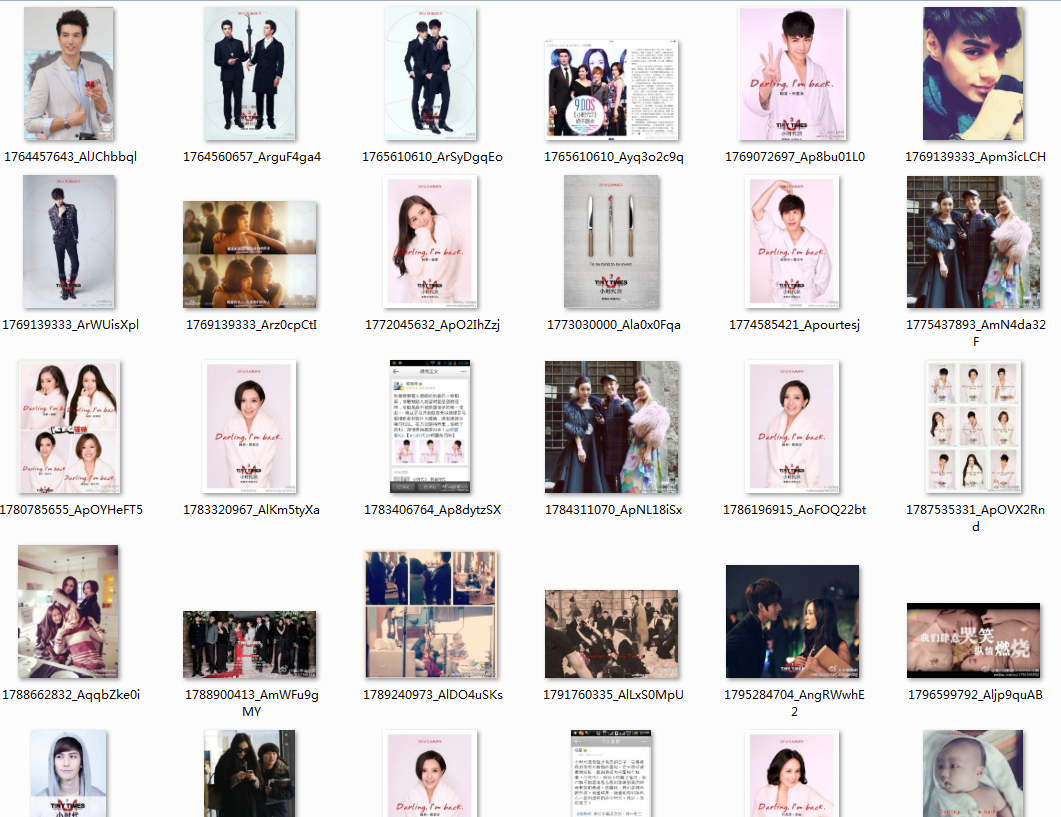
(a) (b) (c)

图3.1 微博话题#雾霾#图片样例

第三是图像情感的话题相关性问题：同一个话题下的图像之间有一定的相关性，但是基于VSO的SentiBank检测子无法利用话题中图像间的相关信息。通过观察发现，图像所包含的视觉情感内容在同一个话题下有一定的相似性，在不同话题间有一定的差异性。比如图3.2（a）-（d）是四个话题的部分图片展示，可以看到各个话题下的图片都有一定的共性，雾霾话题中主体的视觉情感内容是阴霾的天空，或者反向的湛蓝天空和灿烂阳光；星巴克话题中主要是可爱的杯子以及咖啡、蛋糕等；电影小时代3中的图片则主要是电影的海报、演员的剧照。所以如何充分利用话题所包含的潜在相关性信息，从而抽取出表达图像的主要视觉情感要素，对提升分类效果至关重要。

(a) 雾霾话题图片 (b) 星巴克话题图片

(c) 台风话题图片 (d) 小时代3话题

图3.2 微博上不同话题的图片实例

综上所述，为了解决ANP特征的噪声问题、区分度问题，充分利用话题内图像间的相关性信息，对主要表达图像情感语义的ANP进行判别，本文提出了基于话题相关性模型的图像情感语义特征表示，进而从话题的角度更好的理解图像情感分类任务，解决SentiBank特征存在的部分不足，提升图像情感分类的准确率。

## 3.2基于话题相关性模型的图像情感语义特征表示

为了充分利用话题下图像间的相关性，抽取出表现图像情感最重要的语义特征，并且过滤掉一些ANP错误检测造成的噪声，本文提出了话题相关性模型TRM(Topic Relevant Model)，从而对基于原始Visual Sentiment Ontology的SentiBank特征进行优化，得到能更好展示图像情感语义的特征表示。该模型可以根据概率模型得出视觉情感特征与话题的相关性，从而找出类别区分性高的特征，对于图像中检测错误的ANP、图像中不重要的背景内容等信息进行权值降低或过滤，从而更好的提升分类准确率。

### 3.2.1 话题相关性模型构建流程



SentiBank 检测子

图像ANP特征

主题相关特征

构建话题图像数据集

TM BM



**Step 1**

**Step 2**

**Step 3**

图3.3 话题相关性模型构建流程图

话题相关性模型TRM的构建流程如图3.3所示，具体步骤为：

* Step 1：构建话题图像数据集。抓取微博中的热门话题数据，选取合适数量的不同类别的话题，每个话题包含一定数量的图像，对图像进行标注得到话题数据集；
* Step 2：把属于同一个话题的图像进行聚合，使用SentiBank检测子对所有话题的图像进行特征提取(表示一个ANP检测子，m为检测子数量)，得到图像的ANP特征向量，即，其中表示检测子对输入图像*I*的检测响应值(response Probability)。
* Step 3：利用所有话题的视觉情感特征构建话题模型TM和背景模型BM，结合话题模型和背景模型得到话题相关性模型，最终抽取出每个话题中类别区分度高的特征。

### 3.2.2 基于话题模型和背景模型的相关性模型

前面我们介绍了TRM模型的大致流程，其中最重要的是构建相关性模型，而该模型是结合话题概率模型和背景概率模型得到的。首先我们对问题的一些基本定义进行说明：

对于包含t个Topic的数据集，其中每个Topic有其对应的图片集。图片使用SentiBank检测子得到的ANP特征空间为，每个ANP对应的特征检测响应值为。TRM模型的目标是发现特征与话题的相关性，从而找出有区分性的特征。因此，如果获得了话题模型参数，特征与话题的相关性便可以根据概率进行估计。所以概率模型的关键在于如何估计话题模型参数，本文采用最大似然估计来进行求解，根据最大似然估计的原理，该问题可以用公式表达为：

 (3-1)

假设每个图像情感特征满足多项式分布，那么公式(3-1)可以转化为：

 (3-2)

其中表示特征在话题中的情感响应值，n为话题包含的图像数量。在对公式2进行极值点求解之后，可以根据下面这个公式进行估计：

 (3-3)

虽然我们可以根据话题概率模型直接估计特征与话题的相关性，但是就如文本分类中的TF-IDF理论，在某个话题下概率高，在其他话题下概率低的特征才是最具类别区分性的，所以单独的话题概率模型在找到区分性高的特征的同时，还可能把类间区分性低的特征包含进来，所以有必要构建一个背景模型。与话题概率模型的构建相似，背景模型的公式为：

 (3-4)

基于话题模型T和背景模型B，最终话题相关性模型可以表示为：

 (3-5)

### 3.2.3 基于话题相关性模型的图像情感分类

在话题相关性模型构建完成后，图像情感分类流程为：

* Step 1: 利用SentiBank检测子获得训练图像和测试图像的SentiBank特征；
* Step 2: 根据话题相关性模型对SentiBank特征进行权值调整，从而挑选出具有话题区分性的ANP特征。假设图像所属话题为，则根据话题相关性模型得到图像最终的情感语义特征

；

* Step 3: 使用分类算法训练分类模型，对测试图像进行预测。

## 3.3 实验结果及分析

### 3.3.1 数据集构建

为了对话题相关性模型的性能进行测试评估，我们从新浪微博热门话题排行榜中抓取了29个话题共9289张图片，然后由项目组的3位学生进行标注根据投票方法决定图像最终的极性标签(正极性1,中性0,负极性-1)，图3.4分别给出了不同极性标注的样例，三张图所属的话题分布为：#雾霾#，#爸爸去哪儿#，#iOS7#。图片所属的话题分布情况及对应的图片数量如表3.3所示，图像分属不同极性类别的数量分布如表3.4所示。

(a) 标注：-1 (b) 标注：1 (c) 标注：0

图3.4 不同极性的图片样例

表3.3 实验数据的话题分布和图像数量信息

|  |  |  |  |
| --- | --- | --- | --- |
| 话题分类 | 话题数量 | 话题举例 | 图像数量 |
| 社会 | 3 | #雾霾#、#台风天兔# | 1660 |
| 体育 | 2 | #厦门马拉松# | 695 |
| 综艺娱乐 | 9 | #爸爸去哪儿#、#我是歌手# | 1735 |
| 影视 | 10 | #卑鄙的我#、#时代3# | 2562 |
| 科技 | 2 | #iOS7#、#魅族MX3# | 693 |
| 餐饮 | 1 | #星巴克# | 1471 |
| 游戏 | 1 | #节奏大师# | 293 |
| 书 | 1 | #破碎的时光# | 180 |

表3.4 实验图像数据极性分布情况

|  |  |
| --- | --- |
| 类别 | 数量 |
| 正极性 | 3939 |
| 负极性 | 1067 |
| 中性 | 4283 |

### 3.3.2 评测方法及评价指标

评测方法采用了五折交叉验证（five-fold cross-validation），将实验数据随机划分为五个大致相等的子集。分别进行五次训练和测试。每次四个子集作为训练数据，一个子集作为测试数据。

实验中图像底层特征的抽取和ANP中间层特征的抽取实验环境为Matlab，底层特征抽取函数使用vlfeat-0.9.16开源图像处理函数库。TRM模型的构建及最后的特征表示使用Java代码编写，实验环境为Eclipse。最终分类算法Logistic Regression、Naïve Bayes使用java环境下的weka中的分类函数，SVM分类算法采用Chih-Jen Lin的线性支持向量机Liblinear包。

评价指标采用了准确率、精确率、召回率和F-score四种方法：

准确率

精确率

召回率



上式中的，，，是混淆矩阵中的实例划分类型， (True Positive)表示被模型预测为正的正样本， (False Positive)是被模型预测为正的负样本， (False Negative)是被模型预测为负的正样本， (Ture Negative)是被模型预测为负的负样本。

### 3.3.3 图像情感分类效果

为了验证基于话题相关性模型TRM的有效性，在几个常用分类器中选出分类效果相对最优的分类器，并分析TRM模型在各个话题下的分类效果，同时由于所构建的数据集各个类别数据不均衡，所以还给出了均衡数据情况下的实验结果，因此本节采取了三个部分的实验进行说明：

* 实验第一部分：此部分实验通过对比基于TRM模型的特征与SentiBank特征的图像情感分类效果，验证话题相关性模型TRM的有效性。对比实验中采用了三种分类器：逻辑回归、朴素贝叶斯支持向量机，分析在各个分类算法下TRM模型是否都有效，同时选出分类效果最优的特征表示和分类算法。对比实验给出了正负极性2分类和正负中性3分类下两种特征的分类效果展示，说明TRM模型无论是2分类问题还是3分类问题都是有效的。
* 实验第二部分给出了数据集平衡时的分类效果，分析数据不平衡和数据平衡情况下实验效果是否有较大差异。
* 实验第三部分给出了TRM模型在各个话题下的实验效果，从而分析该模型对各个话题是否都能起到有效的提升作用。

**(1) 基于TRM模型的特征与SentiBank特征的对比**

为了验证话题相关性模型TRM的有效性，此部分实验选取了基于VSO的原始Sentibank检测特征作为比较基准，与本章提出的基于话题相关性模型的语义特征进行对比。实验选取了常用的朴素贝叶斯 (Naïve Bayes，NB)和逻辑回归 (Logistic Regression，LR)和支持向量机(SVM)三种分类器来获得分类效果，最终对比得到效果相对最优的分类器。正、负极性2分类的实验对比结果如表3.5和图3.5所示，2分类实验数据集共5006张图像，其中正极性图像3939张，负极性图像1067张)；正、负、中性3分类的结果如表3.6所示，3分类实验数据集共9419张图像，正极性图像为3939张，负极性图像为1067张，中性图像为4285张。

从图3.5中我们可以清楚的看到，三个分类算法中，基于TRM模型的特征都比原始SentiBank特征分类效果好，说明了TRM模型的优越性。而三个分类算法的效果对比发现SVM表现最佳，在TRM模型上的分类准确率从原始SentiBank的82.38%提升到90.47%，提升了8个百分点。这些实验结果都充分说明了基于TRM模型的特征表示对图像情感分类的效果有较大的提升。表3.6说明了在正、负、中三分类任务下话题相关性模型同样有很好的性能。表3.7给出了TRM模型特征表示的SVM情感分类各个类别结果。从表中我们可以看到正类和中性类别的F-Score略高于负类，这主要是由于微博中数据分布不够均衡导致，负类图像数量较少。

表3.5 图像情感2分类结果

|  |  |
| --- | --- |
| **方法** | **Accuracy** |
| SentiBank+Logistic Regression | 0.762 |
| TRM+Logistic Regression | **0.803** |
| SentiBank+ Naïve Bayes | 0.791 |
| TRM+ Naïve Bayes | **0.886** |
| SentiBank+SVM | 0.822 |
| **TRM+SVM** | **0.905** |

图3.5 TRM特征与SentiBank特征在各个分类算法下的2分类结果对比

表3.6 图像情感3分类结果

|  |  |
| --- | --- |
| **方法** | **Accuracy** |
| SentiBank+Logistic Regression | 0.57 |
| TRM+Logistic Regression | **0.64** |
| SentiBank+Naïve Bayes | 0.567 |
| TRM+NaïveBayes | **0.68** |
| SentiBank+SVM | 0.642 |
| **TRM+SVM** | **0.714** |

表3.7 基于TRM+SVM方法的情感分类各个类别结果

|  |  |  |  |
| --- | --- | --- | --- |
| **类别** | **Precision** | **Recall** | **F-Score** |
| 正类 | 0.72 | 0.737 | 0.728 |
| 中性 | 0.713 | 0.736 | 0.725 |
| 负类 | 0.684 | 0.537 | 0.602 |

**(2) 数据均衡实验展示**

从表3.4的数据分布可以看出，所构建的数据集在各个类别下是不均衡的，不平衡的数据分布有可能造成训练的分类器有偏。为了分析不均衡数据与均衡数据下的分类效果是否偏差较大，我们从数据集中选取了正负类等量的样例来做对比实验。实验共选取了2134张图像构成新的训练测试集进行实验，正类和负类的各有1076个样例。实验结果如表3.8所示，从结果中我们可以看到，在原始SentiBank特征下其分类准确率比5006张图像的整个数据集分类结果稍差，但是在话题相关模型表示的特征下其分类效果有了不错的提升，与全部数据集相比较相差不大。该结果说明无论是原始SentiBank特征，还是话题相关模型优化的特征，在不平衡的数据集上都能取得良好的分类效果。

表3.8 均衡数据下的图像分类效果

|  |  |
| --- | --- |
| **方法** | **Accuracy** |
| SentiBank+SVM | 0.735 |
| TRM+SVM | **0.896** |

**(3) TRM模型特征与SentiBank特征在各个话题下的分类效果对比**

我们对所有29个话题图像的情感分类效果进行了对比，如图3.6所示。可以看到在大部分的话题中，TRM模型的效果都比原始SentiBank好，表明了话题相关性模型的有效性。从图3.6中我们发现并非所有的话题在TRM模型下性能都是比较优的，通过对数据的分析发现，TRM性能表现较差的话题，其图像数量都很少，大约在10~50张之间，由于图像数量较少，特征与话题没有统计得到正确的相关性，反而形成了噪声，导致准确率相比原始SentiBank有所下降。所以在数据量相对充足的情况下，TRM模型可以拥有更好的性能。

图3.6 29个话题中图像情感极性2分类的分类效果对比

### 3.3.4 实验案例分析

为了说明为何基于话题相关性模型的特征表示对图像分类性能有较好的提升，下面将给出两部分案例分析，第一部分展示了TRM模型对于ANP的区分度作用，第二部分给出了模型对于图像情感倾向性分类的作用。

1. **TRM模型对于ANP的区分度作用**

图3.7是数据集中话题#雾霾#根据TRM模型得到的相关性特征按降序排列的曲线图，其中大于0的部分表示特征与话题正相关，值越大表示相关性越大，小于0 的部分表示特征与话题负相关，我们只选取与话题相关性为正的特征作为该话题内图像的特征。从图中可以看到相关性为正的特征不到900个。与话题相关性很高的特征有cloudy\_view(阴沉的景色), heavy\_clouds(厚重的云), misty\_trees(), dusty\_sky(灰尘的天空), wicked\_storm(恶劣的暴风雨)等，这些特征都与该话题的视觉内容有很强的联系，具有很好类别区分度，所以TRM模型选出了类别区分度高的ANP特征；而负相关的特征有strong\_body(强壮的身体), awesome\_cars(漂亮的车), lovely\_smile(可爱的笑容)等，TRM模型这些类别区分度低的特征进行了过滤。

图3.7 话题#雾霾#根据话题相关性模型得到的相关性特征降序排列

1. **TRM模型对图像情感倾向性分析的作用**

在表3.9中，我们给出了6张图片来展示基于话题相关性模型的特征与SentiBank特征的分类结果对比。图像1，图像2，图像3和图像4基于Sentibank特征的预测结果是错误的，但是基于TRM模型特征的预测结果是正确的。仔细观察可以发现，这四张图像在特定话题下所展现的主要情感视觉内容不一样，具有更多的语义息，图像1更加凸显的是灰色的环境，图像2是浑浊的积水，图像3是灿烂的阳光，在话题模型下这些更加有区分度的ANP特征可以被筛选出，使得分类结果正确。

图像5两种方法预测都是正确的，图像6两种方法预测结果都是错误的，图像6分类错误的原因是图像的ANP特征检测结果主要显示为adorable\_guy，innocent\_eyes，其情感语义特征在于眼神的细节，无论是SentiBank特征还是基于话题模型的特征都无法正确获取该视觉语义。

表格 3.9 基于话题相关性模型与SentiBank的分类结果对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 图像编号 | 图像 | 所属话题 | SentiBank预测结果 | TRM+SVM预测结果 | 正确标签 |
| 1 |  | 北京雾霾 | 正 | 负 | 负 |
| 2 |  | 台风天兔 | 正 | 负 | 负 |
| 3 |  | 等风来 | 负 | 正 | 正 |
| 4 |  | 李云迪钢琴梦 | 负 | 正 | 正 |
| 5 |  | iOS7 | 正 | 正 | 正 |
| 6 |  | 小时代3 | 正 | 正 | 负 |

## 3.4 本章小结

本章提出了基于话题相关性模型的图像情感语义特征表示方法，充分利用微博同一话题下图像间的相关性信息，对基于SentiBank的特征进行选取和调整，得到图像最终的情感语义特征表示。该方法一定程度上解决了基于SentiBank的ANP特征的检测噪声问题和ANP区分度问题，最终实验验证了基于话题相关性模型的图像情感特征表示的有效性，本文的特征表示相比原始SentiBank特征在基于SVM分类算法的图像正负极性二分类任务中，准确率从82.2%提升到90.5%，在正负中性三分类中，准确率从64.2%提升到71.4%，实验充分说明了话题相关性模型的有效性。该图像情感语义特征的有效性让结合文本和图像的微博情感倾向性分析成为可能。

# 第四章 基于文本和图像信息融合的情感倾向性分类模型

## 4.1 引言

以往的微博情感分析研究工作大都只集中在文本处理上，但是随着用户表达习惯的改变，微博中图像数量占据了越来越大的比例，很多用户都开始用文本和图像结合的方式来表达自己的观点和情感，比如图4.1(a)-(b)是两条包含文本和图像的微博消息，左图表示的是负极性，右图为正极性，但是可以看到消息中的文本信息不足，极性无法单独从文本中判断得出，而要综合文本和图像才能判别。所以在微博情感倾向性研究中，非常有必要将文本信息和图像信息结合处理。基于此，本文提出了基于文本和图像信息融合的情感倾向性分类模型，在一定程度上解决了微博情感倾向性分类中文本信息不足的问题。

本文基于文本和图像信息融合的情感分类模型分别包含基于特征层融合的跨媒体情感分类模型和基于决策层融合的跨媒体情感分类模型，模型的分类流程分别如图4.2和4.3所示。从图4.2和4.3中可以看到，模型的关键问题为三个：(1)文本的特征表示；(2)图像的特征表示；(3)文本特征和图像特征的融合方式。文本特征表示方面，本文在主流文本情感特征表示的基础上，采用了多信息结合的文本词袋特征表示，并利用特征选择算法选出了合适数量的特征作为最终的文本特征；图像特征采用了上一章中话题模型得到的话题相关性图像特征；特征融合方面，本文采用了特征层融合和决策层融合两种方法，对两种融合方法的情感分类效果进行对比。

今晚的柳州#台风海燕#



#魅族MX3# 微距功能是吃货专用。



(a) (b)

图4.1 包含文本和图像的微博消息示例

文本特征

抽取

文本图像特征融合

包含文本和图像的微博数据集

图像特征

抽取

分类器训练



图4.2 基于特征层融合的跨媒体情感分类模型流程

文本特征

抽取

文本情感

分类器训练

包含文本和图像的微博数据集

图像特征

抽取

图像情感

分类器训练

文本情感

分类结果

图像情感

分类结果

分类结果融合



图4.3 基于决策层融合的跨媒体情感分类模型流程

## 4.2 文本特征抽取

本文在主流文本特征表示方法的基础上，提出了多信息结合的文本词袋特征表示(Multi-Info Bag-of-Words, MIBOW)，并采用有效的特征选择算法进行特征选择，得到文本最终的特征表示。

### 4.2.1 多信息结合的文本词袋特征

中文情感倾向性分类中，词袋模型(或称向量空间模型、一元文法特征)是最为常用的特征表示方法之一，它将文档看作是由独立单词组成的集合。以往的微博情感倾向性分析研究表明，词袋特征的分类效果较好，本文采用词袋模型作为文本的特征表示，并将标点信息和表情信息作为“情感词汇”加入词袋，构建出多信息结合的文本词袋特征表示。特征抽取流程如图4.4所示，具体步骤如下：

* 分词：中文单词之间没有分隔，需要将字符串文本分成有意义单词片段，本文采用NLPIR[40]分词系统来进行分词，并加入情感词典优化分词效果；
* 去除无用信息：去掉停用词、代词、助词、量词等功能词，停用词词典为人工选取的20多个词，去掉话题标签##中的内容，去掉用户昵称；
* 构建词袋：对分词后留下的词语、标点符号进行统计，将互不相同的词语和特殊标点符号放入词袋，将训练数据集中的表情放入词袋，最终得到融合多种信息的词汇向量，特征权值采用布尔权值方法进行计算；
* 特征选择：使用特征选择算法对原始词袋特征进行特征选择，选取适当数量的特征作为最终的特征表示。

**特征选择**

**分词**

**过滤信息**

**构建词袋**

微博消息

这/r 电影/n 那么/r 好看/a ，/w 原来/b 编剧/n …

表情



文本



图4.4 基于词袋模型的多信息结合特征抽取流程

1. **加入用户词典的分词效果优化**

中文分词是词袋构建的基础步骤，分词效果的好坏直接影响到后续的词袋特征表示。NLPIR分词系统[40]是目前被广泛使用且分词效果良好的开源系统，但是该系统在情感短语和网络新词的识别上还有一定缺陷，比如词语“尼玛”在现在的网络词语中是一个负面的词语，但是NLPIR的分词结果是“尼/b 玛/n”，分词错误会直接影响后续的特征表示效果，所以有必要对分词系统的分词结果进行优化，一个简单有效的方法是在系统中加入用户自定义的词典。本文选取了“中文情感词汇本体库”中的词语作为用户词典加入分词系统。中文情感词汇本体库是大连理工大学信息检索研究室整理和标注的，共包含情感词27466个，除了以往常见的情感词外，还收录了近几年比较常出现的网络新词。

1. **多信息结合的词袋构建**

文本信息中表达情感的内容除了情感词汇、情感短语外，还会有标点符号来增强情感表达，同时微博文本中常常会表情信息，所以本文选取了词汇特征、标点特征和表情特征三种信息来构建文本情感词袋，最终词袋中包含了三方面的信息，如表4.1所示，具体描述如下：

词汇特征：在汉语中同一个词可以有不同的词性，同一个词在不同词性下可能会有不同的语义，所以本文将词和词性进行组合，形成词袋中的一个单词。

标点特征：中文表达中某些标点符号有增强情感表达强度的作用，比如多个感叹号、多个问号或者连续的句号，表达无奈的语气时会有用户喜欢用连续三个或以上数量的句号“。。。”来表达，所以本文将感叹号“！”、问号“？”、波浪号“~”、连续出现三次或以上的句号“。。。”、省略号“…”作为特征放入词袋。

表情特征：微博中表情是用户极为常用的情感表达方式。对198570条微博热门话题数据进行统计，发现含有表情的微博消息有58776条，占29.6%，可见表情被用户的使用频次较高。表情特征比较直观的反应了用户的情感倾向，是非常重要的微博情感特征。微博上的表情在网页中的格式是文本格式，比如表情的文本是[哈哈]，本文将一个表情当作文本中的一个词语放入词袋。表4.2给出了微博中最常用的部分表情展示。

表4.1 构成词袋的三种特征

|  |  |  |
| --- | --- | --- |
| 序号 | 词袋 | 描述 |
| 1 | 词汇特征 | 微博文本中分词后的词语 |
| 2 | 标点特征 | 感叹号，问号，波浪号，连续句号等 |
| 3 | 表情特征 | 微博语料中出现的表情 |

表4.2 微博中最常用的部分表情

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| [爱你] | [泪] | [哈哈] | [心] | [嘻嘻] | [偷笑] |
| [鼓掌] | [抓狂] | [花心] | [太开心] | [害羞] | [威武] |
| [怒] | [good] | [晕] | [挖鼻屎] | [衰] | [生病] |

### 4.2.2 特征选择

原始词袋构建后特征维度很大，需要利用特征选择算法选出具有类别区分力的特征，对特征进行降维。在文本分类中，信息增益算法和点互信息算法都是比较有效的算法，本文选用两种方法进行了实验对比，选取合适数量的特征作为最终的文本特征表示。

1. **基于信息增益的特征选择**

信息增益通过评估文档中特征项出现与否对分类性能的影响，来对特征的区分度进行排序。特征的信息增益计算公式为：



 (4-1)

特征项的信息增益值越大，则表示该特征越重要。为每个类别在训练集中出现的概率，为特征在训练数据中出现的概率，表示包含特征项的图像属于类别的概率，表示训练数据中特征不出现的概率，表示不包含特征项的文档属于类别的条件概率。

1. **基于点互信息的特征选择**

在特征选择任务中，点互信息通过衡量特征项和类别之间的依赖程度来对特征进行选择。如果相互依赖程度越大，则该特征项越重要。特征项与类别的互信息计算公式：

 (4-2)

如果PMI值越大，则特征与类别的相关性越大，该特征越重要。

图4.5 文本特征选择方法比较

利用两种特征选择算法，我们在所构建的微博数据集上进行了实验，结果如图4.5所示。从图中我们可以看到，信息增益方法表现出比较好的性能，且特征在2000维时达到了F-Score值性能达到最佳，所以将信息增益选取的前2000维特征作为最终的文本最终的词袋特征表示，n=2000。

## 4.3 图像特征抽取

根据上一章的话题相关性模型，图像情感语义特征抽取流程如图4.6所示：

底层特征抽取



SentiBank

检测子

TRM模型



图4.6 基于话题模型的图像情感语义特征抽取流程

## 4.4 结合文本和图像的跨媒体特征融合

主流的特征融合方法主要有两种，一种是特征层的融合，一种是决策层融合。本文采用了特征层融合和决策层融合两种方法对文本和图像两种媒体数据进行结合，从而对微博消息的情感倾向性进行预测，分析哪种融合方法的情感极性分类性能更好。

### 4.4.1 特征层融合模型

特征层融合最为典型的方法是将不同的特征进行串联，形成一个新的融合特征，然后采用分类算法对融合特征训练分类模型。本文中文本和图像两个模态的特征层融合采用了简单串联的形式。根据上述2000维文本词袋特征空间和2089维图像话题情感语义特征空间，对微博数据进行文本特征抽取和图像特征抽取，得到文本特征和图像特征，其中n=2000，m=2089。

1. 图像特征形式转化

特征的串联要考虑两种特征的形态是否相似，特征维度是否相当。本文中文本特征采用了布尔权值计算特征值，所以文本特征值，分别取值为0或1。图像特征是SentiBank中2089个ANP的检测置信度值向量，即图像特征值，是0到1之间的概率值。两种特征向量形式不同，需要进行转化。

由于图像特征值代表了图像对于特征项的检测置信度，置信度越高，则图像包含内容的概率越大。根据对图像的ANP特征向量进行统计，发现检测置信度在Top 10中的ANP集合中，图像大多会包含ANP对应的图像内容，基于该统计，我们将图像特征进行转化，置信度在Top 10中的特征项值置为1，其余置为0，形成新的图像特征向量，文本和图像的特征形式达到了统一。

1. 基于融合特征的情感极性分类模型(Feature-level Fusion Model，FFM)

将文本特征和转化后的图像特征进行融合，得到最终的图文特征表示。在得到新的融合特征向量后，本文采用了支持向量机分类算法训练分类模型，得到特征融合的分类结果，与决策层特征融合方法进行对比，分析得出最好的特征结合方式。

### 4.2.2 决策层融合模型

对包含文本和图像的微博消息进行观察发现，文本的情感极性、图像的情感极性以及消息的总极性存在着一定的规则。如果能挖掘出这些规则，则文本和图像两部分的决策融合便可以根据规则进行融合，从而达到更好的融合效果。

Apriori算法是挖掘关联规则的经典算法，它可以找出数据集中频繁出现的规则项，因此我们采用Apriori算法对微博数据集中文本极性、图像极性、总极性的关联规则进行分析。

1. **基于Apriori算法的文本图像极性组合规则挖掘**

Apriori的理论基础是：如果项集X是频繁集，那么它的所有非空子集也是频繁集。反言之，只有子集是频繁集，那么包含该子集的项集X才有可能是频繁集，所以Apriori算法采用自底向上的方法，从1-频繁集逐步得到k-频繁集。

采用Apriori对微博数据集的文本极性、图像极性、总极性进行规则发现的伪代码如下：

输入：文本极性、图像极性、总极性组成的微博标注集D，最小支持度min\_sup

输出：频繁集L

L1=find\_frequent\_1\_itemsets(D); //找出1-频繁集

for(k=2;Lk-1!=0;k++){

Ck=apriori\_gen(Lk-1); //产生k-候选频繁集

for each item in D{

Ct=subset(Ck,item); //Ct为Ck中被item包含的候选集集合

for each candidate c in Ct

c.count++;

}

Lk={cCk|c.count>=min\_sup};

}

return L=所有频繁项集

procedure apriori\_gen(Lk-1)

for each 项集I1  Lk-1

for each 项集I2  Lk-1 {

if(I1[1]= I2[1])&& (I1[k-2]= I2[k-2])&& (I1[k-1]= I2[k-1])

c= I1 join I2; //连接两个项集

if has\_infrequent\_itemset(c, Lk-1) //如果c的子集包含了非频繁项集

delete c; //删除不可能产生频繁集的候选

else add c to Ck

}

return Ck;

Procedure has\_infrequnet\_subset(c, Lk-1)

for each(k-1) subset s of c //对于c项集的每个k-1子集

if( s 不属于 Lk-1) return true;

return false;

根据Apriori算法挖掘得到的规则如表4.3所示，其中1表示正向极性，0表示中性，-1表示负向极性。

表4.3 文本图像极性组合规则

|  |  |  |
| --- | --- | --- |
| 序号 | 规则 | 置信度 |
| 1 | 文本=1 图像=1 ==> 总极性= 1  3=-1 756 conf:(1)  3. 1=0 2=0 1556 ==> 3=0 1549 conf:(1)  4. 1=1 4749 ==> 3=1 4724  总极性=1 | 1 |
| 2 | 文本=-1 图像=-1 ==> 总极性=-1 | 1 |
| 3 | 文本=0 图像=0 ==> 总极性= 0 | 1 |
| 4 | 文本=1 ==> 总极性= 1 | 0.99 |
| 5 | 文本=1 图像=0 ==> 总极性=1 | 0.99 |
| 6 | 文本=-1 图像=0 ==> 总极性=-1 | 0.99 |
| 7 | 文本=-1 ==> 总极性=-1 | 0.97 |
| 8 | 图像=1 ==> 总极性= 1 | 0.95 |
| 9 | 文本=0 图像=1 ==> 总极性= 1 | 0.91 |
| 10 | 图像=-1 ==> 总极性=-1 | 0.9 |
| 11 | 文本=0 图像=-1 ==> 总极性=-1 | 0.83 |

1. **基于规则的决策层特征融合模型**

本文提出基于规则的文本和图像特征决策层融合模型(Rule-based Decision-level Fuison Model，RDFM)，对微博消息进行情感极性分类，分类流程如图4.6所示，具体思路为：

训练过程：

对微博训练集中的样本分别抽取出文本词袋特征和图像话题情感语义特征；

采用SVM算法分别训练文本情感分类器和图像情感分类器。

测试过程：

对于测试样本，通过文本情感分类器和图像情感分类器得到文本分类结果和图像分类结果，然后采用Apriori算法得到的表4.3中的组合规则进行融合，得到微博消息最终的分类结果。

微博训练集

文本词袋

特征抽取

图像语义

特征抽取

文本情感

分类器训练

图像情感

分类器训练

文本分类

结果

图像分类

结果

文本图像极性组合规则

最终分类结果

微博消息

文本情感

分类模型

图像情感

分类模型

**训练过程**

**测试过程**

图4.6 基于规则的决策层特征融合模型(RDFM)情感极性分类流程

基于规则的信息融合方法优点是对文本和图像的分类结果根据有效的规则进行结合，在单独文本分类模型的基础上有效提升分类准确率；但是也存在缺点：文本和图像分类模型中任意一个模型分类错误，都会导致最终结果错误。考虑到该缺陷，在基于规则融合的基础上，本文提出了结合规则和二次分类的决策级融合模型(Rule-based and Secondary classification Decision-level Fusion Model，RSDFM)，在基于规则融合的基础上，对置信度低的样例采用二次分类模型得到分类结果。

1. **结合规则和二次分类的决策级融合模型(RSDFM)**

对基于规则的信息融合方法的实验结果进行观察，发现在测试样例中，最终结果分类正确的样例，其文本预测结果类别置信度和图像预测结果类别置信度均较高，而分类错误的文本类别置信度或图像类别置信度至少有一个较低。

针对该问题，对于文本分类和图像分类置信度低于一定阈值的样例，构建SVM分类模型对其进行二次分类修正。二次分类模型（Secondary classification Decision-level Fusion Model, SDFM）的训练流程如图4.7所示。

包含文本和图像的微博消息

抽取文本词袋特征

抽取图像话题语义特征

文本情感分类器

图像情感分类器







SVM分类器训练

二次分类模型

图4.7 二次分类模型训练流程

结合规则和二次分类修正的决策级融合模型的情感分类流程如图4.8所示，具体步骤如下：

Step 1： 特征抽取。对微博消息分别抽取文本词袋特征和图像话题情感语义特征；

Step 2：文本和图像分类结果。采用文本情感分类模型和图像情感分类模型分别得到文本分类结果和图像分类结果；

Step 3: 分类决策。对分类结果置信度根据设定的阈值进行判断，文本分类结果中最大类别置信度大于阈值并且图像分类结果中最大类别置信度大于阈值：

判断为真，则根据文本和图像分类结果，采用规则方法得到最终分类结果；

判断为假，则将文本和图像情感分类结果的置信度向量和连接为特征向量，利用二次分类模型得到最终分类结果。

其中文本置信度阈值和图像置信度阈值分别设置为0.72，其设置方法是采用部分测试样例进行基于规则方法的测试，对测试结果中分类正确样例的和分类错误样例的最大阈值进行统计，得出在该阈值以上的样例分类准确率高。

最终分类结果

**True**

包含文本和图像的微博消息

抽取文本词袋特征

抽取图像话题语义特征

文本情感分类模型

图像情感分类模型







二次分类模型

文本图像分类结果

最终分类结果

**False**



文本图像极性组合规则

图4.8 结合规则和二次分类的决策级融合模型情感极性分类流程

## 4.5 实验结果及分析

### 4.5.1 实验数据集构建

本文通过爬虫技术对微博每日Top 10个热门话题数据进行采集，抽取出微博消息的相关内容，包括消息的ID、文本、表情、图像等内容。人工选取合适数量话题的数据，对数据进行去重整理，然后通过项目小组人员进行人工标注得到数据集。标注内容为单独文本极性，单独图像极性，文本加图像极性，最终得到9230条微博数据，各个类别的数量分布如表4.4所示。数据存储于Mysql数据库中，部分内容展示如图4.9所示。

表4.4 跨媒体微博数据极性分布情况

|  |  |
| --- | --- |
| 类别 | 数量 |
| 正极性 | 5492 |
| 负极性 | 2067 |
| 中性 | 1671 |
| 总计 | 9230 |



图4.9 微博情感分类部分实验数据展示

### 4.5.2 实验设计

在实验中，我们首先对文本特征的有效性进行了说明，与谢丽星等人提出的基于情感词典的数值统计特征(dictionary-based feature)[17]进行了对比，证实了本文所采用的文本特征具有较好的优势。同时在本文所提出的特征上进行了分类器对比，证实SVM分类器效果最佳。

然后，为了验证基于文本和图像信息融合的情感倾向性分类模型的有效性，给出了融合模型与单独文本分类的对比实验效果。同时对特征层融合和决策层融合方法进行对比，分析哪种融合方式更能提升分类效果。

最后给出了实验效果的案例分析，更好的说明为何跨媒体模型相对单独文本分析模型效果更好。

### 4.5.3 实验结果展示

1. **文本特征分类效果实验对比**

我们从所构建的数据集中随机选取了1500条正向文本，1500条负向文本和1500条中性文本进行该部分的文本实验分类效果对比。实验采用了五折交叉验证的方式进行。本文提出的MIBOW文本特征与谢丽星等人提出的基于情感词典的特征实验对比结果如表4.5所示。

从表4.5中我们可以看到本文提出的文本词袋特征MIBOW无论是在正负极性2分类下还是在正负中性三分类下的分类效果都明显比统计数值统计特征好。

表4.5 文本情感倾向性分类准确率

方法  2分类 3分类

dictionary-based +SVM 0.7890 0.6507

**MIBOW+SVM 0.8240 0.6813**

同时我们还给出了MIBOW特征在不同分类算法下的分类准确率，如表4.6所示。从三种算法的分类结果中我们可以看到，SVM取得了最高的准确率。

表4.6 基于MIBOW特征的分类器效果对比(3分类)

Naïve Bayes Logistic Regression SVM

MIBOW 0.6458 0.6722 **0.6813**

上述实验结果表明，在文本情感倾向性分类中，基于MIBOW特征和SVM分类算法构建的文本分类器表现出了不错的实验效果，所以在跨媒体模型的对比实验中我们采用了基于MIBOW特征和SVM分类算法构建的分类器作为文本部分的分类模型，在此简称为单独文本分类模型(Text-Based Model，TBM)。

1. **跨媒体模型与单独文本分析的实验对比**

为了说明跨媒体决策层模型相比单独文本分析模型的有效性，我们给出了跨媒体决策层融合模型与单独文本分类的实验对比，验证了跨媒体特征融合模型相比传统的基于单独文本分析的微博情感倾向性分类，具有较好的提升效果。同时还对各种跨媒体特征融合方法进行了对比，得出最佳的特征融合方法。该模块选取了数据集中所有9230条微博消息作为实验数据，采用3折交叉验证的方式得到实验结果。以下实验均为正、负、中性三分类的实验结果。

表4.8决策层特征融合模型情感倾向性分类效果对比

|  |  |
| --- | --- |
| 方法 | 准确率 |
| SDFM(基于二次分类的决策融合) | 0.7239 |
| RDFM(基于规则的决策融合) | 0.7342 |
| **RSDFM**(结合规则和二次分类的决策融合) | **0.7502** |

表4.8中给出了三种决策层特征融合模型的情感倾向性分类效果对比，实验表明结合规则和二次分类的决策层融合模型RSDFM取得了最好的分类效果。

表4.9 单独文本分类模型和跨媒体分类模型的情感倾向性分类效果对比

|  |  |
| --- | --- |
| 方法 | 准确率 |
| TBM(单独文本分类模型) | 0.7108 |
| FFM(特征层融合模型) | 0.7261 |
| **RSDFM**(结合规则和二次分类的决策融合) | **0.7502** |

表4.9给出了单独文本分类模型TBM、跨媒体特征层融合模型FFM、跨媒体决策层融合模型RSDFM的情感倾向性分类结果对比。可以看到本文提出的跨媒体决策层融合模型和特征层融合模型相比单独文本模型分类准确率都有所提升，验证了跨媒体模型的有效性。同时也说明了决策层特征融合模型效果优于特征层融合模型。

表4.10 单独文本分类模型和跨媒体模型在各个情感类别下的效果对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | 方法 | Precision | Recall | F-Score |
| 正向情感消息 | TBM | 0.7708 | 0.8919 | 0.8269 |
| **RSDFM** | **0.7829** | **0.9132** | **0.8430** |
| 负向情感消息 | TBM | 0.6478 | 0.5312 | 0.5837 |
| **RSDFM** | **0.7519** | **0.5718** | **0.6496** |
| 中性情感消息 | TBM | 0.4784 | 0.3375 | 0.3958 |
| **RSDFM** | **0.5803** | **0.4345** | **0.4969** |

表4.10给出了结合规则和二次分类的决策融合模型RSDFM与单独文本分类模型TBM在正向、负向和中性三个情感类别下的精确率、召回率和F-Score值结果对比。从表中可以看出，在正向、负向和中性三种情感类别下，跨媒体模型相对单独文本模型，分类性能均有所提升。

图4.10各个话题下跨媒体模型与单独文本模型的分类效果对比

从图4.10中可以看到，跨媒体模型在各个类别话题下相比单独文本模型的情感倾向性分类提升效果不一。社会类、综艺类、影视类和餐饮类的微博消息分类提升效果明显，分析发现这三个类别话题下的图像内容对文本有很好的补充作用，且这些话题类别的图像情感分析准确性较高，比如社会类话题中的#雾霾天气#话题和#台风天兔#话题，有较多用户通过简短的文字或者直接贴图的方式来表达自己的情感，如图4.11(a)所给出的社会类微博消息示例，图像的情感极性分类结果为负，对文本起了补充作用。

游戏类和科技类话题的跨媒体模型性能相比单独文本反而有所下降，原因在于这两类话题下图像的语义在计算机中比较难理解，比如图4.11(b)的微博消息，用户通过图片展示了文本中所提到的系统更新缺陷，但是这张图像的语义目前计算机还无法解析，所以加入图像分析后反而形成分类噪声，性能有所下降。



#北京雾霾#北京又开始抽烟了

更新后第一次bug，主屏幕不显示信号时间电量等信息了



(a) 社会类话题微博消息示例 (b) 科技类话题微博消息示例

图4.11 话题类微博消息示例

### 4.5.4 实验案例分析

下面给出一些微博消息在单独文本模型TBM和跨媒体模型RSDFM下的情感倾向性分类结果对比，说明为何跨媒体决策模型相比传统的单独文本分析更能提升分类准确率。

表4.11 跨媒体模型分类正确文本分类错误的微博消息示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 消息编号 | 微博消息 | | 单独文本模型TBM | 跨媒体模型RSDFM | 正确标签 |
| 文本 | 图像 |
| 1-1 | 今晚的柳州#台风海燕# |  | 中 | 负 | 负 |
| 1-2 | 今天出门忘带GPS了怎办？ #北京雾霾# |  | 中 | 负 | 负 |
| 1-3 | #星巴克# 解压解压 |  | 中 | 正 | 正 |
| 1-4 | #魅族MX3# 微距功能是吃货专用。 |  | 中 | 正 | 正 |

表4-11给出了单独文本模型分类错误，但是跨媒体决策模型分类正确的四条微博消息。从这四条消息中可以看到，只根据微博的文本信息无法判断出用户想要表达的情感倾向性，其想要表达的内容主要在图像中得以体现，所以单独文本模型得到了错误的分析结果，而跨媒体模型结合了文本和图像信息后，得到了准确的分类结果。

表4.12跨媒体模型和文本分类都正确的微博消息示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 消息编号 | 微博消息 | | 单独文本模型TBM | 跨媒体模型RSDFM | 正确标签 |
| 文本 | 图像 |
| 2-1 | #小时代3# 期待你们的华丽回归~ |  | 正 | 正 | 正 |
| 2-2 | #天兔#已经跳走这么多天了，可是还到处是垃圾堆成小山 |  | 负 | 负 | 负 |
| 2-3 | 重点在哪里？#星巴克# |  | 中 | 中 | 中 |

表4.12给出了跨媒体模型和单独文本模型都分类正确的微博消息示例。这些微博消息中，文本信息内容便足以判断出其情感倾向性类别。

表4.13跨媒体模型和文本分类都分错的微博消息示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 消息编号 | 消息内容 | | 单独文本模型TBM | 跨媒体模型RSDFM | 正确标签 |
| 文本 | 图像 |
| 3-1 | 就说这两天电话都没响过，原来是这么回事，这是要演哪出？#iOS7# |  | 中 | 正 | 负 |
| 3-2 | 壁纸清晰度未优化好。音乐横屏点开专辑图片模糊。好用的只有上拉功能栏。 |  | 正 | 正 | 负 |
| 3-3 | #小时代3#好期待小时代3，虽然寒假还没放呢，我能不能说一句“暑假快点到” |  | 负 | 负 | 正 |

表4-11给出了跨媒体分类模型和单独文本分类模型都错分的微博消息示例，

可以看到第一个示例中文本使用疑问结合图像展示的方式表达了负向情感，但是图像表达的语义使用目前的技术无法分析得到；第二个示例中，文本表达了多个观点，但是由于出现了多个正极性词汇，该条消息的跨媒体模型分析结果是以文本为主，所以两个模型都分类错误；第三条消息中用户对电影表达了多种情感，需要进行主题相关的情感识别才能正确判断。

## 4.6 本章小结

本章提出了结合文本和图像的跨媒体情感倾向性分类模型，该模型是微博情感倾向性分析研究的一个全新尝试，顺应了微博用户表达方式从以往的文本为主转变为如今的图文结合的潮流。该模型在一定程度上解决了微博文本信息不足的问题，提升了情感倾向性的分类准确率。同时，在基于特征层融合和基于决策层融合的文本图像特征融合方法中，发现基于决策层融合的方法效果更佳，合适的特征层融合方法还有待探索。

结合文本和图像的微博情感分析还存在着一个问题：文本信息和图像信息融合的跨媒体模型并不是在所有情况下都适用，因为图像情感倾向性分析本身是一个比较复杂的任务，对某些类别的话题图像分析效果好，而语义信息过于细节和丰富的图像则很难得到其正确的分析结果，这样的图像与文本结合后反而会形成噪声，造成分类性能的下降。所以如何构建出不同话题或不同领域的跨媒体模型，让两个模态信息的结合最大程度的提升情感倾向性分析的效果，是未来需要探究的内容。

# 第五章 总结与展望

## 5.1 工作总结

自2009年新浪推出中文微博服务以来，基于微博的社交便迅速蔓延到广大中国网民中，每日都会有大量用户会在微博中发表自己对事物的观点或者心情感受，对这些数据进行情感倾向性分析，有助于政府对舆情的了解和把握、或者公司产品的反馈分析，意义重大。

现在用户的情感表达方式从传统的文本为主转化为图文结合，微博中图像数据越来越多，而文本存在着信息不足的问题。针对该特性，本文创新性的进行了文本和图像结合的跨媒体情感极性分类相关研究，主要工作总结如下：

1. 针对现有基于SentiBank的图像情感特征ANP的检测噪声问题、区分度问题和情感话题相关性问题，本文提出了基于话题相关性模型的图像情感语义特征表示。话题相关性模型充分利用微博中同一话题下图像间的相关性和不同话题下图像间的差异性信息，对原始SentiBank特征进行选择和优化，一定程度上解决了ANP特征的检测噪声问题和ANP区分度问题，最终实验验证了基于话题相关性模型的图像情感特征表示的有效性。在图像正负中3分类问题上，相比SentiBank特征准确率提升了7.3%。该图像情感语义特征的有效性让结合文本和图像的微博情感倾向性分析成为可能。
2. 针对微博中短文本信息不足和用户表达方式转变的问题，本文提出了基于文本和图像信息融合的情感倾向性分类模型。该模型在文本词袋特征和图像情感语义特征表示的基础上，采用了特征层融合和决策层融合两种方法对文本特征和图像特征进行结合。通过图像和文本两部分信息的融合解决微博中表达方式从单文本到图文结合的转变问题，同时图像信息在一定程度上弥补了文本信息的不足。实验表明本文提出的融合模型在微博情感倾向性分类任务中，相比传统的单文本模型，分类准确率有所提升，在正负中性三分类问题中，准确率达到75.02%。同时对比了特征层融合和决策层融合两种方法，实验表明目前采用决策层融合方法可以获得相对较高的准确率。

## 5.2 工作展望

尽管本文验证了文本和图像结合的跨媒体模型在微博情感倾向性分析任务中的有效性，但是还存在着一些需要完善的问题，未来的研究工作也可以从这些方面进行开展，具体问题概括如下：

1. 微博实验数据的构建问题。本文的实验数据来自于课题组抓取的新浪微博热门话题，在数据的构建过程中发现微博数据很多为娱乐性话题数据，分析意义不大，同时还有很多广告消息，对这些消息的过滤耗费了较大的人力。在今后的研究中，如何自动选取有分析意义的微博数据，去除垃圾消息，构建出一定规模的可供广大研究者共享的微博数据集，对微博情感分析相关研究尤为重要。
2. 缺乏中文微博图像情感倾向性分析数据集。目前图像情感倾向性研究还处于起步阶段，且大部分研究工作集中在国外的Twitter、Flickr、Pinterest等网站的图片数据上。本文的图像情感特征基于D Borth等人的Flickr数据集，在应用到中文微博上存在一定的领域适应性问题。要对微博图像情感倾向性分析进行深入研究，较大规模的微博图像情感分析数据集是非常必要的。
3. 文本图像特征层融合方法。本文中文本和图像的特征层融合方法效果相比决策层融合方法差，原因是文本和图像属于两个不同模态的信息，采用典型的简单串联法无法达到很好的效果。所以探索合适的不同模态特征融合方法，比如采用特征相关性分析，将两个模态特征映射到新的特征空间中，进行信息互补，是今后可以研究的问题。
4. 缺乏话题分类模型。本文的话题是基于微博中的热门话题标签进行分类，但是微博中存在着很多非话题数据，如何对这些非话题数据采用Latent Dirichlet Allocation等潜在语义分析模型进行话题分类，从而得到话题相关性特征，是未来需要完善的工作。
5. 跨媒体模型的话题适用问题。文本和图像融合的跨媒体模型并不是在所有情况下都能取得很好的分类效果，因为图像情感倾向性分析本身是一个比较复杂的任务，对某些类别的话题图像分析效果好，而语义信息过于细节和丰富的图像则很难得到其正确的分析结果，这样的图像与文本结合后反而会形成噪声，造成分类性能的下降。所以如何构建出不同话题或不同领域的跨媒体模型，让两个模态的信息结合最大程度的提升情感倾向性分析的效果，也是未来需要探究的内容。

# 参考文献

1. Brendan J, Subhabrata B, Shih-Fu C. Predicting Viewer Perceived Emotions in Animated GIFs, ACM MM’14, November 3-7, 2014, Orlando, Florida, USA

# 硕士期间发表论文及参与项目情况

1. Min Wang, Donglin Cao, Lingxiao Li, Shaozi li, Rongrong Ji. Microblog Sentiment Analysis Based on Cross-media Bag-of-words Model. Proceedings of the 6th International Conference on Internet Multimedia Computing and Service,2014.
2. Yunhang Shen, Rongrong Ji, Donglin Cao, Min Wang, Shaozi li. Hacking Chinese touclick CAPTCHA by Multi-Scale Corner Structure Model with Fast Pattern Matching. ACM International Conference on Multimedia 2014.
3. 参与厦门大学智能多媒体实验室-微博跨媒体舆情分析系统项目。

# 致谢

不经意间，研究生时光就这样步入尾声，回首三年前入学时，恍如昨日。还记得那时对未来没有规划的自己，对学术研究没有概念的自己，对自身不够自信的自己，在这三年时光的洗礼后，总算有所改变，有所醒悟，有所收获。这段时间所取得的成绩和收获的成长，离不开众多老师和同学的指导、帮助、支持和鼓励，在此表达我无限的感激之情。

首先，我要由衷的感谢我的恩师曹冬林老师。从大学的课程老师、参赛项目导师、毕业设计导师，到研究生阶段的指导老师，曹老师对我的学习和生活产生了至关重要的影响，给予了我莫大的帮助和启迪。是曹老师持续不断的耐心指导和鼓励支持，让天资不够聪颖的我、基础比较薄弱的我可以顺利完成学业，并且取得让自己还算满意的成绩。曹老师认真严谨的治学态度、一丝不苟的工作作风、对学生的无私付出，让我非常敬佩且深受感染。曹老师在这些年中对我的指导与支持，让我不断进取，收获良多。在此谨向曹老师表达我最崇高的敬意和最诚挚的感激。

同时，我还要感谢李绍滋老师、纪荣嵘老师、郭锋老师、苏松志老师、林贤明老师、林达真老师在这些年中对我的指导和帮助。李绍滋老师渊博的学识和严谨的作风让我们在松懈时能够有所警醒，鞭策自己不断追求上进，培养良好的学习生活习惯。纪荣嵘老师在学术研究中给予了我很多指引和启示，让我认识到自己很多的不足，扩宽知识面的同时锻炼了实践创新能力。感谢所有老师，你们让我能够有更多的提升与收获。

感谢智能多媒体实验室的师兄师姐、师弟师妹们，是你们让这个实验室充满了乐趣与欢笑，让科研不再那么艰难与枯燥。感谢雅萍、力玄、董洁、敖翔、德东、伟源，我们七人组一起学习、娱乐，共同度过了无数的快乐时光，共同熬过了学习生活中的艰苦迷茫期。感谢舆情分析小组的同学们，大家在项目中相互配合，共同进步。感谢我的室友雅萍和力玄，一起度过的七年时光，是如此的难得和珍贵。特别感谢雅萍同学，在我痛苦迷茫时，给了我很多支持和鼓励。

感谢2012级智能硕士班的全体同学，研究生生活因为有你们而变得更加精彩和温暖。感谢学院的郑辉阁老师，刘晓峰老师，韩海雄老师，教会了我很多做人做事的道理。

最后，特别感谢我的爸爸妈妈和弟弟妹妹，因为有你们，无论学习生活遇到什么困难，我都可以更加坦然和勇敢的面对。你们的关心和牵挂，赐予我无限动力，你们的支持和理解，让我可以更加安心的做好自己的事。谢谢你们给了我一个温馨和谐的家，希望你们一直健康快乐。

最美的七年，是在厦门大学，成长最多的三年，亦是在厦门大学，这里承载了太多的回忆，感恩，感谢！愿母校、所有我感激的人、自己都有一个越来越好的明天。