# 摘要

随着计算机和网络技术的不断深入发展，如今的互联网已经迈进了社交媒体（social media）时代。一大批带有SNS（social network service）性质的网站、工具和产品，比如Facebook、Twitter、微博等等，正成为互联网的新鲜力量，它们担负起实现真实社会与虚拟空间无缝连接的重大使命。用户既可以在社交媒体上对新闻时事、社会现实、消费产品等话题发表客观或主观的看法，又可以把身边的实时见闻、个人的生活琐事等发表在社交媒体上面。深入挖掘社交媒体中潜在的有用信息，对热点事件发现、舆情监控、商品口碑反馈等方面有重要应用。

本文主要处理面向社交媒体中的热点发现及对热点相关文本的情感分类问题，其中，热点包含突发事件、新产品发布等情况，文本情感分析包含消息级文本情感分类和评价对象明确的文本情感分类。本文的前两章对该问题的研究现状和基本技术进行了详细的介绍，然后针对现有研究的不足之处，在第三到五章分别针对热点发现任务、消息级情感分类任务和评价对象明确的情感分类任务提出了解决方法：

1. 面向热点发现任务，提出了基于融合特征和绝对聚类的在线热点发现算法。在对热点相关文本的表示中，本文使用规则方法和机器学习算法进行热点词汇的发现与选择，然后通过融合特征来对文本进行向量化表示。同时，本文提出基于绝对聚类的在线热点发现算法，能够及时有效的发现社交媒体中出现的热点。
2. 面向消息级文本情感分类任务，针对文本情感表达中的情感偏移现象和含蓄表达问题，提出了基于反转算法的对偶模型，提高了消息级文本情感分类的准确率。首先提出了一种样本情感反转算法，对原始样本构造情感反转样本；然后本文提出了对偶训练模型，来对原始样本和反转样本联合建模学习；最后通过集成算法，对原始模型、对偶模型进行综合考虑来获取更好的效果。
3. 面向评价对象明确的文本情感分类任务，针对评价对象中往往包含多个词的现象，提出了评价对象独立表示的双边注意力神经网络模型，在标准数据集上取得了比较大的效果提升。

关键词： 社交媒体，热点发现，情感分析，对偶模型，神经网络

# Abstract

With the deepening development of computer and network technology, the Internet has entered the era of social media. A large number of websites, tools and products with SNS (social network service) features, such as Facebook, Twitter, Weibo and so on, are becoming the fresh force of the Internet, and they are shouldering the major mission of realizing a seamless connection between the real society and the virtual space. Users can either objectively or subjectively view social media issues such as news and events, social realities, consumer products, etc., and publish on social media such as real-time information and personal life experiences. In-depth mining of potentially useful information in social media has important applications in the discovery of hot events, public opinion monitoring, product reputation feedback and so on.

This article mainly deals with hotspots found and sentiment classification of hot-related texts in social media. Among them, hotspots include emergencies and new product launches, etc. Text sentiment analysis includes sentiment classification of message-level and target-level. In the first two chapters of this paper, the related and basic technology about this problem are introduced in detail. Then for the shortcomings of existing research, in the third to the fifth chapters, we put solutions about the hot spot discovery task, the message-level sentiment classification task and target-level sentiment classification task:

1. For hot spot discovery tasks, an online hot spot discovery algorithm based on fusion features and absolute clustering is proposed. In the representation of hot-related texts, this paper uses ruled methods and machine learning algorithms to find and select hot-topic words, and then uses the fusion features to represent the texts. Then this paper proposes an online hot spot discovery algorithm based on absolute clustering, which can discover the hot spots in social media timely and effectively.
2. For message-level textual sentiment classification task, this paper proposes a dual model based on reversal algorithm and improves the accuracy of message-level textual sentiment analysis, aiming at the problem of sentiment shift and implicit expression in textual sentiment expression. Firstly, a sentiment reversed algorithm of sample is proposed to construct sentiment reversed sample for the original sample. Then a dual training model is proposed to model the original and reversed samples jointly. Finally, by using the integrated algorithm， we consider the original model and dual model to get better results.
3. For target-level sentiment classification task, in view of the phenomenon that target often contains multiple words, a bilateral attention neural network model is proposed and a large improvement is achieved on the standard data set.

**Keywords:** social media, hotspot discovery, sentiment analysis, dual model, neural network

# 1 绪论

## 课题研究背景及意义

随着计算机和网络技术的不断深入发展，如今的互联网已经迈进了社交媒体（social media）时代。一大批带有SNS（social network service）性质的网站、工具和产品，比如Facebook、Twitter、微博等等，正成为互联网的新鲜力量，它们担负起实现真实社会与虚拟空间无缝连接的重大使命。用户既可以在社交媒体上对新闻时事、社会现实、消费产品等话题发表客观或主观的看法，又可以把身边的实时见闻、个人的生活琐事等发表在社交媒体上面。这些客观或主观信息具有巨大的潜在应用价值：（1） 重要的社会价值：对社交媒体中反应的事件或热点进行自动的检测跟踪、信息抽取和情感分析，可以监控引导社会的舆情走向；（2）显著的商业价值：检测跟踪社交媒体中对某些公司或某些产品的热点事件或热点产品的评价，并对评价进行情感分析等分析，对于企业正向引导产品评价、提高产品有重要的作用。然而这些文本信息数量庞大，仅仅依靠人力难以应对和处理爆炸性增长的评论信息，迫切需要运用计算机快速获取和整理这些信息，基于数据统计与机器学习的方法应运而生。

## 社交媒体热点发现研究现状

本论文中的热点发现任务主要是做热点事件的检测，既包含政府更加关注的舆情事件检测，又包含企业更加关注的特定商品的大规模评论发掘。

热点发现主要对社交媒体文本数据流中的突发话题进行检测。与传统话题检测相似，根据突发特征的识别顺序，突发话题检测方法可分为以文本为中心和以突发特征为中心。前者是先进行文本聚类，再在类中抽取出突发特征，进行突发事件的识别; 后者是先抽取出突发特征，再对突发特征进行分组，使用突发特征组进行突发事件的识别。然而，社交媒体的语言不规范等问题给这两种方法带来了挑战。此外，结合社交媒体中的特殊信息，突发事件检测的以特征为中心可进一步分为以文本特征为中心和以非文本特征（如用户关系）为中心的方法。

以文本为中心的方法先将社交媒体文本按照发布时间划入不同的单位窗格内，并对单个窗格内的微博文本进行聚类，一个类代表一个突发事件，再从每个类中抽取突发特征，用突发特征表示突发事件，以达到突发事件检测的目的。 Diao等[]通过观察发现两个规律，相同时间发布的消息更有可能拥有相同的话题，同一个人发布的消息（posts）往往更有可能针对同一个话题。他们认为前一个规律有助于帮助我们发现驱动话题发展的重要消息（posts）,后一个规律有利于过滤掉个人消息（"personal" posts）。于是他们将这两规律同时融入话题模型，从而解决社交媒体文本中的噪声问题。邱云飞等[]针对社交媒体文本语言不规范问题，提出一种基于动态滑动窗口的微博突发话题检测方法。他们利用窗口提取具有潜在突发性的信息，采用TD-IDF以及空间向量模型完成文本表示，最后使用改进的Single-Pass 聚类算法完成话题检测。童薇等[]综合考虑了微博数据的文本特征（转帖、评论、内嵌链接、用户标签hashtag、命名实体等）、语义特征、时序特性和社交关系特性，提出了一种有效的基于微博数据的事件检测算法（event detection in microblogs，EDM）。还提出了一种通过提取事件关键要素，即关键词、命名实体、发帖时间和用户情感倾向性，构成事件摘要的方法。以文本为中心的方法针对传统的文本，效果较好，但是社交媒体文本中含有很多垃圾信息，先进行聚类再进行突发特征识别会引入很多噪声信息，并且在文本聚类时还存在较多阈值的选取问题，阈值选取大多是根据经验值来设定，对聚类结果比较敏感。

以特征为中心的方法可分为以文本特征为中心、以非文本特征（如提及、粉丝）为中心。Kleinberg等[]首次采用基于突发关键词的方法进行突发话题检测，他们通过自动机模型模拟特征词的状态以及转台之间的转换，不同状态表示词的不同的出现频率，状态间的转换表示突发的产生或者消亡。宋丹等[]通过对用户查询的分析在用户产生的数据（User Generated Content, UGC）流上检测事件。该方法的基本思想是：首先，从查询日志中找出具有突发性的查询词。然后，使用这些突发词的查询结果构建事件。该文章使用查询日志和博客数据进行了实验并且取得了较好的事件检测效果。郑斐然等人[]通过词频和词语增长速度构造复合权值，在增量式聚类算法中融合上下文相关度模型，有效地从大量微博数据中检测出新闻话题。一些研究者们通过融入社交媒体特殊信息，提高社交媒体中突发事件检测的性能。Du等[]在统计术语权重时考虑了用户权重以及文本的回复数量、收藏数量等信息，并且为每个术语建立生命周期模型，最后采用无监督学习算法检测突发话题。SONG等[]受到认知权威理论以及社交网络理论的启发，（Cognitive Authority Theory and Social Network Theory）提出基于用户兴趣的方法进行社交媒体话题检测，他们首先根据用户兴趣以及用户的粉丝关系（following relationship）构造话题图，接着采用link-based ranking algorithm计算话题的热度。以突发特征为中心的方法避免了数据稀疏问题，但对突发特征的抽取结果较为依赖，且易受到垃圾信息的影响。

## 社交媒体情感分析研究现状

文本情感分析（Sentiment Analysis）又被称为观点挖掘（Opinion Mining），是指用自然语言处理、文本挖掘以及计算机语言学等方法来识别和提取源素材找那个的主观信息。情感分类问题在情感分析领域被研究得最多最广泛，也是本文主要研究的问题。情感分类问题就是讲给定的文本按照其情感倾向分为正向类别、负向类别或者其他情感类别。从情感分类粒度方面，情感分类任务可以分为文档级、句子级和属性级，我们从这三个方面进行相关工作的介绍

文档级情感分析（Document-level Sentiment Analysis）是情感分析中最浅层的分析，其前提是整个文本只表达了一种情感，即积极的或者消极的情感。在目前的研究中，进行文档级情感分析的主要方法是有监督机器学习和无监督机器学习方法，随着神经网络和深度学习的发展，越来越多的研究者把目光放在了使用神经网络来解决文本情感分类问题。Mikolov等[]扩展了word2vec模型，通过无监督的方式在学习词向量的同时学习文档的表示，为文档学习到一个固定长度的表示向量，然后送到分类器中对文本进行分类。考虑到文档具有层次结构，即词到句子到文档，许多研究者尝试在建模中利用到文档的层次结构信息，Tang等[]提出了一种自底向上的文档表示方法，使用卷积神经网络或者递归神经网络表示句子，然后对句子进行同样的操作，把句子表示作为输入通过设计的门控递归神经网络来学习文档的表示，取得了当时最好的效果；考虑到在句子的表示中不同的词对句子的表示起到不同的作用，比如情感词对句子情感判定更重要，在文档的表示中不同的句子对整个文档的表示起到不同的作用，比如某些句子不含有情感信息，所以Yang等[]在Tang的基础上，进一步提出了层次注意力模型，在词-句子和句子-文档的层级表示中分别加入了注意力机制来更好的表示句子和文档，取得了明显的效果提升。

句子级情感分析（Sentence-level Sentiment Analysis）相比于文档级情感分析层次更加深人，关注某一句话的情感倾向，其前提是单个句子只表达一种情感，即积极的或消极的情感。句子情感分析可以分为两个步骤：主客观分类和将主观句进行情感倾向分类。主客观分类是把句子分为主观句和客观句，主观句中包含情感倾向，客观句是对事实的陈述，不包含情感信息，但是目前所说的句子级情感分析通常是指将主观句进行情感倾向分类。句子级情感分析方法经历了从最开始的基于情感词典的无监督方法，到基于特征的有监督机器学习方法，到现在主流的神经网络方法。Socher等[]利用递归神经网络（Recursive Neural Network）模拟句子解析器的功能，尝试发掘句子中含有的语法信息，对词组和句子进行表示后进行情感分类；Kalchbrenner等[]提出了动态卷积神经网络，使用动态的k-max池化层，更好地发掘句子中N-gram信息，在各种数据集上都取得了很好的效果，可以算作把卷积神经网络方法用在文本分类中的开山之作；针对情感分析任务的特殊性，Qian等[]人在使用神经网络对句子进行表示的基础上，使用情感词典、否定词和程度词对模型进行正则化，有效的降低了情感偏移对情感判断的影响。

属性级文本情感分析（Aspect-level Sentiment Analysis）又被称为方面级文本情感分析，是更细粒度的文本情感分析，考虑到在单独句子中评价者对不同评价对象的表达出不同的情感，比如在商品评论中评价者对商品的不同方面往往会给出不同的评价，这些是句子级文本情感分类没有考虑到的问题。在这里“属性”不单单指实体的属性，或者称为“评价对象”更为合适。属性级文本情感分析主要分为两个步骤：评价对象的抽取和评价对象的情感判别，本文主要关注于评价对象的情感判别，对评价对象的抽取不再详述。同时需要说明的是，评价对象包含隐式评价对象和显示评价对象，隐式评价对象即评价对象不在句子中出现的，显示评价对象即以名词或名词词组的形式在句子中出现的，本文所研究的内容都是关于显示评价对象的，即显示评价对象的文本情感分类。Tang等[]在2015年提出了评价对象连接的LSTM模型（Target-connection LSTM），把评价对象词向量和评价语句中的每个词进行拼接一同作为输入，但是他们忽略了评价对象可能包含多个词，简单地通过词向量的平均作为评价对象的表示不能很好地表现评价对象。同时我们对现有的两个数据集进行了统计发现，在餐馆这个数据集中超过25%的评价对象中包含多个词，在手提电脑这个数据集中超过35%的平均对象中包含多个词。Wang等[]在2016年提出了基于注意力的LSTM模型，在建模的过程中考虑不同词起到不同的作用，情感指示词理应在情感分类中起到更重要的作用，但是最终的效果和基于SVM的方法还是有较大的差距。同时，Tang等[]在2016年设计了一个深度记忆网络，同时考虑到评价对象的位置信息和评价对象的上下文信息，提出了基于位置和基于内容的注意力模型，取得了在标注数据集上的目前最好的效果。

## 对研究现状的分析与思考

从热点发现和文本情感分析两个方面来表达，分析当前主流方法的优点与不足，同时尝试给出可行性解决方案。

对于热点发现的分析与思考。虽然有一些研究者提出了一些社交媒体热点发现算法，但是并没有一个标准的、通用的数据集；同时热点发现主流方法是基于聚类算法，但是对于在线热点发现，热点的数目是不可知，所以对于需要设置聚类数的算法往往是不适用的。针对这两个问题，本文尝试构造了一个较为规范的热点事件数据集，同时使用在线聚类算法而不是离线聚类算法，来更好地适用社交媒体热点发现这个任务。

对于社交媒体文本情感分类的分析与思考。文本情感分类分为不同的粒度，不同的粒度反应了不同的情感信息，单一粒度的情感分析往往不能充分体现评价人群的情感，所以本文从消息级文本情感分类和评价对象明确的文本情感分类两个角度上进行热点相关的文本情感分析，针对不同粒度所存在的不同问题分别设计算法来提高分类的准确率。

## 本文主要研究内容与论文结构安排

### 1.5.1论文主要研究内容

本文研究了社交媒体中的热点发现及热点相关文本情感分类问题，具体为利用人们在社交媒体发表的关于社会中发生的热点问题的描述、看法等信息，通过算法自动发掘其中的热点话题，并分析热点相关的文本的情感倾向性（包含文档级、句子级、属性级多个粒度）。为了解决研究现状中存在的部分问题，本文提出了一下几点主要研究内容：

针对热点发现，提出了一种基于融合规则和机器学习的绝对聚类在线热点发现算法，发挥了人工规则和机器学习各自的优点，更好的表示社交媒体文本，同时提出基于绝对聚类的在线文本聚类方法，能更及时地发现社交媒体中的热点事件。

针对文本情感分析，从消息级和评价对象明确级两个粒度入手，分别改进了现有算法，进一步提高了模型在标注测试集上的效果。在第四章中，从消息级文本情感分析入手，尝试解决文本表达中的情感偏移现象和含蓄表达现象。在第五章中，从评价对象明确的文本情感分析入手，更加深入挖掘文本中对评价对象不同方面的不同评价，本文提出了上下文和属性词分开表示的双边注意力神经网络模型，极大地提高了在标注数据集上的最好准确率。

### 1.5.2论文组织结构

本文共六个章节，按照如下结构安排：

第一章 绪论。介绍了本课题的研究背景和研究应用意义。然后分别介绍了热点发现和文本情感分析在社交媒体上的研究现状和存在的问题，最后介绍了本文的研究内容和尝试解决的问题。

第二章 相关技术介绍。介绍了热点发现和文本情感分类的基本概念和使用的相关技术，包含文本表示技术、文本聚类技术、文本情感分类相关的神经网络技术等。

第三章 一种基于融合特征和绝对聚类在线热点发现方法。先介绍了人工规则和机器学习两种特征发现方法，及如何对两种特征进行融合，然后改进了现有在线聚类算法中的距离计算方法，提出了一种适用于社交媒体热点发现的文本距离度量方法，最后通过实验证明了方法的有效性。

第四章 一种基于文本情感反转和对偶模型的消息级文本情感分类算法。本章首先介绍了一种构造文本情感倾向反转算法，然后提出了利用原始样本和反转样本进行联合训练的对偶算法，最后通过对偶预测算法来更好地判别消息级文本的情感倾向。

第五章 一种基于双边注意力的神经网络模型。针对评价对象明确的文本情感分类任务，从模型的出发点、实际的效果和模型的可解释性等角度对模型进行了介绍。

第六章 总结与展望。对全文的研究内容进行总结，对该研究内容进行展望，同时指出了现有工作的不足之处及未来可能的研究方向。

# 2 相关技术介绍

本章主要介绍论文中使用的相关技术，包含文本表示相关的技术和神经网络相关的技术，为后续章节做铺垫。

## 2.1文本表示技术

文本表示在表示粒度上可以分为对词的表示和对句子的表示，对词的表示主要有两类：独热表示和分布式表示，对句子的表示方式比较多，主要介绍两种：向量空间模型和基于神经网络的分布式表示。

词的独热表示

One-hot Representation，直译过来是独热表示，就是使用一个向量来表示一个词，向量的长度为词典的大小，向量中只有一个位置为1，其他全部为0，1的位置对应该词在词典中位置。这种表示方式比较简单直接，但是存在一些问题：词向量的维度等于词典中包含的词数，但词典往往比较大；词与词之间相互独立，不能计算词之间的相似性（任意两个词向量的乘积为零）。

词的分布式表示

Distributed Representation，翻译为分布式表示，最早由Hinton在1986年提出，尝试来解决独热表示的缺点。基本思路是通过训练将每一个词映射成固定长度的短向量，所有这些向量一起形成一个词向量空间，每个向量为该向量空间中的一个点，在这个空间上引入“距离”，可以根据词之间的距离来判断它们之间的相似性。Mikolov在Hinton的基础上进行了改进和简化，提出了Word2Vec，通过简化神经网络结构和使用层次分类，使利用大规模语料来训练词的分布式表示成为现实，成为现在词向量训练的主流方式。同时近几年有研究者提出可以使用词的共现矩阵来获取词的语义表示，其中以Glove为代表，其作者给出了在大规模社交媒体语料上得到的词向量，被很多研究者作为基准词向量使用。

向量空间模型

向量空间模型（Vector Space Model，VSM）由G.Slaton等人在1975年提出，是一种把文本表示为标识符向量的代数模型。文档D通过向量空间模型可以表示为：

其中，是特征项，是SVM的最小表示单元，可以把文本中字符、词、短语等任意级别的元素以及它们的组合作为VSM的特征项；为对应特征项的权重。使用向量空间模型需要构造特征项和进行特征项权重计算，权重既可以为布尔型也可以为浮点型，其中最有代表性的是词频-逆文档频率（Term Frequency – Inverse Document Frequency，TF-IDF）权重。向量空间模型隶属于词袋模型（Bag of Words，BOW），因为文本被看做是无序的词汇集合，忽略语法语义规则，特征项之间相互独立。向量空间模型的优点在于模型简单、构造文本表示直观方便，通过简化关键词之间的关系来表示文本，虽然使模型具备可计算性，但缺点在于特征项之间的独立性导致无法进行语义相关的判断。针对向量空间模型的缺陷，研究者提出了很多解决方案，但随着神经网络的火热，通过神经网络模型来进行文本的表示逐渐成为现在的主流。

句子的神经网络分布式表示

对句子的分布式表示主要分为两种方式：有监督和无监督的。对于无监督的表示方式，在获得词的分布式表示之后，可以通过对句子中的词向量求均值来获得句子的向量表示，除此之外Mikolov在Word2Vec的基础上进行了修改可以通过词的分布式表示训练一起学习到句子的分布式表示。对于有监督方式，通常伴随着不同的任务设计不同的模型来对句子进行建模，根据有监督信息训练来得到句子的表示，是本文中文本情感分类模型中主要研究的问题。

## 2.2文本聚类技术

文本聚类（Text Clustering）主要是依据著名的聚类假设：同类的文档相似度较大，而不同类的文档相似度较小。作为一种无监督的机器学习方法，聚类由于不需要训练过程，以及不需要预先对文档进行手工标注类别，很适合热点发现任务。

由于聚类算法的多样性很难对聚类算法进行简洁的分类，大体包含以下几种聚类方式：划分法、层次法、密度算法、图论聚类法等，针对本文所解决的问题，现在主要介绍一下划分法。对于给定的含有N个文本的数据集，划分法将这N个样本划分到K（K<N）个分组中，每一个分组代表一个聚类，其中需要满足：每一个分组至少包含一个文本，每一个文本只属于一个分组。对于分组数目K，根据具体划分算法的不同可以事先指定或不指定。主要思路是基于距离，不断把样本划分到距离最近的分组中，直到所有样本都有属于自己的分组。对于样本与分组之间距离的度量有很多中方式，如欧氏距离、夹角余弦。下面详细介绍一下本文使用到的Single-pass在线聚类算法。

Single-pass算法

Single-pass算法又称为单通道法或单遍法，是流式数据聚类的经典方法。对于一次到达的数据流，按照输入顺序每次处理一个数据，依据当前数据与已有类的匹配度大小，将该数据划归为已有的类或者新创建的类簇，实现流式数据的增量和动态聚类。可以看到Single-pass算法是一种增量算法，适合对流式数据进行聚类，而且算法的时间效率高，其缺点是该方法对数据的输入顺序有依赖性，即对于同一个聚类对象如果按照不同的次数输入可能会出现不同的聚类结果，所以如何设计度量算法来衡量聚类对象和类簇的匹配度起到至关重要的作用。

## 2.3有监督情感分类技术

监督学习指通过已有的带标注信息的样本（包含输入和输出信息）去训练得到一个最优模型，再利用这个模型对不带标注的测试样本进行类别标签预测的过程。文本情感分类问题是一个给定文本信息去判断情感极性的分类问题，很自然地可以用有监督学习来进行学习。Pang等最先将有监督机器学习模型用在电影评论情感分类任务中，测试比较了三种经典机器学习分类算法在不同的特征权重下的不同表现，确立了文本情感分类任务基于有监督机器学习的发展方向。同时，随着神经网络和深度学习的发展和其强大的表示能力，其在越来越多的任务中取得了超越以往经典机器学习算法的结果。在近几年，神经网络技术在文本情感分类任务上取得了长足的发展，已经成为了主流方法和研究方向。在下文中介绍一下本文中使用的分类算法：逻辑斯特回归、朴素贝叶斯模型和神经网络相关模型。

### 2.3.1逻辑斯特回归模型

逻辑斯特回归（Logistic Regression）是一种广义线性回归，其在线性回归的基础上套用了一个逻辑函数，使其变成一个非线性模型，虽然其被称为“回归”，但是解决的分类问题，既可以用于二分类也可以用于多分类。



图2.1 Sigmoid函数

逻辑斯特回归的核心在于其使用的逻辑函数，即Sigmoid函数：

如图所示，其值域为(0,1)。对于二分类问题，逻辑斯特回归对不同类别的预测概率分别为：

把上面两个公式进行联合可以得到：

基于对数最大似然准则，可以推导出逻辑斯特回归模型的目标函数为：

最后使用一般的数值优化算法（比如随机梯度下降算法）均可对模型参数进行优化求解。

### 2.3.2朴素贝叶斯模型

朴素贝叶斯（Naïve Bayes）模型是基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入，利用贝叶斯定理求出后验概率最大的输出。朴素贝叶斯模型实现简单，学习与预测的效率都很高，是一种常用的方法。

朴素贝叶斯模型通过训练数据集学习联合概率分布*P(X,Y)*。具体地，学习以下先验概率分布及条件概率分布。先验概率分布为：

其中，*K*为*Y*可能的取值个数。条件概率分布为：

其中，[,…,]表示输入的*n*个特征。

朴素贝叶斯模型对条件概率分布作了条件独立性的假设。由于这是一个较强的假设，朴素贝叶斯模型也因此得名。具体地，条件独立性假设是：

最终朴素贝叶斯模型的分类形式为：

### 2.3.3神经网络技术

神经网络以其强大的表示能力在文本情感分类任务上取得了显著的效果，下面主要介绍一下本文中使用到的神经网络相关技术。

**多层感知器**

多层感知器（Multi-layer Perceptron，MLP）是一种前向结构的人工神经网络，通过全连接的方式映射一组输入向量到一组输出向量。将单个神经元组织在一起，便组成了神经网络，如图2.2所示，左边表示单个神经元，右边表示一个三层神经网络结构。

|  |  |
| --- | --- |
|  |  |

图2.2神经元与多层感知器结构图

相比于单个神经元，多层感知器增加了隐藏层（Hidden Layer），同时在隐藏层中的神经元中增加了激活函数，来进行非线性变换，从而使多层感知器能够表示所有的函数。

**卷积神经网络**

卷积神经网络（Convolutional Neural Network，CNN）是一种前馈神经网络，但其相比多层感知器网络参数更少，在结构上有三个特性：局部连接、权重共享和空间上的次采样，由卷积层-池化层的多次拼接叠加和最后的全连接层构成。

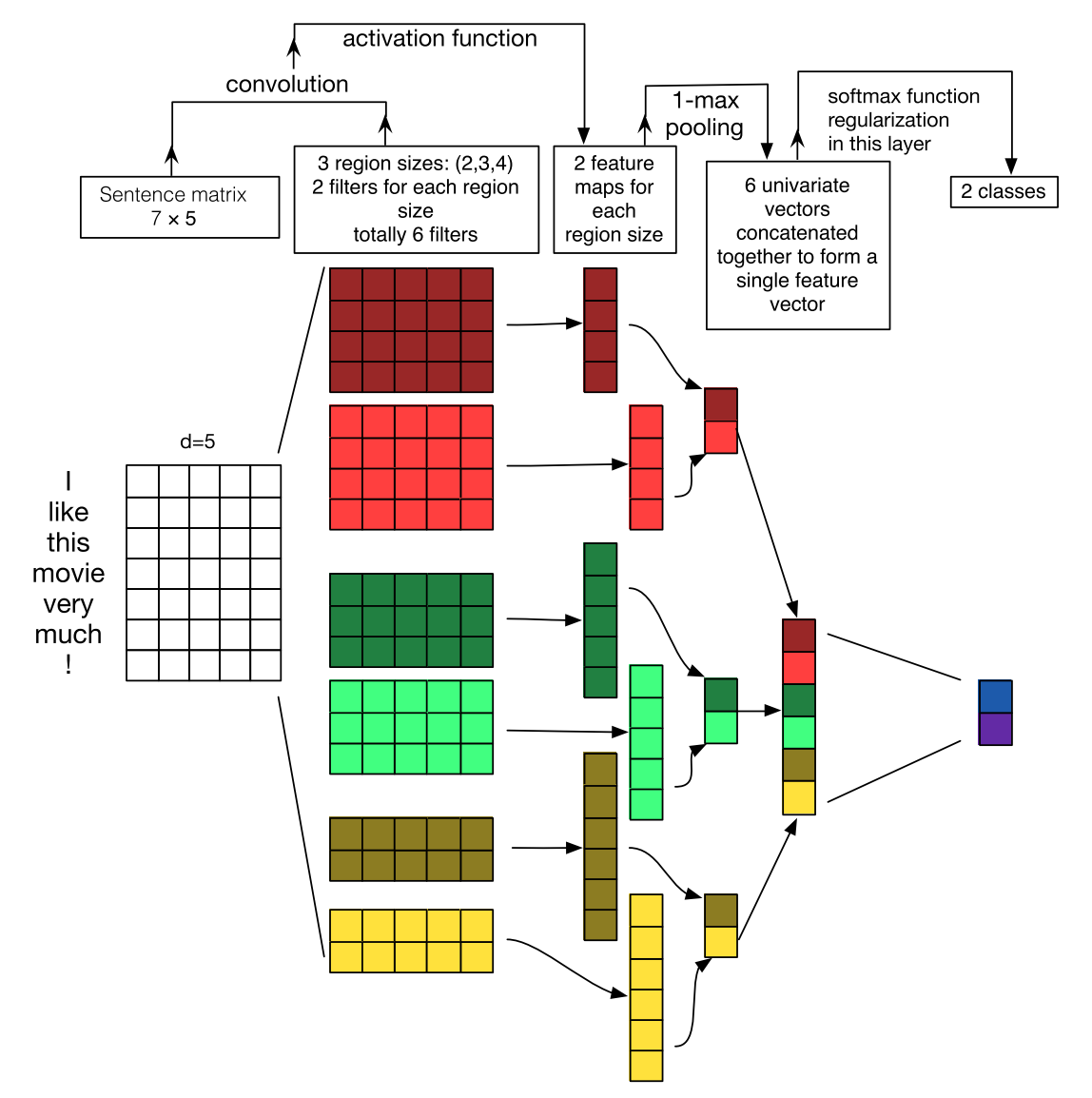


图2.3CNN在文本表示中的应用示例图

如图2.3所示，使用CNN对文本进行向量表示时，先通过不同大小的卷积核对句子向量矩阵进行卷积操作，然后通过池化层，进行max pooling或者mean pooling操作，对于不同长度的文本可以得到相同长度的向量表示，最后可以通过Softmax进行文本情感类别的分类。

**递归神经网络**

递归神经网络是两种人工神经网络的总称，一种是时间递归神经网络（Recurrent Neural Network），另一种是结构递归神经网络（Recursive Neural Network, RNN），在本文中如无特殊声明所有的递归神经网络都是指时间递归神经网络。

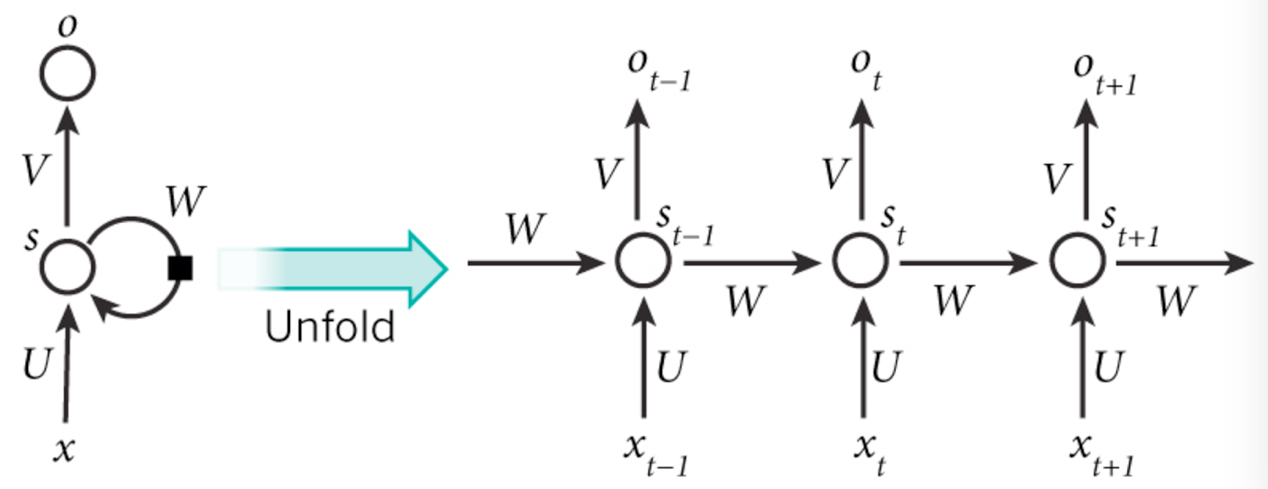


图2.2 递归神经网络展开模型

下面详细介绍一下RNN的内部网络结构，如上图所示。为网络t时刻的输入，为模型t时刻的输出，可以看到t时刻的输出不仅仅与t时刻的输入有关同时和上一个时刻t-1的隐层状态有关：

是权重矩阵，分别对应输入到隐层、隐层到隐层和隐层到输出的连接，*f*是非线性函数，通常为*tanh*。

**长短时记忆网络**

针对RNN在处理长序列输入时容易出现梯度消失/爆炸问题，Hochreiter等在1997年提出了长短时记忆（Long Short-Term Memory，LSTM），在RNN的基础上增加了单元状态（cell state）和三个门控机制：输入门、忘记门和输出门。LSTM中的核心是“cell state”，通常被称为单元状态或者细胞状态，也可以被理解为传送带，作为整个模型中的记忆空间，随着时间而变化，但是传送带本身是无法控制哪些信息是否被记忆，记忆控制是通过三个控制门来实现的。控制门的结构是一样的，通过一个sigmoid函数跟点乘操作组成，sigmoid函数的值为0-1之间，点乘操作决定多少信息可以被传送过去，当为0时，不传送，当为1时，全部传送。

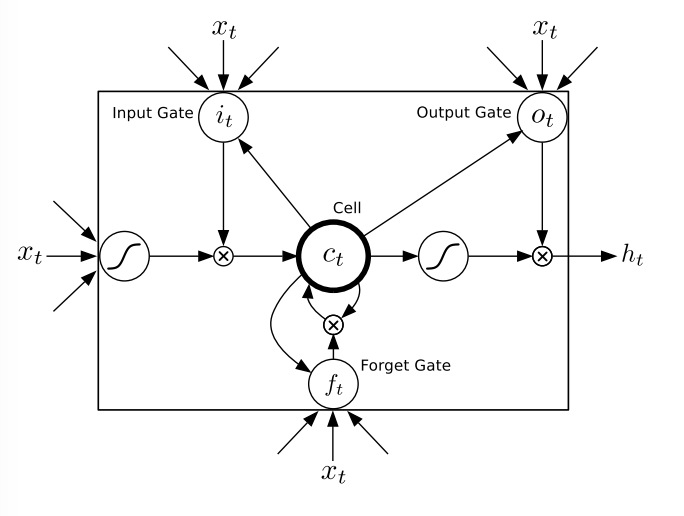


图2.3 LSTM内部结构示意图

输入门：选择记忆现在的某些信息：

忘记门：选择忘记过去的某些信息：

将过去与现在的信息进行合并：

输出门：选择输出信息：

表示两个向量元素对应相乘, 是sigmoid函数,  是参数变量。

**注意力机制**

在处理文本情感分类问题时，RNN和LSTM主要用来对文本输入进行编码表示。如果输入为句子：首先需要对句子中包含的词进行词向量的初始化；然后送入RNN/LSTM中进行编码，从上图我们可以得知RNN/LSTM对句子中每个词从左到右依次进行编码表示，得到每个词对应的隐层状态向量，直到句子中所有的词被处理完；最后，可以使用最后一个隐层状态向量来作为句子的最终向量表示，也可以使用句子中每个词对应的隐层状态向量的均值作为句子的最终向量表示。所以对句子的最终表示可以为：

或者，

其中，T为总时刻数即句子中词的个数。通过RNN对句子进行编码表示后，可以直接对句子进行情感倾向性分类，也可以在句子编码的基础上继续对文档进行编码。

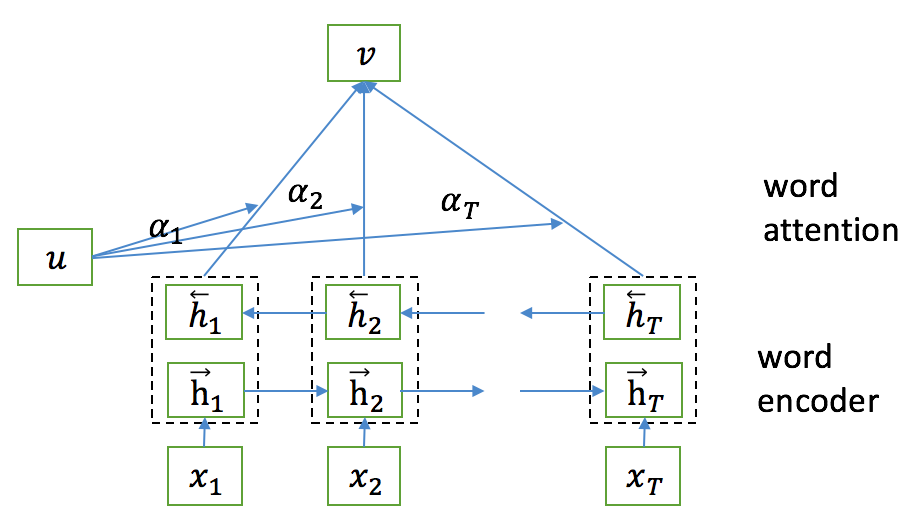


图2.4注意力机制在句子表示中的应用示例

在对句子进行向量表示时，考虑到不同的词对文本表示的重要性不同，特别是针对文本情感分类这个任务，情感指示词明显应该起到更重要的作用。受到神经翻译中注意力机制的启发，注意力机制被应用到文本情感分类任务中的文本表示上面。通过对输入序列中的每个词学习到不同的权重值，通过加权平均的方式得到最终的句子表示。如图所示，具体的公式计算：

其中，u表示上下文向量，用于在注意力机制中计算权重值，在模型训练中学习得到，为隐层状态的权重向量，*r*为文本的最终向量表示。

**总结**

神经网络技术在文本情感分类任务中主要的作用是对文本进行向量表示，针对情感分类这个任务的特殊性，在基础神经网络（比如CNN和LSTM）的基础上，各种新式的网络结构都是为了更好地对文本进行任务相关的表示。在对文本表示完成后，通常的做法是把文本向量送入到Softmax层进行情感标签的分类预测，构成一个完整的网络结构进行有监督的模型训练。

# 3 社交媒体热点发现

由于社交媒体文本数据具有实时性、数据海量、热点数目不确定等因素，本文采用基于特征的在线聚类方式，即先进行热点特征的构造和选择去表示文本然后再进行热点聚类。所以在本章节中，先介绍本文使用的特征构造和文本表示方法，然后介绍本文提出的在线聚类算法。

## 3.1基于机器学习方法和规则融合的文本表示方法

由于社交媒体文本内容简短，传统的基于词的特征向量方法容易导致向量的稀疏性问题和空间高维性问题。所以，本文使用热点词来作为社交媒体文本的特征，构建基于热点词的社交媒体文本特征向量。下面给出热点词的定义：

对于在某个时间段内，某个词的某项热点评价指标高于某个阈值，则称这个词为热点词。

比如，在某个小时内“南京”这个词的词频高于我们设定的10000这个阈值则把“南京”看成这个时间段内的热点词。在下面几个小节中，详细介绍一下本文所使用的热点词发现算法。

### 3.1.1简单规则的热点特征发现方法

本节主要介绍一下本文所使用的几种简单的热点词发现规则：词频、词增长率和词频-文档频率。

1. 词频计算

词频是最能体现一个词的热度的指标，对于当前时间段统计每个词出现的次数。

1. 词增长率计算

词的增长率表示当前时间段中的词k相对于上一个时间段的增长率，计算方法如下：

当时，即当前时间段内出现的某个词k频率很大但是上一个时间段内没有出现时，的值为一个比较大的常数NaN。

1. 词频-文档频率权重计算

词频-逆文档频率（Term Frequency - Inverse Document Frequency）算法是信息索引领域常用的权重算法，其中IDF的计算倾向于在别的文本中出现更少“区分度”较高的词。而在热点检测中，出现热点词的文本数比较多，使用TF-IDF算法不利于热点词的权重计算。所以本文提出词频-文档频率算法：

其中，为词k在时间段中的词频-文档频率权重，为词k在时间段 内的文档频率。

根据热点词的定义，如果在某个时间段内，某个词的以上三个评价指标都高于我们事先设定的阈值，则认为这个词为热点词。

### 3.1.2基于隐马尔科夫模型的热点特征发现方法

基于规则的热点词选择算法在于其简单性和灵活性，但往往不能很好地反应特征之间的内在联系。所以在本节中，我们提出了一种基于隐马尔科夫模型的热点特征发现算法。对于某一个词假设其以生成式的方式产生，在其出现的时刻只可能有两个状态：正常状态q0和热点状态q1，这两个状态由词相邻两次出现的时间间隔决定，同时随着热点的发生与平息不同状态之间可以相互转移，所以我们以离散的隐马尔科夫模型为基础提出了以时间间隔为连续值的隐马尔科夫模型来进行热点词的发现。

基于隐马尔科夫模型的热点词发现算法

对于某个特定的词，我们可以计算出其出现的n个时间间隔，令X对应的状态序列为。假设词的两个相邻间隔x满足指数分布密度函数：

当词处在正常状态时，间隔x有密度函数；当词处在热点状态时，间隔x有密度函数，显然>。令状态q0转移到状态q1的概率为p（）。所以对于状态序列q的密度函数为：

令b为状态序列q中状态的转换次数，所以q的先验概率为：

所以，

其中，。对上式中两边同时取对数后取反，得：

为求得最可能的状态序列，只需使等式右边最小，而等式右边后两项与状态序列q无关，所以：

现在只需使c(q|x)最小。

为了简化模型，令，T为时间段的总时长，从状态q0转移到状态q1代价设置为，>0，从状态q1转移到状态q0代价为零。令是对于输入以状态结尾的最小代价值，则：

并且。

如何使用？

对时间段T内每个词，首先根据词所在的消息得到其出现的时间点和次数，计算得到其出现的时间间隔序列，然后送入到我们提出的模型中，计算得到时间间隔对应的状态序列，如果在某个时间点上其状态为热点状态，则这个词为此时间点上的文本消息的热点词，可以用此词作为其所在文本消息的特征来表示文本消息。

### 3.1.3基于特征融合的文本表示方法

在前面两个小节中，我们首先使用基于规则的方式得到每个时间段内的热点词，然后使用基于隐马尔科夫模型的算法来得到每条消息文本中的热点词。我们可以看到通过隐马尔可夫模型得到的热点词是针对每一条文本数据的，而通过规则方式得到的热点词是针对某个时间段内的，所以我们在融合这两种热点词的时候以隐马尔可夫模型得到的热点词为基础，如果文本消息中含有基于规则得到的热点词则同样把其看做此条文本的特征。

对于某个时间段内中的任意文本消息，可以用热点词特征来进行向量化表示：

其中，N表示热点词的数目，如果表示时间段*i*内的第*j*个文本，如果热点词*k*出现在文本中则令，反正则令。例如，如果热点词集为{南京，宝马，肇事}，文本中包含的热点词为{南京，肇事}，则。

在新闻领域，描述新闻的要素一般包括：何时(When)、何地(Where)、何人(Who)、何事(What)、何故(Why)、如何(How)，用英文首字母简称为“5W1H”，被称为新闻六要素。但是由于社交媒体文本数据往往受字数的限制，在描述热点的时候往往不能包含完整的热点描述。通过对社交媒体数据的分析，我们认为如果一条社交媒体文本中如果包含了一个热点事件或话题，那么它应该至少包含三个要素，仅仅通过少于三个要素很难表述一个事件或话题。基于这个原则，一个描述热点的社交媒体文本数据应该至少包含三个热点词，所以，本文过滤掉语料中所有包含热点词少于三个的文本。

## 3.2基于绝对距离的single-pass聚类算法

前面小节我们介绍了如何进行文本的表示，在获得了文本的表示之后就需要对文本进行聚类，来获取相关的热点事件或话题。由于社交媒体数据的稀疏性和特征的多样性，如何有效地计算文本间、文本和聚类中心点的距离是与传统的文本聚类算法有所不同的。因此，本文在*single-pass*在线聚类算法的基础上，改进了距离度量算法，提出一种“绝对聚类”算法，来更好地适用社交媒体文本聚类。

首先，我们给出文本距离度量算法。对于两个使用热点词表示后的文本和，定义它们之间的距离为:

其中，表示文本中热点词的个数，表示两个文本向量的数量积。

然后，我们给出文本相似度量规则。如果且，当时，则认为文本和文本相似，否则认为不相似；如果或，当时，则认为文本和文本相似，否则认为不相似。

最后，我们给出基于“绝对聚类”的*single-pass*在线聚类算法。“绝对聚类”的核心思想是：如果某个文本属于一个存在的类簇，那么它应该和这个类簇中的每一个文本都相似，即“绝对”属于这个类，否则不属于这个类。具体算法如下：按照时间顺序依次选择文本，第一个文本自动划归为第一个类簇，然后对后面的每个文件计算和之前类簇中每个文本的聚类，满足相似度规则则认为它们相似，如果只和其中一个类簇绝对类似则把其加入到此类簇，如果和多个类簇绝对类似则把其加入到文本数最多的类簇中，如果不和任意类簇绝对类似则把其归为新的类簇，直到所有文本划分完毕。

## 3.3实验与分析

### 3.3.1数据集构建

本实验数据来源于新浪微博，利用爬虫下载工具获取到2015年6月20日到2015年9月20日间和南京相关的约三千万条微博，在这些微博的基础上本文构建了两个数据集用于评估本文提出算法的有效性。

数据集一 为了验证算法对热点事件/话题发现的及时性，本文以“南京肇事宝马案件”事件为基础，在获取的三千万微博中人工标记出与此案件有关的七个关键时间节点，以能否检测出这七个时间节点和检测出的时刻和真实时刻的时间差作为评估标准。

数据集二 为了验证算法对热点事件/话题发现的全面性，本文在获取的三千万条微博中人工筛选出2015年9月13日到2015年9月18日之间的688140条微博，通过本文算法和人工筛选总结出在此时间段中发生的热点事件。

对于数据集一，我们使用检出率和及时值来衡量算法的性能：

对于数据集二，我们使用正确率(Precision)、召回率(Recall)和F值(F-measure)来衡量算法性能。

### 3.3.2实验结果

我们在3.1节中介绍了三种热点特征构造方法，在特征构造的基础上在3.2节中介绍了本文提出的基于绝对距离的Single-pass在线聚类算法。根据热点特征构造方法的不同，产生三种不同的热点发现模型，分别简称为简单规则模型、隐马尔科夫模型和融合特征模型，在本节中对这三种不同的热点发现模型分别在数据集一和数据集二上进行了实验。需要说明的是，由于3.1.1节中热点特征发现算法需要设定时间间隔，所以在实验中设定时间间隔为一小时。

表3.1 模型在数据集一上的结果

|  |  |  |
| --- | --- | --- |
| 模型 | 检测率 (%) | 及时值 (小时) |
| 简单规则模型 | 100.00 | 4.0 |
| 隐马尔科夫模型 | 100.00 | 2.5 |
| 融合特征模型 | 100.00 | 1.0 |

由表3.1可知，不论是基于简单规则的热点特征提取，还是隐马尔科夫模型都能获得很好的检测率，但是在热点检测及时值这个指标上，我们提出的基于融合特征的Single-pass算法取得了明显的优势，证明了模型的有效性。

表3.2 模型在数据集二上的结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 正确率 (%) | | 召回率 (%) | F1值 |
| 简单规则模型 | | 57.14 | 50.00 | 0.5333 |
| 隐马尔科夫模型 | | 61.85 | 75.00 | 0.6779 |
| 融合特征模型 | | 63.64 | 87.50 | 0.7368 |

由表3.2可知，隐马尔科夫模型相比简单规则模型有比较大的提高，融合特征模型相比于其他两种模型都有提高，分析原因有如下两点：

1. 相比于简单规则，隐马尔科夫模型在热点词提取的过程中不需要对输入数据划分时间段，只需要计算词间隔即可，排除了时间段划分粒度不同对结果的影响；

2) 在进行热点发现时，本文采用的是在线聚类算法。

## 3.4本章总结

在本章中，针对热点发现这个任务，本文提出了基于特征融合的文本表示方法，同时在Single-pass聚类算法的基础上改进了距离度量方法，提出了基于“绝对聚类”的Single-pass算法，在实验中取得了不错的效果，证明了算法的有效性。同时，由于本文采用的是在线聚类算法，在线聚类算法固有的对输入顺序敏感的缺点也困扰着我们，是后面需要改进的方向。

# 4 基于对偶模型的消息级文本情感分析

在文本表达中存在一种语言现象：极性转移（Polarity Shift），即句子中存在的否定词、程度副词等会使文本的情感表达发生转移，极性转移主要有三种情况：强调、否定和转折。“强调”是通过程度副词的修饰来增强情感表达，比如中文中的“很”、“十分”和英文中的“so much”、“very”通过修饰动词来增强情感表达。“否定”是通过否定词，比如中文中的“不”和英文中的“don’t”，来反转情感表达的极性。“转折”往往是在同一个句子中使用转折词，比如“但是”，在不同子句中表达不同的情感极性。

## 4.1 情感反义样本构造方法

在本节中，我们详细介绍一下如何构造情感反义样本技术。在反义词典的基础上，对于每一个原始的样本，反转样本构造方法如下：

1. 文本反转。如果存在否定词，则需要先检测否定词的有效范围。所有否定词有效范围之外的情感词替换成其反义词；在否定词的有效范围内，否定词被移除，情感词保留下来。
2. 标签反转。对于每一个训练样本，其标签被替换成其反义标签，比如“积极”替换成“消极”，“消极”替换成“积极”。

在表4.1中给出了两个创建情感反义样本的例子。对于原始的训练样本，“I don’t like this book. It is boring.”，其标签为“消极”，获得其反转样本需要三步：1）把情感词“boring”替换成“interesting”；2）把否定词“don’t”移除，同时因为情感词“like”在否定词的有效范围内不需要替换；3）把情感标签从“消极”转换成“积极”。需要说明的是，在测试集中我们只需要反转文本即可，使用反转样本和原始样本来进行联合预测。图4.1给出了原始样本和反转样本的词袋表示。

通过反转规则创建的情感反转样本在语法和表达上可能不如原始样本，但是原始样本和反转样本都是通过词袋模型来进行文本表示，词序和语法结构都被忽略掉，所以不管在原始样本还是在反转样本中，对语法质量的要求相对较低。同时，我们使用一个参数来综合原始样本和反转样本，对于反转样本对应的模型预测给与较小的权重，来削弱反转样本语法不完美对结果的负面影响。

表4.1创建反转样本示例

|  |  |  |
| --- | --- | --- |
|  | 文本 | 标签 |
| 原始样本 | I don’t like this book. It is boring. | 消极 |
| 反转样本 | I like this book. It is interesting. | 积极 |

## 对偶训练模型

在训练阶段，所有的原始训练数据利用我们提出的反正规则进行情感反转，分别命名为“原始数据集”和“反转数据集”，对于每一条原始样本都会有一条对应的情感反转样本。分类器通过最小化原始训练样本和反转训练样本的联合对数似然函数。为了更好地说明，在本节中我们以逻辑斯特回归（Logistic Regression，LR）为例，详细介绍一下如何进行联合训练。

我们给出本章中用到的符号表示，如表4.2所示。和分别表示原始训练集和反转训练集，和分别表示原始样本和反转样本的向量表示，表示原始类别标签，表示反转类别标签，N是样本总数。令w为特征权重参数，为损失函数。

逻辑斯特回归在二分类问题上被广泛使用，是一个十分经典的基于统计的有监督机器学习分类算法。逻辑斯特回归使用逻辑斯特函数去预测特征向量的所属类别概率：

其中，w为特征权重参数，在模型训练中学习得到。在标注的逻辑斯特回归模型，使用训练样本的对数似然函数（log-likelihood function）作为损失函数：

类似的，在对偶训练模型中，原始样本和反转样本被联合使用，反应在损失函数上为：

考虑到类别标签同样被反转，即有，所以对损失函数进一步化简为：

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |

对比标准逻辑斯特回归损失函数和本文提出的对偶训练模型损失函数，我们可以得到：

1. 如果是一个正向的训练样本，在标准的逻辑斯特回归中其似然值为，在对偶模型中其似然值为，即在对偶模型中进行参数学习时，不仅仅考虑原始样本的正向程度，同时考虑反转样本的负向程度。
2. 如果是一个负向的训练样本，在标注的逻辑斯特回归中其似然值为，在对偶模型中其似然函数值为，即在对偶模型中进行参数学习时，不仅仅考虑原始样本的负向程度，同时考虑反转样本的正向程度。

## 对偶预测模型

在预测阶段，我们对每个原始测试样本x构造其反转测试样本，使用和分别表示原始测试样本和反转测试样本的预测概率。在对偶预测中，最终的预测标签根据样本的正反两面来确定：

1. 当我们衡量一个测试样本的正向程度时，不仅考虑原始测试样本的情感倾向正向程度，同时考虑反转样本的负向倾向程度；
2. 当我们衡量一个测试样本的负向程度时，不仅考虑原始测试样本的情感倾向负向程度，同时考虑反转样本的正向倾向程度。

正负两个情感倾向预测概率通过权重参数联合起来共同进行原始测试样本的情感标签概率，即对偶预测：

其中，（）是权重参数，用来平衡原始样本和反转样本的对最终情感标签预测的重要程度。在我们的实验中，以常量的形式出现，因为我们发现当时往往能获得更好的表现。当标签时，对上式进行联合表示，即：

为了降低糟糕反转样本对预测结果的影响，在进行最终的情感标签预测时，不单单使用对偶模型来进行预测，同时引入原始模型进行预测修正，即：

其中，-，t为阈值。在实验中，我们把t设置为零，即如果原始模型中对于情感标签预测的置信度更高则选择原始模型最为最终结果，否则使用对偶模型的结果作为最终结果。

为什么我们的模型能够解决情感偏移问题？下面以表4.1中的示例为例，进行简单的解释。我们假定“I don’t like this book. It is boring”为原始样本，“I like this book. It is interesting”为反转样本。在传统的词袋模型中，“like”在预测情感倾向时会被赋予更高的正向权重，尽管其前面有否定词“don’t”，所以有很大的概率原始样本会被预测为正向情感。在我们的对偶模型中，通过构造反转样本去掉了否定词，由于“like”的存在所以反转样本有大概率被预测为正向样本。所以通过一正一反这种操作，能够在一定程度上降低情感偏移现象对情感倾向预测的影响。

## 实验与分析

### 4.4.1数据集与实验设置

为了更方便地进行模型对比，我们使用相关比赛和论文中提到的标准数据集，包含中文数据和英文数据，具体如下：

1. NLPCC2013：自然语言处理与中文计算会议在2013年举办的评测任务2中的中文微博情绪分类的语料，本文将所有正向情绪合并为积极类别，将所有的负向情绪合并为消极类别；
2. MDSD：多领域情感数据集（Multi-Domain Sentiment Dataset）是一个英文数据集，包含Book、DVD、Electronics和Kitchen四个领域内的评论数据，原始情感标签分为1-5星，本文将所有1和2星的数据合并为消极类别，将所有4和5星的数据合并为积极类别。

详细的数据集统计信息如下表4.1所示。

表4.1数据集统计信息

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集 | #积极 | #消极 | 平均长度 |
| Book | 1000 | 1000 | 201 |
| DVD | 1000 | 1000 | 197 |
| Electronics | 1000 | 1000 | 126 |
| Kitchen | 1000 | 1000 | 105 |
| NLPCC2013 | 2500 | 2500 | 35 |

在实验中，对所有数据进行5-fold交叉验证，所有的实验结果为5次实验结果的均值，使用逻辑斯特回归模型和朴素贝叶斯模型作为基础分类模型来验证我们提出的对偶模型的有效性。对于模型的特征选择，我们使用两种特征：unigram和bigram，分别测试了单独使用unigram特征和同时使用两种特征。需要说明的是，我们不是为了比较不同的基础分类模型的差异，而是为了验证对偶模型的有效性。

### 4.4.2结果分析

为了验证对偶模型的有效性，在本章中和以下几个模型进行了对比：

1. Baseline，使用基础分类模型（LR和NB）进行情感分类；
2. DS，论文[]中提出的模型，否定词被拼接到其有效范围内的情感词的后面，比如“The book is not interesting”被转换成“The book is interesting-NOT”；
3. LSS，论文[]中提出的模型，把文本分成有极性转移部分和无极性转移部分，分别进行情感分类，使用两部分的情感组合来综合判断原始文本的情感倾向；
4. DSA，Dual Sentiment Analysis，即本文提出的模型。

表4.2 LR模型作为基础分类器时的分类正确率

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 特征：unigram | | | |  | 特征：unigram和bigram | | | |
| Baseline | DS | LSS | DSA |  | Baseline | DS | LSS | DSA |
| Book | 0.771 | 0.775 | 0.784 | **0.809** |  | 0.779 | 0.789 | 0.809 | **0.823** |
| DVD | 0.785 | 0.800 | 0.815 | **0.826** |  | 0.801 | 0.802 | 0.823 | **0.831** |
| Electronics | 0.803 | 0.815 | 0.823 | **0.842** |  | 0.826 | 0.833 | 0.844 | **0.857** |
| Kitchen | 0.835 | 0.841 | 0.851 | **0.875** |  | 0.851 | 0.858 | 0.872 | **0.886** |
| Avg. | 0.798 | 0.808 | 0.818 | **0.838** |  | 0.814 | 0.821 | 0.837 | **0.849** |
| NLPCC2013 |  |  |  |  |  |  |  |  |  |

表4.3 NB模型作为基础分类器时的分类正确率

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 特征：unigram | | | |  | 特征：unigram和bigram | | | |
| Baseline | DS | LSS | DSA |  | Baseline | DS | LSS | DSA |
| Book | 0.779 | 0.783 | 0.792 | **0.818** |  | 0.811 | 0.815 | 0.822 | **0.837** |
| DVD | 0.795 | 0.793 | 0.810 | **0.824** |  | 0.824 | 0.826 | 0.837 | **0.844** |
| Electronics | 0.815 | 0.828 | 0.824 | **0.844** |  | 0.841 | 0.857 | 0.852 | **0.859** |
| Kitchen | 0.830 | 0.847 | 0.840 | **0.864** |  | 0.878 | 0.879 | 0.883 | **0.895** |
| Avg. | 0.804 | 0.813 | 0.817 | **0.838** |  | 0.838 | 0.844 | 0.848 | **0.859** |
| NLPCC2013 |  |  |  |  |  |  |  |  |  |

表4.2和表4.3分别为逻辑斯特回归模型和朴素贝叶斯模型作为基础分类器时不同模型的分类正确率。

## 本章总结

# 5 基于双边注意力模型的评价对象明确的文本情感分析

## 5.1 细粒度情感分析相关介绍

### 5.1.1 相关概念定义

为了统一表达方式，本文对实体、方面、评价对象这几个概念进行了定义与说明。

定义 1 实体：实体可以是产品、服务、主题、人物、组织、或者事件等。实体由其组成部分和属性集组成，同时组成部分也有可能包含属性和更细粒度的组成部分。比如，对于“电脑”这个实体，其组成部分有“屏幕”、“电池”、“键盘”等，其属性有“价格”、“重量”、“体积”等，同时对于“电池”这个组成部分同样具有属性，如“电池寿命”、“电池重量”等。

定义 2 方面（Aspect）：实体的组成部分和属性统称为方面，同时也被称为特征（features）。方面在表达中通常为名词和名词词组，但是也可以为动词、动词词组、形容词和副词。方面表达分成显式方面表达（explicit aspect expression）和隐式方面表达（implicit aspect expression）。显式方面表达为名词和名词词组，隐式方面表达为名词和名词词组以外的词或词组。比如，“电脑有点贵”，“贵”就是隐式方面表达，其表示的是“价格”这个方面；“电脑的电池很不错”，“电池”就属于显式方面表达。显示方面是明确出现在表达语句中的，隐式方面则通过表达语句中的其他词汇体现。

定义 3 评价对象（Opinion Targets）：指某段评论语句中情感表达所指的对象，可以为实体、实体的组成部分或者实体的属性，可以认为和方面等价。

### 5.1.2 任务介绍

细粒度情感分析任务包含两个子任务：评价对象抽取任务和评价对象情感分类任务。评价对象抽取是从评论语句中提取出评价对象，包含显式评价对象抽取和隐式评价对象抽取。评价对象情感分类是对提取出的评价对象给出情感倾向的判定。一般情况下把这两个任务看做独立的两个子任务来处理，在进行评价对象情感分类的时候评价对象作为先验知识输入，也有把评价对象抽取任务和评价对象情感分类联合起来的做法，利用两个任务的相关性同时获取评价对象及其情感倾向。

本章节关注的任务是评价对象情感分类任务，评价对象同评价语句一起作为输入，评价对象的情感倾向作为输出。对于隐式评价对象不论是提取任务还是情感倾向分类任务现在都没有相对成熟的算法和较为通用的公共评测数据集，所以本章节主要关注显示评价对象的情感倾向分类问题，即评价对象以名词或名词词组的形式明确出现在评论文本中的情况。

### 5.1.3 相关工作

在早起的研究中，对评价对象明确的情感分类任务的处理方式和其他文本情感分类任务类似，研究者设计各种特征（比如情感词典特征、语义特征等）通过各种统计机器学习模型来进行训练分类器（Kiritchenko et al. 2014; Wagner et al. 2014; Vo and Zhang 2015）。由于人工设计特征需要花费很多体力劳动并且随着特征设计的精细化几乎到达了性能瓶颈，在近几年中越来越来的研究者采用更高级的神经网络算法来进行情感分类任务。利用神经网络强大的表示能力，精心设计的神经网络模型可以自动地生成评价对象和其上下文的实值低维向量，来表达其语义和情感信息，并且获得了目前最好的正确率。

Jiang等在2011年的论文中指出，在推特情感分类中约40%的错误是由于没有考虑评价对象。随着研究工作的深入进行，越来越多的研究者认识到评价对象信息在文本情感分类任务中的重要性，充分利用评价对象信息已经成为共识。但是之前的方法中，对评价对象的建模还存在许多缺点。Tang等人在2015年提出了评价对象连接的LSTM模型（Target-connection LSTM），把评价对象词向量和评价语句中的每个词进行拼接一同作为输入，但是他们忽略了评价对象可能包含多个词，简单地通过词向量的平均作为评价对象的表示不能很好地表现评价对象。同时我们对现有的两个数据集进行了统计发现，在餐馆这个数据集中超过25%的评价对象中包含多个词，在手提电脑这个数据集中超过35%的平均对象中包含多个词。Wang等人在2016年提出了基于注意力的LSTM模型，在建模的过程中考虑不同词起到不同的作用，情感指示词理应在情感分类中起到更重要的作用，但是最终的效果和基于SVM的方法还是有较大的差距。同时，Tang等人在2016年设计了一个深度记忆网络，同时考虑到评价对象的位置信息和评价对象的上下文信息，提出了基于位置和基于内容的注意力模型，取得了在标注数据集上的目前最好的效果。

通过对之前方法的分析和对问题的深入思考，对于评价对象明确的情感分类任务，我们认为目前存在两个主要的问题需要解决：一是如何更好地表示评价对象，尤其是当评价对象中存在多个词时；二是如何更好地表示评价对象的上下文，尤其是如何获取对情感判断其重要作用的情感指示词。针对这两个问题，我们提出了对评价对象独立表示的双边注意力神经网络模型，在下面的小节中我们会给出详细的介绍说明。

## 5.2 评价对象独立表示的双边注意力神经网络模型

在本节中详细介绍我们提出的评价对象独立表示的双边注意力神经网络模型，先详细介绍一下模型中使用的基础单元—双边长短时记忆网络(Bidirectional Long Short-term Memory Network, Bi-LSTM)，然后对本文提出的评价对象独立表示的双边注意力神经网络模型(Target-Independent Representation Neural Network with Bilateral Attention Model, TIBA)细节进行详细的说明。

### 5.2.1 Bi-LSTM

标准的RNN容易出现梯度消失/爆炸现象，很多研究者提出了各种各样的解决方案来尝试解决这个问题，Hochreiter和Schmidhuber提出的长短时记忆网络(Long Short-term Memory, LSTM)在不断改进中逐渐成为主流方案。如图5.1所示，相比于RNN，LSTM单元包含三个额外的门：输入门、忘记门和输出门，来分别控制当前信息的输入、历史信息的忘记和下一步信息的输出。LSTM单元的计算过程如下：

(1)

(2)

(3)

(4)

(5)

(6)

其中，表示矩阵中元素相乘（element-wise multiplication），表示sigmoid函数，分别为输入门、忘记门、输出门的参数。

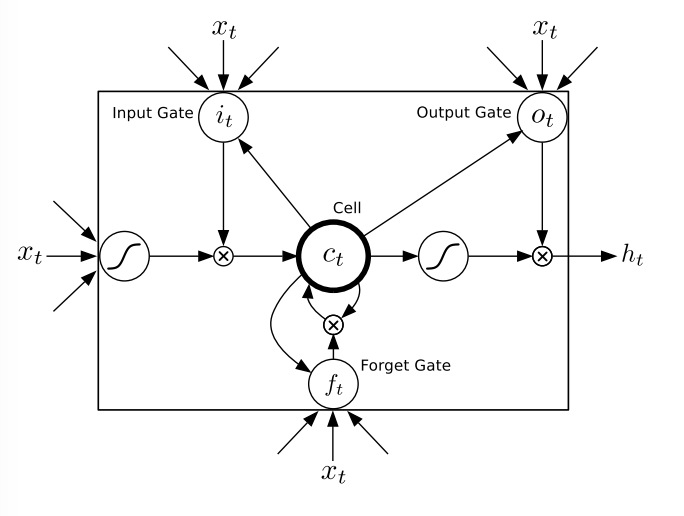


图5.1 LSTM内部结构图

在LSTM中，对于每个词的表示是由其位置t之前的词编码得到的，其位置t之后的词没有对其产生影响，这种编码方式不能很好的利用其上下文信息。所以Graves等人针对这个问题，提出了Bidirectional LSTM，使用两个LSTM分别从前往后和从后往前两个方向对词序列进行编码表示：

(7)

(8)

对于位置t的词，使用前向LSTM得到的隐层状态向量和后向LSTM得到的隐层状态向量的拼接作为词的向量表示，通过这种方式来同时考虑词的左右上下文，如图5.2所示。

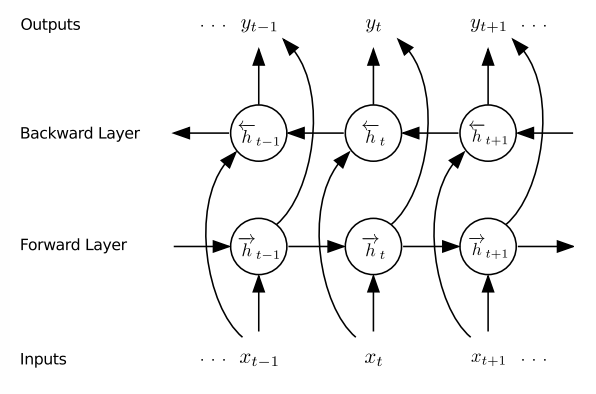


图5.2 Bi-LSTM模型

### 5.2.2 TIBA模型

由图可知，本文提出的TIBA模型主要包含三部分：输入层、表示层、Softmax层，下面我们分别从这三部分详细介绍模型的具体细节。



图5.3 TIBA模型结构

输入层

对于每个输入的句子，首先根据评价对象的位置把其分成三部分：左边上下文、评价对象和右边上下文，表示词*i*的初始词向量，其维度为*d*，可以由Word2Vec训练得到，*N*为句子的长度即句子中词的个数，*L、M、R*分别为左边上下文、评价对象和右边上下文的句子长度。下面通过一个例子来具体说明：

对于句子“I think the computer screen is very nice but the operating system is slow”，其评价对象有两个：“computer screen”和“operating system”，情感倾向分别为积极和消极。根据评价对象的不同输入也是不同的，当评价对象为“computer screen”时：

左边上下文：I think the

评价对象：computer system

右边上下文：is very nice but the operating system is slow

情感标签为：积极

当评价对象为“operating system”时：

左边上下文：I think the computer is very nice but the

评价对象：operating system

右边上下文：is slow

情感标签为：消极

输入时首先把词映射为相应的分布式表示向量，然后把左边上下文、评价对象和右边上下文分别送入到Bi-LSTM中进行编码表示。

表示层

对于三部分输入分别送到三组Bi-LSTs中进行编码，分别得到它们的隐层表示：左边上下文、评价对象和右边上下文。为了更好地表示评价对象相关的句子情感表示，我们引入了双边注意力机制来更好地获取评价对象相关的情感指示词：

1.对于评价对象，使用其内部各个词的隐层向量的均值来作为其最终的向量表示：

2.对于左右上下文，我们首先定义一个得分函数f，使用评价词的向量表示和上下文隐层的向量作为输入，来衡量评价对象词和上下文词的相关性：

其中，和分别为权重矩阵和偏置项，tanh是非线性函数。使用f函数来 分别计算左右上下文隐层向量的权重：

通过权重向量和隐层向量矩阵计算得到上下文句子的最终向量表示：

3.通过对左右上下文和评价对象表示向量的拼接得到最后的评价对象相关的句子向量表示：

Softmax层

把上一步得到的句子向量表示，送入到Softmax函数中，得到最后的情感预测概率：

其中，和分别为权重矩阵和偏置项。

模型训练

模型通过有监督的方式最小化情感分类的交叉熵损失，损失函数定义为：

(17)

其中，C是情感类别数（两类对应积极和消极，三类对应积极、消极和中性），表示第i个类别的one-hot表示方式，表示第i个类别的预测概率，是L2正则化的权值系数，表示模型中所有的参数。

## 5.3 实验与分析

### 5.3.1 实验设置

数据集

由于中文社交媒体中缺乏标准的数据集，为了更好地比较说明TIBA模型的优势与效果，所以采用SemEval 2014 Task 4中的两个英文数据集，包含两个领域的评论数据：手提电脑Laptop和餐馆Restaurant。具体数据统计如下表：

表5.1 数据集统计信息

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集 | 积极类别数目 | 中性类别数目 | 消极类别数目 |
| Restaurant-Train | 2164 | 637 | 807 |
| Restaurant-Test | 728 | 196 | 196 |
| Laptop-Train | 994 | 464 | 870 |
| Laptop-Test | 341 | 169 | 128 |

同时，我们统计了在不同的数据集中评价对象中包含不同词数的比例情况，如下所示：

表 5.2 评价对象包含词数统计信息

|  |  |  |
| --- | --- | --- |
| 数据集 | 单个词评价对象 | 多词评价对象 |
| Restaurant-Train | 2720/75.4% | 888/24.6% |
| Restaurant-Test | 801/71.5% | 319/28.5% |
| Laptop-Train | 1474/63.3% | 854/36.7% |
| Laptop-Test | 351/55.0% | 287/45.0% |

实验参数

在本文实验中，词向量的维度和隐层状态向量的维度都设置为300，并且为了其他系统公平比较，采用300维的GloVe向量来初始化词向量。词表之外的词和权重矩阵都初始化为U(-0.1，0.1), 所有的偏置项设置为0。在模型训练中，学习率设置为0.1，L2-norm设置为1e-5，dropout设置为0.5，使用带动量（Momentum）的随机梯度下降（Stochastic Gradient Descent，简称SGD）来优化模型，动量系数设置为0.9，使用Tensorflow来实现模型。

### 5.3.2 实验结果

为了使比较更有说服力，我们的模型和以下十种方法进行了对比：

1. Majority 作为基线，简单地把训练集中类别数最多的类别赋予测试集样本。
2. Simple SVM 使用uin-gram和bi-gram作为特征的支持向量机（Support Vector Mechine，简称SVM）。
3. Feature-enhanced SVM 使用SVM作为分类器，但构造了复杂的特征模板，包括：N-gram特征、解析器特征和词典特征等。
4. LSTM 使用标准的LSTM来表示句子后进行情感分类。
5. TD-LSTM 采用两个LSTM分别对左边上下文和右边上下文进行编码表示，然后拼接作为句子的表示。
6. AE-LSTM 是LSTM的改进版本，把评价对象的初始向量和句子中的每个词的初始向量进行拼接后送入到LSTM中，编码得到句子的最终表示。
7. ATAE-LSTM 在AE-LSTM基础上，通过注意力机制获得每个词的权重值，加权后得到句子的表示。
8. GRNN-G3 采用RNN而不是LSTM来进行句子编码表示，通过均值Pooling得到左右上下文和评价对象的表示，然后通过门控RNN得到最终的句子表示。
9. MemNet 考虑到评价对象的上下文信息和位置信息，设计了内容注意力和位置注意力来获得句子的表示，然后通过迭代模型层数来获得更好的结果。
10. IAN 考虑到评价对象中每个词对评价对象的最终表示影响不同，在评价对象的表示中加入注意力机制来更好地表示评价对象，同时考虑到在表示的过程中评价对象和上下文的相互影响。

表5.3是实验结果，评估指标使用正确率(Accuracy)。

表5.3 不同模型实验结果

|  |  |  |
| --- | --- | --- |
| 数据集 | Restaurant (%) | Laptop (%) |
| Majority | 53.50 | 65.00 |
| Simple SVM | 73.22 | 66.97 |
| Feature-enhanced SVM | 80.90 | 72.10 |
| LSTM | 74.30 | 66.50 |
| TD-LSTM | 75.60 | 68.10 |
| AE-LSTM | 76.60 | 68.90 |
| ATAE-LSTM | 77.20 | 68.70 |
| GRNN-G3 | 79.55 | 71.47 |
| MemNet | 80.95 | 72.37 |
| IAN | 78.60 | 72.10 |
| TIBA | 81.34 | 74.45 |

从表5.3中我们可以看到，Feature-enhanced SVM获得了十分有竞争力的结果，超过了大多数的神经网络方法，说明传统的机器学习方法依然有其存在的意义；从LSTM到IAN，基于递归神经网络的方法的效果在逐步提高，注意力机制在其中发挥了巨大的作用；MemNet虽然使用前向神经网络，但因为其采用了位置注意力和内容注意力并多层迭代模型，取得了比以往递归神经网络更好的效果；本文提出的基于双边注意力机制的神经网络模型取得了目前最好的效果，证明了模型的有效性。

### 5.3.3实验结果的进一步分析

本模型提出的初衷是解决两个问题：一是解决评价对象中存在多个词的问题，二是为了更好地表示评价对象相关的上下文，获取对评价对象情感起决定作用的情感指示词。为了更好地验证本模型是否能更好的处理这两个问题，在本节中我们针对这两个问题分别对模型进行了分析。

评价对象独立学习的有效性

ATAE-LSTM是一个典型的基于LSTM的用于解决评价对象明确的文本情感分类模型，和本文的模型一样，都采用了LSTM和注意力机制，但是本文模型对评价对象进行了表示学习，以期待能更好地表示含有多个词的评价对象。为了评估TIBA模型在不同长度评价对象上的表现，我们对比了TIBA模型和ATAE-LSTM模型在单词评价对象和多词评价对象上的效果，如表5.4所示。

表5.4 ATAE-LSTM和TIBA在不同长度评价对象上的对比

|  |  |  |
| --- | --- | --- |
|  | Single-word Target (%) | Multi-word Target (%) |
| ATAE-LSTM | 73.50 | 61.32 |
| TIBA | 77.49 | 69.69 |
| Improvement of TIBA | +3.99 | +8.37 |

在表5.2中，我们可以看到，在餐馆数据中超过1/4的样本含有多词评价对象，在手提电脑数据中超过1/3的样本含有多词评价对象。从实验结果中，TIBA在餐馆和手提电脑数据上比ATAE-LSTM获得了4.14%和5.75%的提高，进一步分析可以发现，在手提电脑数据上，当评价对象为单个词时模型提高为3.99%，但是当评价对象为多词时提高为8.37%，这充分证明了本文提出的对评价对象重新表示学习能更好地处理多词评价对象样本。

注意力机制的有效性

为了更直观地理解注意力机制和其产生的效果，本文对权重值进行了可视化，如上图所示，红色表示评价对象，蓝色表示上下文词语，颜色越深表示权值越大。从（a）中可以看到，对于“computer screen”这个评价对象，“nice”获得了更高的权值，表示其对评价对象的影响更大，是符合我们的预期的；从（b）中可以看到，对于“operating system”这个评价对象，“slow”获得了比“nice”更高的权值，证明了本文提出的模型能够比较好地处理同一句话中包含多个评价对象的情况。



a.评价对象: computer screen



b.评价对象: operating system

图5.4 注意力机制的可视化

哪个模块最重要？

为了进一步验证模型的有效性，分析模型的哪一部分对最终的结果影响最大，在双边注意力模型的基础上，通过去掉某些结构设计了一下三个模型：

1.No-Bi 是TIBA的简化版，把Bi-LSTM替换成了LSTM。

2.No-Bi-Attention 在No-Bi的基础上，进一步去掉了注意力机制，使用LSTM最后的隐层状态向量来作为句子的向量表示。

3. No-Target-Learned 是TIBA的简化版，不对评价对象进行表示学习，仅仅使用初始化向量矩阵的均值作为评价对象的表示。

表5.5 TIBA模型及其弱化版本效果

|  |  |  |
| --- | --- | --- |
| 模型 | Restaurant (%) | Laptop (%) |
| No-Bi | 80.30 | 73.09 |
| No-Bi-Attention | 78.18 | 71.00 |
| No-Target\_Learned | 78.01 | 70.06 |
| TIBA | 81.34 | 74.45 |

从表中可以看到，No-Bi比TIBA表现稍差，说明Bi-LSTM能更好地进行语义表示，但是提升是很有限的；进一步去掉注意力机制后，No-Bi-Attention获得了更差的效果，说明在本模型中注意力机制是很重要的一部分；通过比较TIBA和No-Target-Learned模型，我们可以看到效果下降最严重，在餐馆数据和手提电脑数据上效果分别下降了3.33%和4.39%，说明对评价对象的重新表示学习对评价对象明确的情感分类任务十分重要，同时从表5.2中我们可以看到手提电脑中含有更多的多词评价对象，手提电脑获得了更多的效果下降，进一步说明了本文提出的TIBA模型能更好地解决多词评价对象的情况。

## 5.3本章总结

对于评价对象确定的文本情感分类这个任务，本节中针对多词评价对象现象和不同词对情感判断重要性不同现象，提出了对评价对象独立表示学习的双边注意力神经网络模型，通过模型实验与详细比较分析证明了模型的有效性，取得了在相关标准数据集上的最佳效果。

# 6 总结与展望