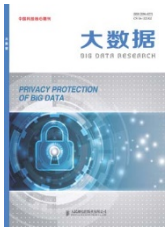

大数据
Big Data Research
ISSN 2096-0271,CN 10-1321/G2

《大数据》网络首发论文

题目：基于社交网络大数据的民众情感监测研究

作者：李爱黎，张子帅，林荫，王秋菊，杨建安，孟炜程，张岩峰

网络首发日期：2022-05-06

引用格式：李爱黎，张子帅，林荫，王秋菊，杨建安，孟炜程，张岩峰．基于社交网络大数据的民众情感监测研究[J/OL]．大数据．
https://kns.cnki.net/kcms/detail/10.1321.G2.20220506.1207.002.html

网络首发时间：2022-05-06 15:05:00
网络首发地址：https://kns.cnki.net/kcms/detail/10.1321.G2.20220506.1207.002.html

基于社交网络大数据的民众情感监测研究

李爱黎[1]，张子帅[1]，林荫[2]，王秋菊[2]，杨建安[1]，孟炜程[1]，张岩峰[1]

1 东北大学计算机科学与工程学院 沈阳市 110000；

2 东北大学外国语学院 沈阳市 110000

摘要： 近年来，新浪微博、Twitter 等社交网络平台逐渐成为反映社会舆情的主要载体之一，为网民发表观点和表达情绪提供了一个便利的平台。一旦突发事件发生，民众借助社交网络平台发布微博、推文等来表达与此相关的态度，这些信息通过社交网络进一步被传播扩散，从而产生一定的社会影响。基于社交网络大数据的舆情监控已经成为新的研究热点，利用各国的社交网络大数据进行民众情感监测，有助于直接掌握国际关系中的民众情感倾向，对我国外交、政治、对外贸易等方面都有很重要的作用。基于此，提出了一种面向中日语料的民众情感监测系统，该系统能够同时分析微博和 Twitter 等社交平台的中日文数据中包含的情感倾向，并以可视化的形式展现给用户。在情感分析算法上，在 BERT 模型基础上结合自扩展的中日文情感词典，提出了一个新的情感分类模型—EmoBERT。实验结果表明，相比于原始 BERT 模型，提出的 EmoBERT 模型在中文情感分类任务和日文情感分类任务上都取得了很好的效果。其中 EmoBERT-C 将中文 BERT 准确率从 89.68%提升至 92.15%，EmoBERT-J 将日文 BERT 模型准确率从 74.73%提升至 78.26%。

关键词： 情感分析；舆情监测；情感词典；中日关系；微博；Twitter

中图法分类号： TP311.13 文献标志码：A

doi: 10.11959/j.issn.2096-0271.2022054

Research on Emotion Monitoring of Chinese-Japanese Public Based on Social Network Big Data

LI Aili[1], ZHANG Zishuai[1], LIN Yin[2], WANG Qiuju[2], YANG Jian'an[1], MENG Weicheng[1], ZHANG Yanfeng[1]

1 Department of Computer Science and Engineering, Northeastern University, Shenyang 110000, China

2 Department of Foreign Language, Northeastern University, Shenyang 110000, China

Abstract: In recent years, social networking platforms such as Sina Weibo and Twitter have gradually become one of the main carriers for reflecting social public opinion, providing a convenient platform for netizens to express their opinions and express their emotions. Once an emergency occurs, people use social network platforms to publish microblogs, tweets, etc. to express their attitudes related to this. These information are further spread and spread through social networks, thus producing a certain social impact. Public opinion monitoring based on social network big data has become a new research hotspot. The use of social network big data in various countries to monitor people's emotions is helpful to directly grasp people's emotional tendencies in international relations, and has a great impact on my country's diplomacy, politics, foreign trade and other aspects. have important roles. Based on this, this paper proposes a public sentiment monitoring system for Chinese and Japanese data, which can simultaneously analyze the emotional tendencies contained in Chinese and Japanese data on social platforms such as Weibo and Twitter, and display it to users in a visual form. In the sentiment analysis algorithm, based on the BERT model and combined with the self-expanding Chinese and Japanese sentiment dictionary, we propose a new sentiment classification model—EmoBERT. The experimental results show that, compared with the original BERT model, the EmoBERT proposed in this paper has achieved good results on both Chinese sentiment classification tasks and Japanese sentiment classification tasks. Among them, EmoBERT-C increased the accuracy of Chinese BERT from 89.68% to 92.15%, and EmoBERT-J increased the accuracy of Japanese BERT model from 74.73% to 78.26%.

Keywords: Sentiment analysis, Public opinion monitor, Sentiment lexicon, Chinese-Japanese relations, Weibo, Twitter

1 引言

互联网的飞速发展改变了人们传统的交流习惯，人们对网络的利用率越来越高。互联网上相继出现了社区、论坛、微博等各种形式的社交网络平台，用户在网上通过这些平台去表达自己对某一事件的看法和态度，这些信息包含了大量的社会热点及情感倾向[1]。因此，在大数据

技术支撑下，如何挖掘社交网络中用户的观点、态度和情感，并服务于社会是一个很有意义的工作。

目前，多数研究主要采用主流社交网络平台的热门数据进行情感分析与监测。Zhao 等人[2]构建了一个面向中文微博的情感分析系统，对异常或突发事件进行监测；Wang 等人[3]利用 Twitter 数据构建了一个针对 2012 美国大选结果的进行实时预测的系统，通过统计美国民众对于四位候选人的情感倾向来预测大选结果；Jennifer 等人[4]提出了一种预测算法，针对 Twitter 上发生的某一事件，预测其发生时间。

然而目前的大多数研究仅仅是对于微博、Twitter 等某单一平台进行舆情数据的情感分析，并且多数是针对中文语料和英文语料的分析，国内使用日语语料进行情感分析的研究极少。与此同时，针对海量的数据，利用人工浏览、打标签的方式来获取用户情感是一件及其复杂且困难的事情。

因此，本文提出一种面向中日语料的民众情感监测系统，该系统能够同时分析微博和 Twitter 等社交平台的中日文舆情数据中包含的情感倾向。当某一焦点事件发生时，自动生成中日两国民众的情感对比，进而供相关舆情部门监测。从而情感分析算法上，本文在 BERT(Bidirectional Encoder Representations from Transformers)模型基础上提出了一个新的情感分类模型--EmoBERT，利用自扩展的情感词典改进了 BERT 的预训练任务，并提出了情感词增强的注意力机制，解决了 BERT 在预训练阶段情感特征提取不充分的缺陷。实验证明，相比于原始 BERT 模型，在中文和日文情感分类任务上都取得了更好的效果。

此外，为了更好地对算法结果进行存储和展示，本文采用 Flask 框架搭建网站，设计并实现了一个自动化情感监测系统，使用 ECharts 库实现情感分析结果的可视化展示，可交互的动态曲线可以让用户实时监测到中日民众对某一事件的情感态度变化。针对情感突变点以及情感差值较大的区间，通过 TF-IDF 算法自动生成该区间的关键词同时给出热度前五的博文。经验证，生成的关键词和博文可以很好的对应情感突变点处发生的事件。

2 相关工作

舆情监测系统是一个网络应用系统，用于监测由热门事件或突发事件引发的有影响力且倾向性强的观点和言论[5]。同时，情感分析技术也被广泛运用于舆情监测的研究中，成为当前舆

情监测研究中的主流方法之一。21 世纪后，情感分析在自然语言处理的各个领域都有广泛的研究，如数据挖掘、文本挖掘、舆情监测和信息检索等。情感分析最早是在 2003 年由 Nasukawa 等人[6]提出，通过对文本中包含的情感进行计算，进而分析用户的情感倾向和观点。

对文本数据进行情感分析主要有三种，分别是：基于有监督的情感词典的情感分析方法、基于无监督的机器学习的情感分析方法以及基于深度学习的情感分析方法。三种方法的对比如表 1 所列。

表 1 情感分析方法对比

方法	优点	不足
基于情感词典	不需要人工标注，可以用简单的统计方法进行情感分类	分类的结果依赖词典的质量和规模
基于机器学习	适用于较小的数据集，泛化能力强	需要投入大量的人工成本来标注数据集
基于深度学习	学习能力强、覆盖范围广、适应力强、可移植性好	计算量大、硬件成本较高、模型设计复杂

2.1 基于情感词典的情感分析方法

基于情感词典的方法，是在标注极性或极性分数单词的基础上，通过比对情感文本中包含的极性情感词，然后采用简单统计的方法或权值算法进行情感分类。因为此方法不需要训练数据，被广泛地应用于传统的文本情感分析中。在 20 世纪 90 年代末期，国外开始了有关文本情感分析研究，Riloff 和 Shepherd[7]基于语料数据构建了语义词典；熊德兰等人[8]基于 HowNet 研究了句子的褒贬性；潘明慧等人[9]提出了基于词典的方法识别出微博表达的 6 种情绪。

2.2 基于机器学习的情感分析方法

基于机器学习的方法也被广泛应用于情感分析领域。该方法首先要建立一个训练集，并根据用户情感去标记数据；然后从训练集中提取特征，构建分类模型，进而预测没有标签的数据，最后通过分类器对未标记的数据进行情感倾向性判定。在国外，Pang 等人[10] 使用了三种机器学习的方法进行对比试验，分别是朴素贝叶斯、最大熵和支持向量机，对电影评论进行了情感极性分类，将其分为积极和消极两种情感，并比较三种方法的实验结果，其中支持向量机方法的分类效果最好。国内，也有很多学者比较不同的分类算法，杨艳霞[11]使用两种机器学习方法对微博数据集进行了情感分析，分别是贝叶斯和支持向量机，同时比较了在分类性能上两种算法的优劣，其中贝叶斯算法的准确率更高。

2.3 基于深度学习的情感分析方法

深度学习最早是在 2011 年 ColloBERT R 等人[12]在解决词性标注等问题时将其应用到了自然语言处理领域。其最大的特点是可以自动学习批量数据，从而挖掘数据中的潜在特征，并通过注意力机制实现对目标内容的增强关注，在训练过程中进行参数的调整[13]。Mousa E Dd 等人[14]引入了一种基于长短期记忆循环神经网络语言模型的新方法，不需要任何特殊的预处理或特征选择。宋婷等人[15]为提取方面级的情感，提出了分层的 LSTM 模型。徐志栋等人[16]提出一种基于胶囊网络的方面级情感分类模型—SCACaps，解决了方面级情感分析中多重情感造成的特征重叠问题。张宝华等人[17]提出了一种多输入模型，该模型结合了 MCNN、LSTM 和全连接神经网络。而深度学习中的迁移学习(Transfer Learning)也常常应用于舆情分析领域，比如利用美团外卖的评论数据作为原始数据集，抽取其特征，建立美团外卖评论的情感分析模型，再将其应用到相应的目标域中，如电影评论的情感分析，以此实现模型的大规模迁移。基于此，迁移学习逐渐成为舆情分析领域的研究热点。Radford 等人[18] 提出了名为 OpenAI GPT 的预训练模型，该模型可以通过的少量的微调后用于各种下游任务。近年来，BERT 模型作为一个强大的预训练模型，首先在大规模的语料库上进行预训练，获取通用的语言模型，然后进行一系列的微调以吸收下游具体任务的相关知识[19,20]。但是在情感分类任务上，BERT 模型还存在一定的提升空间，这是因为 BERT 在预训练阶段并没有考虑任何情感信息。为了解决这个问题，本文将融合情感词典和 BERT 模型，将情感特征引入预训练过程中。

3 数据来源与数据预处理

3.1 数据来源

利用 python 语言进行编程，使用新浪微博 API 接口、日本 Twitter API 接口、网络爬虫等技术完成了整个舆情数据的获取。

（1）中文舆情数据

在中国，新浪微博(简称“微博”)具有用户多、消息数量大、更新快等特点，成为人们获取信息、发表舆论的主要途径，越来越多的民众习惯于在微博这一社交网络平台上交流观点、分享信息。这些信息包含了大量的社会热点及情感，能很好的反映民众对话题的关注和态度。本研究以“日本”为关键词，通过网络爬虫技术，爬取了 2013 年—2021 年的舆情数据。数据来自于科

技、体育、娱乐、经济、疫情五个类别，数据主要包括微博标题、URL、时间、内容、点赞数、评论数及转发数等。

（2）日文舆情数据

在国外，Twitter 无疑是访问量最大的社交网络平台之一。不仅是普通民众，许多名人也都通过 Twitter 发布消息与民众进行互动。在日本，Twitter 作为互联网 Web2.0 时代的最新应用，逐渐影响和改变世界的交流和沟通的方式。因此，Twitter 数据十分适合进行国外舆情分析。本研究以“中国”为关键词，通过网络爬虫技术，同样爬取了 2013 年—2021 年的舆情数据。数据来自于科技、体育、娱乐、经济、疫情五个类别，数据主要包括推文标题、URL、时间、内容、点赞数、评论数及转发数等。

3.2 数据预处理

微博和 Twitter 数据文本含有很多标签、注释等特殊符号，使用 Python 自然语言处理工具包 NLTK 和正则表达式等工具对数据进行清洗。然后，由于情感分析的质量依赖于情感词典，因此必须对清洗后的数据做分词处理，同时移除停用词。本文采用 jieba 作为中文分词工具，MeCab 作为日文分词工具。主要处理过程包括：中文分词、日文分词、提取词元（token）、词根化（stemming）、移除停用词等。

其中微博的中文数据集采用“百度停用词表”进行过滤。由于现有的日文停用词表中停用词较少并且不够全面，因此，我们在现有的停用词基础上扩充了新的日文停用词表用于 Twitter 日文数据集上。

4 基于自扩展情感词典的情感分析模型

微博和 Twitter 中的文本具有领域广、更新速度快等特点，而通用情感词典存在着领域差异、知识覆盖率较低、情感词权值过于固定等问题，因此，本文利用自扩展的中日文情感词典提出一种情感极性量化算法，用以计算文本的情感强度值。通过计算情感词权值来量化该文本的情感强度。具体方案如下：

首先，在通用情感词典的基础上，构建适合于本研究领域的中日文情感词典，并对前面处理好的数据进行情感倾向性分析。然后分别构建中文和日文的程度词表和否定词表，之后对于

特殊标点符号进行量化加权。最后，将点赞数也考虑进来，对情感值进行加权计算，得到最终的情感分值,并进行情感分类。情感分析整体框架如图 1 所示。

4.1 情感词典构建

情感倾向，是用户对某一事物主观的内心喜恶，以及主观评价的一种倾向。不同的情感词或情感语气可以表达不同程度的情感倾向。通常给每个情感词赋予不同的权值。例如：“楽しい”和“嬉しい”，都是表达开心，但是“嬉しい”要比“楽しい”在表达情感程度上要强烈。中文的“讨厌”与“厌恶”都是表达消极情感，但是“厌恶”的情感程度会更强烈。因此，情感词典能否覆盖全面在一定程度上影响着情感分类结果，情感词典的构建是情感分析研究的基础，本研究尽可能构建一个足够大、覆盖面广的情感词典应用于中日舆情研究领域。

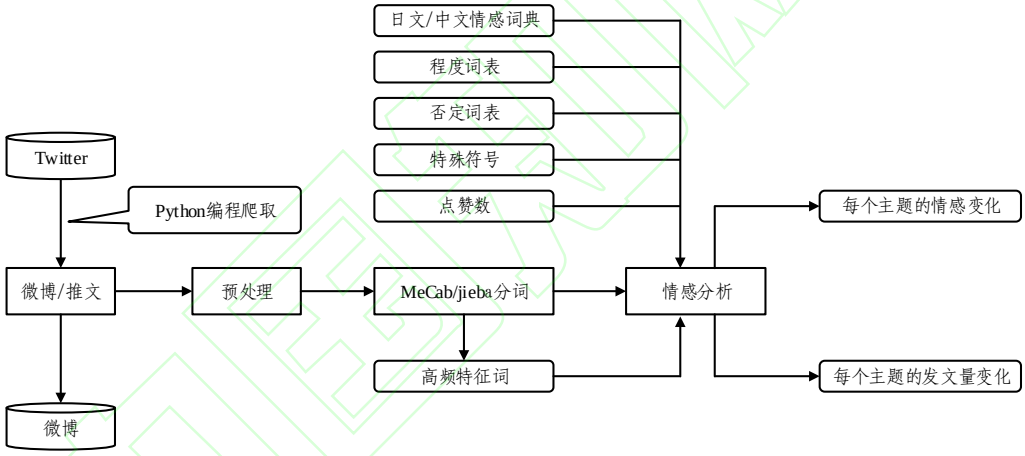

图 1 情感分析框架图

(1)基础情感词典：总结并整理当前已有的情感词典资源。对于中文情感词典，我们采用了 HowNet（知网）情感词典作为中文基础情感词典，如表 2 所示。对于日文情感词典，我们也集成了多个开源的日文情感词典用于 Twitter 数据集。

表 2 情感词典

类别	个数
积极评价词语	3730 个
消极评价词语	3116 个
积极情感词语	837 个
消极情感词语	1255 个

(2)网络情感词典：随着互联网的高速发展，网络用语应运而生。网络用语的形式和传统词语有着很大区别，比如：“yyds”、“绝绝子”、“emo”等等，它们往往具有强烈的感情色彩。这些词语是不包含在基础情感词典当中的，但在判别情感倾向时起着重要的作用。

(3)领域情感词典：从微博和 Twitter 中获取的中日舆情数据集中选取了适合于本研究领域、情感鲜明的词作为基准词，通过基于扩展的情感倾向点间互信息（SO-PMI）算法计算候选词与基准词的相似度，以此判断候选词的情感倾向，将领域情感词自动加入到基础情感词典中，构建了适用于中日舆情领域的中日文情感词典。

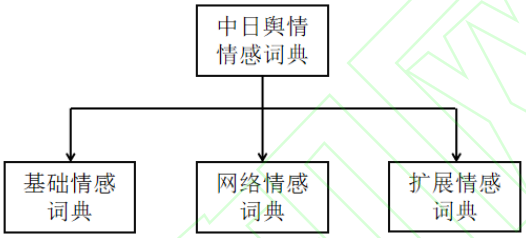

图 2 情感词典组成图

情感词典中积极词语分值为 1，消极词语分值为-1。假设一个句子中包含 *pos_num* 个积极词语和 *neg_num* 个消极词语，那么该句子的情感分值为：

$$pos\_score = pos\_num \tag{1}$$

$$neg\_score = neg\_num \tag{2}$$

$$score = pos\_score + neg\_score \tag{3}$$

4.2 程度词表构建

我们发现在网民发布的微博和推文中，情感词语前大都包含副词修饰，例如在“非常喜欢”、“很喜欢”中，“非常”修饰“喜欢”，“很”也修饰“喜欢”，但是，“非常”所表达的情感强度显然要多于“很”。为了更加准确地计算文本的情感倾向，我们构建了程度词表，并将程度副词分为 5 个等级：极强级、中强级、中级、中弱级、微弱级，并将程度副词的强度取值范围限定在[0，3]。并人工标注这些程度副词语气的强弱，用一个二元组表示为 *level*，其中 *adv* 表示词语名称，*intensity* 表示该词的语气强度，一个副词的语气强度 *inte* 取值范围在 0-3 之间，越接近 0 说明该词表达的情感强度越弱，越靠近 3 说明该词表达的情感强度越强烈。如：“出头”的强度设置为 0.5，“更”的强度设置为 2，“极其”的强度设置为 3。同样，我们也构建了日文程度词表，并人工对这些程度词的语气强度进行了标注，与中文的构建方法相同，此处不再赘述。

如果一个积极词语前后出现程度词,那么：

$$pos\_score = pos\_score * inte \quad (4)$$

如果一个消极词语前后出现程度词，那么：

$$neg\_score = neg\_score \times inte \quad (5)$$

4.3 否定词表构建

在微博和 Twitter 的文本中，否定词也是经常出现的。例如：“不公平”、“拒绝接受”，其中“不”用来否定“公平”、“拒绝”用来否定“接受”。为了使得文本情感倾向性的计算更为准确，我们分别构建了中文否定词表和日文否定词表。对于否定词表，我们不需要对其进行标注，用一个列表表示为 *list*，其中 *deny* 表示否定词名称，当一个积极词语前出现一个否定词，则该词语的情感分值变为原来的相反数。如果是双重否定，则该词语的情感分值不变。消极词语同理。

$$pos\_score = -1 \times pos\_score \quad (6)$$

$$pos\_score = -1 \times -1 \times pos\_score \quad (7)$$

4.4 感叹句

除此之外，我们认为带有“！”的感叹句往往比陈述句的语气更强烈。因此，我们定义了感叹句的加权计算公式：

$$score = score \times 1.2 \times n \quad (8)$$

如果一句话是感叹句，那么对该句的情感分值进行加权，其中 *n* 为感叹号的数量。

4.5 点赞数

对于微博和 Twitter 这类热门的社交网络平台，我们认为，一篇微博或推文的点赞数能很好的说明其他网民对该观点的支持度。即点赞数越多的文本，应该赋予其更高的权重。因此，我们将点赞数映射到 0-1 之间，对句子的情感分数加权计算，可以得到会更准确的分析结果。

4.6 TF-IDF 关键词抽取

TF-IDF（Term Frequency–inverse Document Frequency，词频-逆文档频率），是一种常用来计算一个字或词语对于一篇文档的重要程度的统计方法。如果某个单词在一篇文章中频繁地出现，但在其他文章中很少出现，那么就认为该词或者短语对于该文章具有一定的代表性，适合用来分类。

$$TF = \frac{某个词在文章中的出现次数}{文章的总词数} \quad (9)$$

$$ID = log\left(\frac{语料库的文档总数}{包含该词的文档数+1}\right) \tag{10}$$

$$TF-IDF = TF \times IDF \tag{11}$$

假设一篇微博的共有 100 个单词，而“会议”这一单词出现了 5 次，那么“会议”一词在该微博中的词频(*TF*)就是 5/100=0.05。而计算逆文档频率(*IDF*)的方法是测定有多少篇微博出现过“会议”一词，然后除以数据集中的总微博数。所以，如果“会议”一词在 100 条微博中出现过，而文件总数是 10,000,00 份的话，那么，其逆文档频率(*IDF*)就是$\log\left(\frac{1000000}{4}\right) = 0.2$。最后的 *TF-IDF* 的分数为 0.2。

一篇文章中某个字或词语的 TF-IDF 越大，那么一般来说这个词对于这篇文章越重要，因此通过计算文章中每个词的 TF-IDF，按大小排序，排在最前面的几个词，就是可以代表该文章的关键词。

基于前面的情感分析方法，可以得到每个主题的情感变化情况。接下来，通过 TF-IDF 算法自动生成各个时间区间内词频最高的 10 个关键词，关键词可以体现出情感变化的原因以及引发情感突变的事件。

5 EmoBERT-结合中日文情感词典的情感分析模型

本文以 BERT 预训练模型为基础，利用自采集的中日舆情数据集，提出一种结合中日文情感词典的情感分析模型。

5.1 BERT 模型及其缺陷分析

BERT 模型(Bidirectional Encoder Representations from Transformers)，是一种基于双向 Transformer 的大规模预训练语言模型，其利用了 Transformer 的编码层，并在此基础上加入掩码机制，使其能够自动进行预测训练。“双向”是指它在处理一个词的时候，可以考虑到该词前后的单词的信息，进而得到上下文的语义。本质上是在大规模数据集的基础上，使用自监督学习方法对单词学习进行模型训练[21]。

图 3 展示了 BERT 模型结构，可大致分为输入层、Transformer 层以及输出层三层。其中 E_i 表示 BERT 模型输入的编码向量，T_i 表示 BERT 模型输出的编码向量。T_{rm} 就是 Tranformer 的编码器结构。

BERT 输入层的编码向量包含三种嵌入特征，分别是词嵌入、段嵌入和位置嵌入，如图 4 所示。为了使得 BERT 模型适应下游的任务，在输入时，为每个句子附加[CLS]和[SEP]，这是两个特殊符号：[CLS]用于下游的分类任务，最终输出时可以用来表征整个句子；[SEP]用来分割两个句子，如：[CLS]+句子 A+[SEP]+句子 B+[SEP]。

(1)词嵌入(Token Embeddings)：从词汇表学习得到的每个特定词的嵌入特征，Token Embeddings 层会将每一个 wordpiece token 转换成 768 维的向量,如图四的句子会被转换成一个(10,768)的矩阵。

(2)段嵌入(Segment Embeddings)：用来区别两种句子，E_A 表示第一句话，E_B 表示第二句话。文本中的多个句子被拼接在一起后送入到模型中，BERT 通过 Segment Embeddings 去区分每个句子。

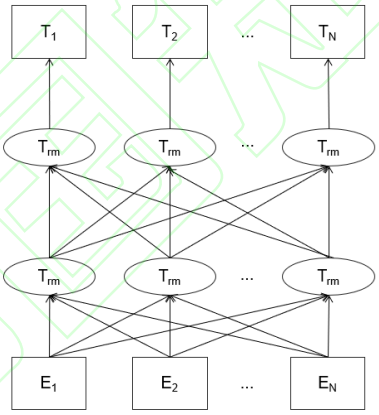

图 3 模型结构

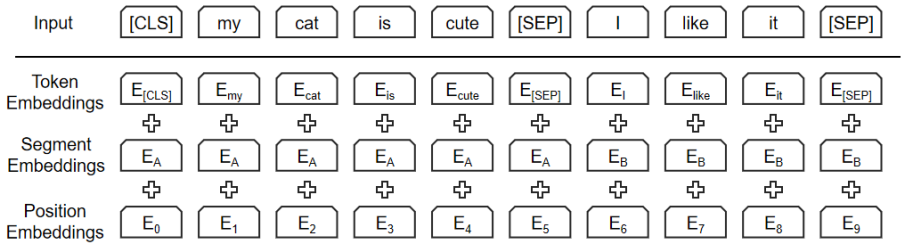

图 4 BERT 模型输入特征向量

(3)位置嵌入(Position Embeddings)：指将单词的位置信息编码成特征向量的形式，将单词位置关系引入模型中，BERT 通过学习得到位置向量，实际上，Position Embeddings 是一个大小为(512,768)的查找表，其中第 i 行是指第 i 个位置上的任意单词的向量表示.

将三种嵌入特征进行简单相加，得到模型的输入向量，同时传递给 BERT 的编码层作为输入表示。

$$\text{Input Embeddings} = \text{TE} + \text{SE} + \text{PE} \tag{12}$$

然而，尽管输入层中嵌入了词向量、段向量和位置向量，使其能够获取一定的句法和语法信息，用以在 Transformer 层进行掩码预测，但是在面对情感分类任务时，由于缺乏情感特征，预测效果并不如其他的分类任务。这使得在情感分析任务上的预测效果还有待优化和提升，尤其是在一些情感词显著的样本中，情感词的特征无法被充分提取，因而无法发挥其价值。

表 3 展示了 BERT 对某两个样本预测时的结果，其中第一个文本中出现了“繁杂混乱”，第二个文本中出现了“危害”，这两个词都是具有明显情感倾向的情感词，但 BERT 在预测时反而给出了相反的结果，这说明 BERT 的 Multi-Head Attention 机制并没有给予情感词更多的关注，导致包含一个乃至多个情感词的文本无法利用其具备情感标签的优势。这使得情感任务的预测准确率并不理想。

表 3 部分预测结果

文本	结果
移动电源已经成为使用智能手机用户常用的配件，但是繁杂混乱的产品却让用户不知如何选择。	积极
抽烟的危害性众所周知，但仍无法做到有效地制止。	积极

事实上，当下游任务为情感任务时，在预训练阶段模型的注意力机制应该把更多注意力放在情感词上，使得模型可以更好地提取整个文本的情感特征。因此，本文提出一种情感词增强的注意力算法。

5.2 情感词增强的注意力机制

Transformer 层利用了 Self-Attention 机制来帮助理解上下文语义，多头是指允许模型可以学习到不同表示的子空间里的相关信息。实际上，Self-Attention 机制是一种分配机制，由于文本

中的每个词都与句子中的其他词进行联系，因此，注意力机制会根据对象的重要程度以及和句子中其他词的关联性，重新分配权重。

基于 Self-Attention 的这一性质，本文利用自扩展的中日文情感词典改进了 Attention 的计算规则，提出了一种更注重情感词增强的注意力算法，以突出对象的情感特征。

图 5 展示了 Transformer 的编码器结构，Transformer 由两个子层组成，分别是多头自注意力机制和前馈神经网络，每个子层后连接了一个规范化层及残差单元对输出进行控制，使向量的标准差和均值均为一个固定的数值。输入层的数据和 Multi-Head Attention 层输出的结果进行残差相加后进行标准化，经过反馈层之后，再进行上述的环节，最后输出结果。其中，多头自注意力机制是 Transformer 层的核心，输入层的三个箭头分别对应 Multi-Head Attention 的三个输入向量，分别是 Q(query)、K(key)、V(value)，这三个向量是由输入层的 Word Embeddings(X) 分别和一个矩阵相乘得到的，三个向量通过计算 *score* 及 SoftMax 归一化处理后得到 Attention 的计算公式如下：

$$Attention(Q,K,V) = softmax(\frac{QK}{\sqrt{d_k}})V \qquad (13)$$

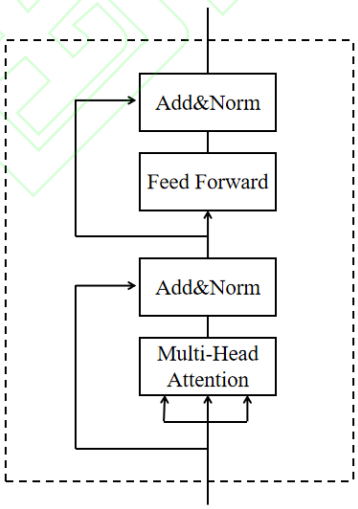

图 5 Transformer 编码器结构

根据 Attention 的公式可已看出，一个词的注意力一定程度上会受到 Q、K、V 三个矩阵的大小因素影响。而 Q、K、V 的值是通过原始的 word embeddings 得到的，它相当于底层的特征信息。因此，在情感分类任务中，为了情感词能得到更大的关注，本文提出一种情感词增强的注意力算法，希望通过以下两个方面增强情感词的注意力。

(1)增强情感词的 word embeddings，即 X 向量，进而增大 Q、K、V 的值，来增强该词的注意力，从而提升预测的准确率。利用在第 4 章构建的目标领域的情感词典作为输入层的 Sentiment Embeddings，向模型中加入外部情感信息，将四种嵌入特征相加后，得到模型的输入向量如下：

$$Input\ \ Embeddings = TE + SE + \ \ PE + S'E \tag{14}$$

(2)将情感词典中情感词的分值赋给模型作为额外的权重。直觉上，情感词表达的强度越强烈，越能代表其所在句子和文本的情感倾向，越应该得到更大的注意力权重。例如，对于“这家餐厅的服务太差劲了，菜再好吃我也不来了”这一样本，由于“差劲”的情感强度较强，因此很大概率是整个文本的情感倾向，应该给予更多的关注。而情感强度相对平和的情感词，虽然本身具备情感倾向，但是相比于情感强度大的词，上下文转折的可能性会更大。例如，对于“这家餐厅的服务不怎么样，考虑到菜品做得太美味了，下次还来”这一样本中的情感词，模型应该把注意力更多放在“美味”上，而非“不怎么样”。因此，本文重新定义了针对于情感任务的 Attention 算法，公式如下：

$$Attention = softmax\left(\frac{QK}{\sqrt{d_k}}\right) * (1 + \lambda)V \tag{15}$$

其中 λ 表示为对情感词的分值归一化处理后的(0,1]之间的数值，给模型的注意力机制量化加权。

5.3 结合情感词典的预训练模型

本文利用自扩展的情感词典提出了一个新的情感分析模型—EmoBERT，模型结构由三部分构成，如图 6 所示。

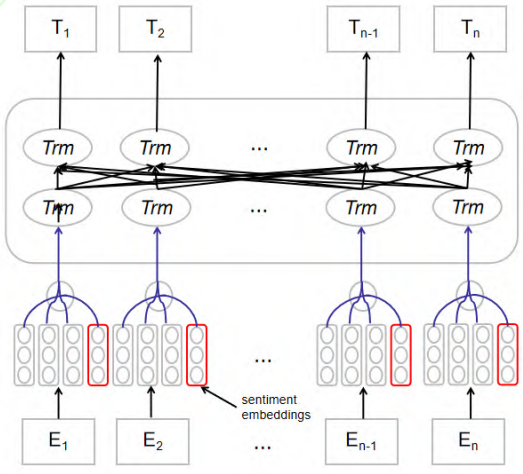

图 6 EmoBERT 模型结构

由于本文改进了 Attention 机制，因此，在原始 BERT 模型基础上，进行了进一步的预训练，预训练任务如下：

(1)输入层：输入层中每个 token 额外加入了情感向量，与其他三个嵌入向量相加，组成了具备情感特征的 Word Embeddings。

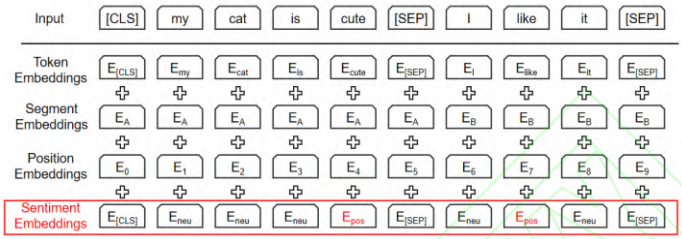

图 7 BERT 模型输入特征向量

(2)Transformer 层：在这一层，根据本文提出的情感词增强的注意力机制，可以求出每个子空间的注意力分值，进而计算出多个子空间的输出结果，公式如下：

$$head_i = Attention(QW^Q{}_i, KW^K{}_i, VW^V{}_i) \tag{16}$$

$$MultiHead(Q,K,V) = Linear\left(W_i concat\left(head_1, head_2, \ldots head_n\right) + b\right) \tag{17}$$

(3)输出层：输入层的 Word Embeddings 和 Transformer 层输出的结果残差相加,再经过进行标准化，经过反馈层之后，再重复上述的过程，最后输出结果。子层最后得到的输出结果如公式 18 所示：

$$SubLayer_{output} = LayerNorm\left(X + SubLayer(X)\right) \tag{18}$$

由于 BERT 是在通用数据集上训练的，在本特定领域的任务上，原始模型无法完全抽取出 token 的内在含义。因此，需要用本领域语料对其进行微调。此外，在微调前使用目标领域数据集的数据对模型进一步预训练，相当于在预训练阶段实现将模型从通用领域向特定领域提前迁移，然后再执行目标领域的任务[22]。

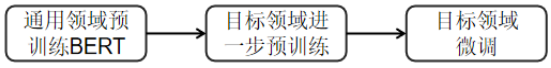

图 8 模型流程图

6 实验结果

为了验证本文提出的EmoBERT模型在情感分类任务上的表现，基于自采集的微博和Twitter数据集进行试验。

6.1 数据集

本文使用自采集的微博数据集和Twitter数据集作为中日舆情研究的数据集，类别包含娱乐、体育、经济、科技和疫情。两平台的数据集的信息如表4所列。

表4 数据集

类别	微博	Twitter
娱乐	40768	8603
体育	19843	8788
经济	16808	6266
科技	35538	8746
疫情	7520	3198

由于自采集的微博和Twitter数据集数据量庞大，并且不包含情感类别标签，所以分别从每个类别提取1000条数据并人工标注其情感极性。在加载好实验所需要的数据集之后，需要对这两个数据集中的数据进行训练集和测试集的划分，划分之后训练集和测试集的比例为8:2。各个数据集的训练集和测试集包含的情感极性信息如表5所列。

表5 数据集情感极性信息

Datasets		Positive	Negative	Total
微博	Train	1894	2106	4000
	Test	473	527	1000
Twitter	Train	1768	2232	4000
	Test	442	558	1000

6.2 超参数设置

预训练阶段，批量大小设置为256，学习率为5e-5，持续训练1000000步。模型中所有 dropout 概率都为 0.1，Adam β1 和 Adam β2 的值分别为 0.9 和 0.999，L2 权重衰减为0.01。

微调过程中，对批量大小、学习率和训练周期数量等进行了一定调整。其中 batch size 的值是每个 epoch 训练的句子数。如果设置过小，会使训练时间延长；如果设置过大，损失函数曲线比较平坦时，将无法得到最优模型。不同的下游任务对应着不同的最佳超参数值，为了让模型得到最佳分类效果，本文经多组实验验证最优的参数值如表 6 所示。

表 6 微调参数设置

	参数名	值
1	batch size	64
2	learning rate	5e-5
3	epoch	3

6.3 实验结果及模型对比

为了评估本文提出的 EmoBERT 模型在情感二分类任务上的分类效果，针对自采集的中日文舆情数据集分别进行了实验，同时和原始 BERT 模型以及领域迁移后的模型进行了对比实验，实验过程中的准确率变化如图 9、图 10 所示。

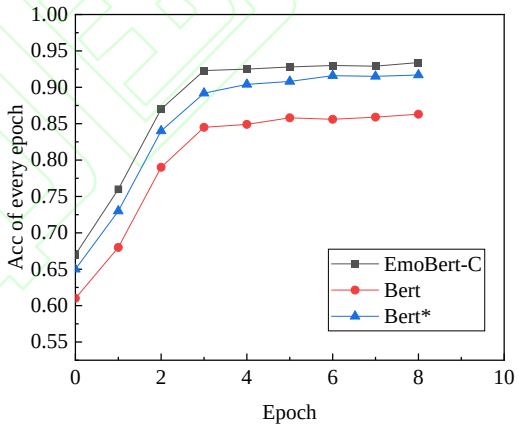

图 9 EmoBERT-C 及原始模型准确率对比图

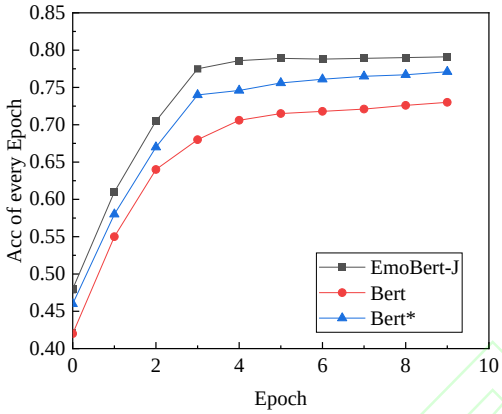

图 10 EmoBERT-J 及原始模型准确率对比图

从图中可以清晰地看出准确率的对比效果，本文提出的 EmoBERT 模型在中文情感分类任务(EmoBERT-C)和日文情感分类任务(EmoBERT-J)中，跟原始 BERT 模型相比都有一定的提升。其中 EmoBERT-C 相比于原始模型，准确率从 89.68%提升到 92.15%。EmoBERT-J 相比于原始模型，准确率从 74.73%提升到 78.26%。其中，EmoBERT-C 是中文模型，EmoBERT-J 是日文模型，BERT*为在目标领域数据集上预训练之后的实验结果。

此外，针对中日文舆情数据集中的积极文本和消极文本也分别进行了实验，同时和原始 BERT 模型进行了对比，其准确率和 F1 值对比情况如表 7 所示。从表中可以看出本文提出的 EmoBERT 模型在积极样本和消极样本中分类的效果均优于原始 BERT 模型，进一步验证了模型的有效性。

表 7 对比实验结果

		Accuracy	F1
BERT-C	positive	89.81	83.26
	negative	89.55	82.83
EmoBERT-C	positive	91.83	83.15
	negative	92.47	84.02
BERT-J	positive	75.92	70.34
	negative	77.04	71.28
EmoBERT-J	positive	77.92	71.53
	negative	78.60	72.96

6.4 中日舆情分析--“新冠疫情”

新冠肺炎疫情带来的影响是多元复杂的。由于病毒的攻击具有无差别性、跨国性且具有极大的不确定性，疫情给全球的经济和金融市场造成了强大的冲击，并且很大程度上催化了国际关系的演变。因此，利用该数据集分析相关的中日舆情是很有意义的。

在实验中，我们以“新冠肺炎”为关键词，采集了疫情初期 2020 年 1 月—2020 年 9 月的舆情数据。模型对数据进行情感分析，结果如图 11、图 12 所示。从图中我们可以看到，两国民众的情感态度普遍是积极的、且有稳步上升的趋势。这是由于疫情的跨国性影响以及在共同抗击疫情目标的驱动下，两国民众对 “命运共同体”的相互认知得到了增强。其中 2 月和 3 月两平台发文量非常大，两平台对应时间的关键词如表 8 表 9 所示，通过分析可知那段时间是春节期间，人口流动巨大且新冠疫苗还未研制成功，因此正是疫情的暴发期，也是民众讨论最多的时期。

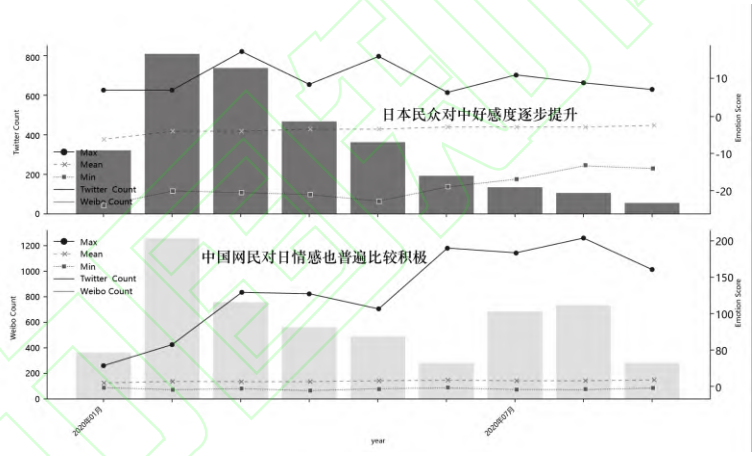

图 11 中日民众情感变化对比图

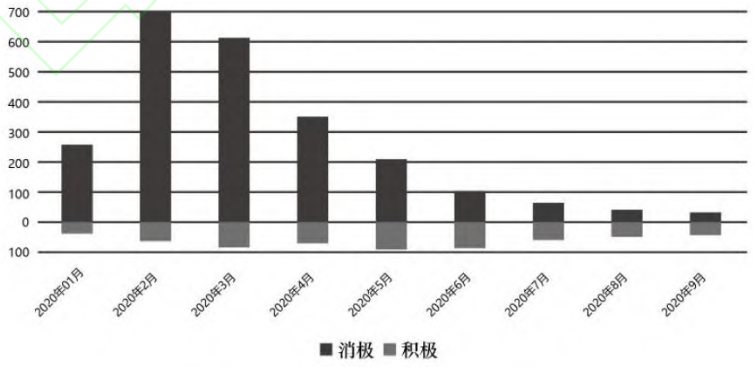

图 12 中日民众情感极性分布图

表 8 微博新冠肺炎关键词

时间	关键词

2020/02	肺炎，新冠，疫情，防控，患者，确诊，工作，病例，医院，武汉，感染，新型，隔离，冠状病毒
2010/03	肺炎，新冠，疫情，确诊，病例，防控，视频，患者，美国，累计，隔离，医院，新增，感染

表 9 Twitter 新冠肺炎关键词

时间	关键词
2020/02	新型，ウイルス，コロナ，感染，肺炎，拡大，武漢，死者，对策，对应
2010/03	コロナ，新型，ウイルス，感染，イタリア，武漢，拡大，肺炎，死者

7 系统构建及数据可视化

通过对中日民众情感检测系统需求整体分析，将该系统功能分成三大模块，分别为：用户交互模块、情感分析模块以及可视化模块。

7.1 用户交互模块

(1)实时采集数据并分析

在系统功能界面提供用户输入数据采集条件的接口，为了实时监测社交平台的数据，本系统服务端连接微博和 Twitter 等社交平台的 API 接口，以响应用户输入的关键词、详细程度、起始时间、结束时间等条件，从平台实时采集数据信息并以数据流的方式传输到服务器端，解析获得合法可分析的数据集。

(2)接收用户上传的数据并分析

为了满足相关研究领域的专业用户，本系统支持用户自行上传数据集。由于服务端情感分析算法的计算方式，数据集应包含每条社交平台微博/推文的具体内容、点赞数（喜欢数）以及发布时间，所上传的数据集最终以表单形式将数据集传给服务器端。

7.2 情感分析模块

服务器端作为自然语言处理的主要执行端，将以 python 为主要语言通过封装好的情感分析算法以及依赖系统预先保存的中日文情感词典、程度词表、否定词表等分析用户所提交或实时采集的数据集，本系统所使用算法将从三个方面进行情感分析。第一，根据用户指定的

关键词和数据集计算中日民众的情感值随时间的变化情况并实现实时监测。第二，对于每个点都能分析提取该时间区间热度最高的前十条微博和推文，方便分析中日民众情感分歧较大的原因事件。第三，根据算法分析结果，给出中日民众对于该主题的情感极性分布情况。三者最终通过 jinja 引擎中的模版函数动态渲染前端 HTML 模版文件中的 JavaScript 脚本变量，再通过 JSON 解析生成可用于展示的格式，最终传递到展示界面，动态生成用户所需要的分析结果。

7.3 可视化模块

本系统前端展示界面基于 css、html、JavaScript 脚本完成开发。展示模块包括三部分，均使用展示可交互的动态曲线图，可交互曲线图通过调用以 JavaScript 为主要语言编写的开源可视化工具包 echarts.min.js。该工具包的使用需要首先实例化曲线图对象，并将 JSON 解析出的统一格式情感分值数据传入曲线图 option 对象的 data 列表，将对应时间数据传入 option 中的 xAxis 的 data 数组，将每月或每日的文章条数分别渲染对应时间点 tooltips 的 formatter 方法，设置好曲线图的曲线颜色，横纵坐标，图像大小后，最终生成出围绕本次主题中日民众的情感值随时间变化的曲线。其中第一部分是微博和 Twitter 单独的可交互数据展示，包括以日/月为单位情感分值随时间变化曲线图，以日/月为单位情感极性随时间变化曲线图，热度前 100 文章情感散点图，内容来源年份分布扇形图以及词云。第二部分是微博、twitter 的情感随时间变化对比图，每一数值点还支持当鼠标点击时具体显示出影响该点的热度最高的前五条微博和推文。第三部分是根据用户输入来进行对微博某一特定关键词实时检测，显示实时变化的情感分值随时间变化曲线图。

7.4 系统架构

系统主体基于 Web 的 C/S 架构设计，以 Flask 为开发框架，其负责将 html 文件、JavaScript 脚本与后端 python 代码进行连接，同时 Jinja2 模版和 Werkaeug WSGI 套件负责前后端的数据传递。使本系统从规模上来说较为轻量，同时 Flask 框架高度的灵活性也降低了后续系统功能拓展和维护的成本。

客户端基于 Web 技术设计，一方面负责处理业务逻辑，另一方面负责返回响应内容。业务逻辑方面：支持用户在页面中输入主题关键词、详细程度、起始时间、结束时间等条件实时采集数据集或接收用户自行上传本地数据集文件，并通过 form 表单将主题关键词字符串或

合法数据集文件通过表单传递给服务器端进行解析处理。返回响应内容方面：在展示结果页面通过 Jinja2 模版引擎渲染模版，与服务端进行数据传递，接收服务器端传来的计算结果，客户端展示结果页面使用可视化工具包 Chart.js 生成可交互式情感值随时间变化曲线图以及情感极性分布图等来展示所分析的结果给用户。

服务器端以 Flask 框架为主体进行开发，通过不同 URL 来区分响应前端视图函数的不同表单请求，完成业务逻辑的具体功能实现。本系统主要包含三个不同 URL 绑定的功能函数。monitor()负责调用社交平台 API 接口，即时采集流式数据并保存到服务端等待系统进行进一步分析处理。uplouder()负责用户将本地数据集上传到服务器端，此外为了防止数据集所包含内容缺失或不兼容于本算法的运算所需条件，在用户上传数据集后首先会进行合法性检测，要求包含算法计算所需数据的完整性，保证了系统的稳定性。analyse()负责调用本地情感词典、程度词表并通过本地情感分析算法计算分析已经上传到服务器端的数据，计算结果通过 Jinjia2 模版引擎渲染动态前端系统客户端的 HTML 模版文件中的内嵌 JavaScript 变量，并进一步进行 JSON 解析，使用 eval()函数进行合法性检测。同时系统使用 Flask—Caching 扩展的缓存技术提高程序运行速度。

7.5 可视化实现

为了更清晰地展示系统的有效性，本文基于自采集的 2018 年以“RNG”为主题的中日文数据集对系统的功能进行演示，“RNG”是当下深受国内外游戏爱好者喜爱的电竞战队。传入数据集后，后台算法开始分析并将结果反馈给前端可视化界面，图 13 图 14 为微博和 Twitter 单一平台对于“RNG”主题的分析结果，包括情感极性分布、年份分布、热度情况分布、词云、每日情感变化图、每月情感变化图、每日发文量变化图以及每月发文量变化图等，图上的点均可以点击，用以和用户交互，可以展示当前点的详细信息，如：情感分值、积极人数、中立人数、消极人数等。

从图 13 中可以看出，中国民众对于“RNG”战队的讨论度很高，数据量达到了 38272 条。通过分析情感变化曲线图可以看出，共有 5 个情感波动较为明显的时间段。在 4 月 28 日、5 月 20 日、8 月 29 日以及 9 月 14 日前后，中国民众对于该主题的情感非常积极，同时通过柱形图也可以看出对应时间的发文量很高，通过点击曲线上对应的点可以看到当天的热门微博。以 9 月 14 日为例，当日的热门微博如图 15 所示，可以看出当天 RNG 战队在赛事中夺冠。其

他的点也根据热门微博验证了系统对于情感分析以及对应事件监测的有效性。

图 13 微博可视化展示界面

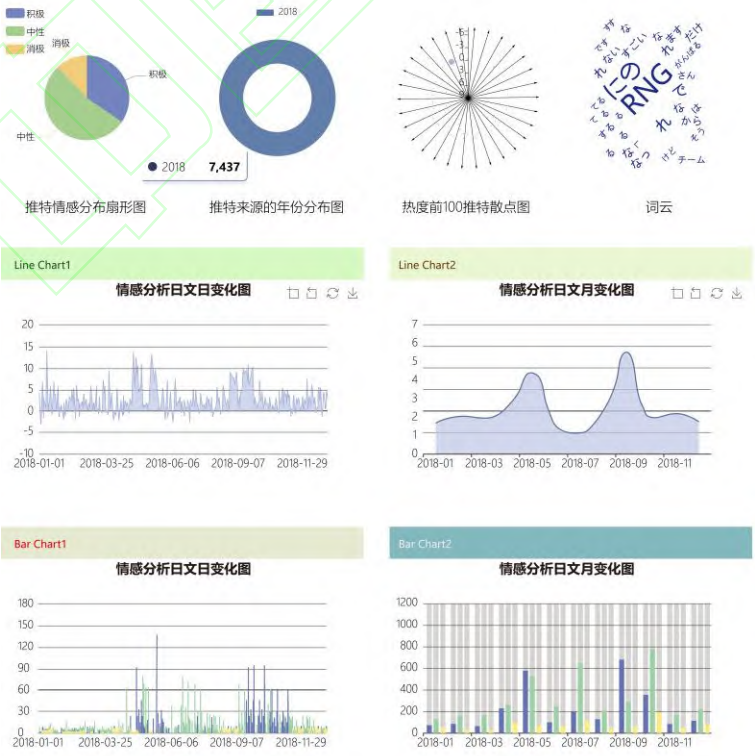

图 14 Twitter 可视化展示界面

热门微博内容	微博热度
#英雄联盟七周年# 【LPL夏季赛总决赛RNG赛后采访“更有冲劲去期待下一场比赛”】恭喜RNG!感谢两只队伍为我们奉上如此精彩的BOS大战，世界赛加油！OLPL夏季赛总决赛RNG赛后采访“更有冲劲去期待下一场比赛”	324
#心疼Rookie#rng两次夺冠，恭喜小狗大满贯！karsa别哭，你的很好！吧友神评：心疼一波rookie，明明就差一电，课冠军总是失之交臂，再坚持一下，这只崭新的ig值得期待。	192
rng夺冠了我很开心但是如果ig拿冠军我也不会难过就是了第五把看的我心脏不好但后面我看着看着觉得两个队的人都在拼命我觉得冠军给谁都很名副其实了看微博上那些因为ig暂停了游戏开玩笑说是王思聪打我的人觉得你真是个弱智，你这是在损自己喜欢的队伍人品，还是在看不起对面队伍的事例？脑子呢我喜欢rng，但不会不尊重他的对手希望两只队伍都能S8加油希望这次韩国的决赛场上能是两个中国队伍rng加油 ig加油	129
Drng他们是彼此忠贞不渝的光	120
再次恭喜RNG 今天大家辛苦了!!!S8冲冲冲！#RNG夺冠#英雄联盟7周年#2南京	107

图 15 情感突变点当日热门博文

从图 14 中可以看出，日本民众对于“RNG”主题的讨论度较少，一年中共爬取到 7437 条相关数据。同时，通过分析情感变化曲线可以看出日本民众对于该主题的态度普遍比较积极，且比较平稳。图 16 展示了微博和 Twitter 两平台中日双语料的对比分析结果，针对中日民众情感分歧较大的点，同样可以通过点击对应的点查看当日两平台关于该主题热度最高的博文，便于用户进一步了解情感分歧的原因事件。

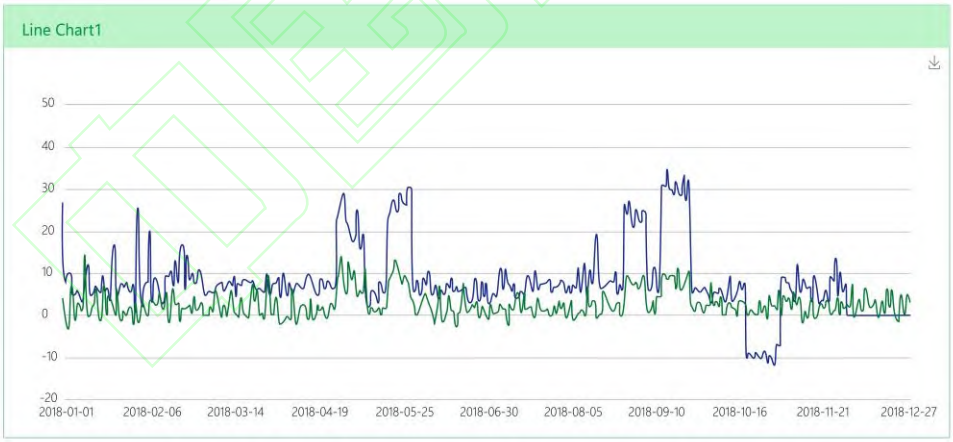

图 16 两平台对比结果可视化展示界面

此外，本系统还提供话题情感的实时监测功能，针对用户输入的关键词进行实时监测、实时分析，并实时呈现给用户，如图 17 所示。

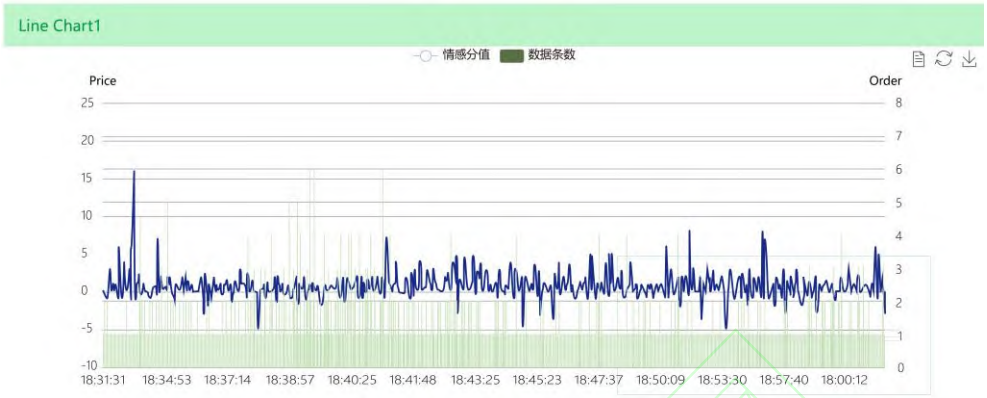

图 17 话题情感实时监测界面

8 结论及进一步工作

微博、Twitter 等社交网络平台的流行使得其中蕴含了丰富的情感信息，通过对这些平台上的用户发布的内容进行情感分析，可以挖掘其中有意义的社会价值。我们采用了基于自扩展的情感词典结合改进的 BERT 预训练模型进行了实验，我们的系统可以同时分析中日舆情数据，并自动生成中日民众情感态度对比和情感极性分布，针对情感突变点也能通过分析热点微博和推文来有效地分析出相应的事件，并将分析结果以可视化的形式展现给用户。因此，我们的系统是合理的、有效的且解决了目前单一平台、单一语料舆情监测的缺陷。未来工作将从以下方面进行：

(1) 社交网络平台用户的情绪比较丰富，应从多个方面分析情感词，不能局限于积极和消极的二分类，应进一步延伸对情感情绪的等级判断。

(2) 舆情监测系统也不应只局限于中文和日文两种语料，未来会支持更多种语料进行分析。

(3) 只对文本进行了情感分析，但现实世界中除文本外的，如图片、视频、语音等信息也会包含强烈的情感倾向，系统应实现多模态的情感分析。

参 考 文 献

[1] 敦欣卉，张云秋，杨铠西. 基于微博的细粒度情感分析[J]. 数据分析与知识发现，2017，001(7)：61-72.

Dun X, Zhang Y, Yang K. Fine-grained Sentiment Analysis Based on Microblog[J]. Data Analysis and Knowledge Discovery, 2017, 001(7): 61-72.

[2] Zhao J, Dong L, Wu J, et al. MoodLens: An emoticon-based sentiment analysis system for chinese tweets. ACM, 2012.

[3] Wang H, Can D, Kazemzadeh A, et al. A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle[C]// Acl System Demonstrations. 2012.

[4] Williams J, Katz G. Extracting and modeling durations for habits and events from Twitter[C]// Meeting of the Association for Computational Linguistics: Short Papers. 2013.

[5] 李忠俊. 基于话题检测与聚类的内部舆情监测系统[J]. 计算机科学，2012，39(12)：237-240.
LI Z J. Internal Public Opinion Monitoring System Based on Topic Detection and Clustering [J]. Computer Science, 2012, 39(12): 237-240.

[6] Yi J, Nasukawa T, Bunescu R, et al. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques[C]// Third IEEE International Conference on Data Mining. IEEE, 2003.

[7] Riloff E M, Shepherd J. A Corpus-Based Approach for Building Semantic Lexicons[J]. 1997.

[8] 熊德兰，程菊明，田胜利. 基于 HowNet 的句子褒贬倾向性研究[J]. 计算机工程与应用，2008，(22)：143-145.
Xiong D, Cheng J, Tian Shengli. A Study of Sentence Praisal and Derogation Tendency Based on HowNet [J]. Computer Engineering and Applications, 2008, (22): 143-145.

[9] 潘明慧，牛耘. 基于多线索混合词典的微博情绪识别[J]. 计算机技术与发展，2014，(9)：28-32.
Pan M H, Niu Y. Microblog Emotion Recognition Based on Multi-cue Hybrid Dictionary [J]. Computer Technology and Development, 2014, (9): 28-32.

[10] Pang B. Thumbs up Sentiment Classification Using Machine Learning Techniques[J]. Proc. EMNLP, Philadelphia. PA, USA, 2002.

[11] 杨艳霞. 基于分类的微博情感分析算法研究及实现[J]. 计算机与数字工程，2017，45(2)：197-197.
Yang Y X. Microblog Sentiment Analysis Algorithm Research and Implementation Based on Classification [J]. Computer & Digital Engineering, 2017, 45(2): 197-200.

[12] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing(almost)from scratch[J]. The Journal of Machine Learning Research, 2011, 12:2493-2537.

[13] Al-Rifaie M M, Bishop J M. Swarmic Sketches and Attention Mechanism[J]. Springer, Berlin, Heidelberg, 2013, 85–96,.

[14] Mousa E D, Vryniotis V. Sentiment analysis and opinion mining: on optimal parameters and performances[M]. John Wiley & Sons, Inc. 2015.

[15] 宋婷，陈战伟，杨海峰. 基于分层注意力网络的方面情感分析[J]. 大数据，2020，6(5)：10.
Ting SONG, Zhanwei CHEN, Haifeng YANG. Aspect sentiment analysis based on a hierarchical attention network[J]. Big Data Research, 2020, 6(5): 10.

[16] 徐志栋，陈炳阳，王晓，等.基于胶囊网络的方面级情感分类研究[J]. 智能科学与技术学报，2020，2(3)：284-292.
XU Z D, CHEN B Y, WANG X, et al. Research on capsule network-based for aspect-level sentiment classification[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(3): 284-292.

[17] 张宝华，张华平，厉铁帅，等.基于多输入模型及句法结构的中文评论情感分析方法[J]. 大数据，2021，7(6)：41-52.
ZHANG B H, ZHANG H P, LI T S, et al. Chinese comment sentiment analysis method based on multi-input model and syntactic structure[J]. Big Data Research, 2021, 7(6): 41-52.

[18] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.

[19] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in neural information processing systems, 2017, 30.

[21] Sun C, Qiu X, Xu Y, et al. How to Fine-Tune BERT for Text Classification[J]. Springer, Cham, 2019.

[22] 杨晨，宋晓宁，宋威. SentiBERT:结合情感信息的预训练语言模型[J]. 计算机科学与探索, 2020, 14(9):8.

Chen Y, Xiaoning S, Wei S. SentiBERT: Pre-training Language Model Combining Sentiment Information[J]. Journal of Frontiers of Computer Science & Technology, 2020, 14(9): 8.

作者简介

李爱黎，出生于 1995 年，东北大学研究生，主要研究方向为情感分析与数据挖掘。
张子帅，出生于 2000 年，东北大学本科，主要研究方向为数据挖掘、机器学习。
林荫，出生于 1984 年，硕士，讲师，主要研究方向为中日文化比较研究。
王秋菊，出生于 1962 年，博士，教授，主要研究方向为中日文化比较研究、科技与文化（STC）研究。
杨建安，出生于 2002 年，东北大学本科，主要研究方向为数据挖掘、机器学习。
孟炜程，出生于 2002 年，东北大学本科，主要研究方向为数据挖掘、机器学习。

张岩峰，出生于 1982 年，博士，教授，CCF 高级会员，主要研究方向为大数据挖掘、大规模机器学习、分布式系统。