文章编号：2095-2163(2022)08-0060-06 中图分类号：TP391 文献标志码：A

基于文本挖掘的弹幕情绪分析研究

江 涛，黄昌昊，孙 斌
（西北民族大学 中国民族语言文学信息技术教育部重点实验室，兰州 730030）

摘 要：随着大数据和新媒体时代的到来，短视频+短评论的风潮愈发流行。相比于普通的视频评论，弹幕具有更强烈的情感倾向，蕴含着丰富的情绪价值。本文意在研究弹幕的情绪分类，依据大连理工大学情感词汇本体库，对爬取的 B 站弹幕进行人工数据标注，采用 Bert 预训练语言模型与循环卷积神经网络 BiLSTM_CNN，充分提取弹幕文本信息和句子语义特征，最终通过 *softmax* 函数得出弹幕文本的情绪分类，分类精度达到 84.6%，相比传统模型准确率得到显著提升。

关键词：弹幕；情绪分析；文本分类；预训练语言模型；卷积神经网络

Research on sentiment analysis of Danmaku based on text mining

JIANG Tao, HUANG Changhao, SUN Bin
(Key Laboratory of Information Technology for Ethnic Languages and Literature, Ministry of Education, Northwest Minzu University, Lanzhou 730030, China)

【Abstract】 With the advent of big data and new media era, the trend of short video and short commentary are becoming more and more popular. Compared with ordinary video commentary, Danmaku have stronger emotional tendency and contain rich emotional value. This paper intends to study the emotional classification of Danmaku, According to Dalian University of Technology emotional vocabulary, performs manual data annotation of the crawled BiliBili Website Danmaku, realizes full extraction of Danmaku text information. Sentence semantic features using Bert pretrained language model with recurrent convolutional neural network BiLSTM_CNN. Therefore, *softmax* function is used to classify the emotion of Danmaku texts. The classification accuracy reaches 84.6%, which is a significant improvement compared with the traditional model.

【Key words】 Danmaku; emotional analysis; text classification; pretrained language models; CNN

0 引 言

研究表明，弹幕（Danmaku/Bullet Screen/Overlaid Comments）起源于日本，是当今新型的一种即时评论方式。用户在观看视频的过程中，通过弹幕来表达自己的情绪感受，深受年轻用户的喜欢。相对于普通的评论文本，弹幕具有更强的情感倾向，由于弹幕比较短，所以情绪较为单一，蕴含了有待挖掘的情绪价值。弹幕经常会出现同词不同义，或者一词多义的情况。例如“东西”这个词，在句子“这个东西很好看”中表示一种喜爱的情绪，但在“就这篇作文，你写的什么东西!”这句话中，却表达了一种极度反感的情绪。这给情感判别带来了极大的挑战，也是传统基于情感词典的弊端。为了解决弹幕情绪分类问题，诸多研究人员从不同角度进行分析，提出相应的解决方案。

从弹幕文本分析的角度来看，文献[1]中提出了搭建 Albert-CRNN 模型，对弹幕进行积极情感与消极情感的二分类。该模型采集了上下文的语义信息，并且在哔哩哔哩弹幕视频网、爱奇艺视频和腾讯视频等平台的弹幕数据集上进行对比实验，实验数据约为 1 万条左右。通过对比实验，验证了该模型在弹幕的情感二分类任务中有较高的分类精度效果。该模型最大的特点，是通过对文本进行卷积操作，将得到的语义特征向量传入到 RNN 模型中，使其能够充分利用上下文语义信息。文献[2]对弹幕的口语化问题，构建了弹幕情感词典。根据对比词典中的词汇计算其情感值，并融合词汇的情感强度，综合计算出一条弹幕的情感值。文献[3]提出了一种融合了注意力机制的 LSTM 情感二分类模型，通过利用 Word2Vec 构建词向量传入 LSTM 中来融合上下文信息。文献[4]使用双通道卷积神经网络对

基金项目：中央高校基本科研业务费青年教师创新项目(31920210083)；2021 年甘肃省自然基金(21JR1RA199)。

作者简介：江 涛（1983-），男，博士，副教授，CCF 会员，主要研究方向：自然语言处理、舆情分析、情感计算等；黄昌昊（1996-），男，硕士研究生，CCF 会员，主要研究方向：自然语言处理、大数据；孙 斌（1999-），男，硕士研究生，CCF 会员，主要研究方向：自然语言处理、大数据。

通讯作者：江 涛 Email：190205978@qq.com

收稿日期：2022-01-08

弹幕进行文本分类。2 个通道中，一个为字向量的通道，另一个为词向量通道。这里，字向量通道主要用来辅助词向量的通道，词向量通道能够捕捉上下文的语义信息。由于增加字向量通道后能够将字和词的语义特征同时融合，为单一的卷积做了充分的补充。文献[5]中主要使用集成学习的方法，将弹幕分为带情感词和未带情感词。其中，带情感词汇的弹幕，在使用 BS-CAL 算法的基础上，再集成朴素贝叶斯(NB)算法计算其中的情感值；未含情感词汇的弹幕使用 ATT-GRU 结构进行预测。该方法融合了 3 个模型的优点。总的说来，基于 Word2Vec 的 ATT-GRU 分类模型，充分利用了词之间的语义和位置关系；BS-CAL 方法擅长处理包含情感词的弹幕，在对包涵很强情感极性的弹幕进行文本分类上，具有较高的性能；基于情感词典的朴素贝叶斯方法，充分考虑了不同情感词组合带来的隐含影响。文献[6]提出了基于奇异值分解 SVD 算法的卷积神经网络模型，该方法替代传统 CNN 模型中的池化层进行特征提取和降维。文献[7]在弹幕情感分类和情绪分类方面，利用多头注意力机制的卷积神经网络框架进行情绪分类，但实验精度仅在 60%左右。

从短文本分类的角度，文献[8-13]采用 CNN、多头注意力机制、语义抽取等相关的深度学习技术，融合了上下文信息进行短文本分类。文献[14]采用融合注意力机制的 BiLSTM，对短文本中的关键词权重进行优化，使准确率得到了提升。文献[15]采用 LDA 模型在新闻文本中进行主题抽取，拓展了单一的文本表示方法，取得了良好的结果。文献[16]提出了一种融合卷积神经网络和多头自注意力机制(CNN-MHA)的模型，使用多头注意力机制降低了文本噪声，该模型在搜狐新闻数据集上有明显的提升。

从弹幕情绪应用的角度，文献[17]通过构建隐含狄利克雷分布(LDA)的弹幕词语分类模型，判断出弹幕中的词语在视频片段中的多种情绪，再根据视频片段之间的情感依赖关系推荐视频的情感片段。文献[18]通过改进的协同过滤算法对视频进行推荐，将弹幕的情感作为一种视频的特征，再计算视频间的相似度，选择相似度高的进行推荐。

由以上关于弹幕分类的研究可见，基本都只是情感的二分类，并未对弹幕的情绪进行分析，然而弹幕的情绪具有丰富的研究价值。本文意在将弹幕文本按照大连理工大学情感词汇本体库进行情绪分类，按照乐、好、怒、哀、惧、恶、惊 7 个情绪标签进行数据标注后，输入到模型进行训练。Bert 预训练语言模型能充分学习上下文语义相关特征，提高在文本多分类的精度。

1　前期准备

1.1　情绪价值研究

视频的弹幕反映了观看者当下的瞬时观看体验，能够反映出观看者当下的情绪反映或是视频的高光时刻。图 1 是国外一个乐队的演唱会视频，弹幕大部分的情绪是好看和好听的弹幕，能给人一种愉悦和欢乐的感觉，此时若结合弹幕的数量、再依据弹幕的情绪就可以具体定位到歌曲的气氛高潮时刻。弹幕展示见表 1。表中，在视频的 00:32 min 左右，视频中的歌手开始登台演唱，弹幕中几乎都是对歌手的夸赞和赞美，弹幕的情感几乎都是好的情感。

图 1　歌曲桥段

Fig. 1　The movie "Midnight Bell" bridge

表 1　弹幕展示

Tab. 1　Danmaku display

时间	弹幕	情感
00:32	全体起立	好
00:35	听了一遍又一遍	好
00:31	Up 主对我们太好了	好
00:30	终于等到了	好
00:33	嗨起来！	好

再如图 2，这是一个乒乓球赛事视频，由于运动员打出一个不可思议的旋转球，弹幕中大多数情绪偏惊讶，比如："哇！这球太厉害了"，"诡异的，弧圈"，"这球怎么旋转的"，"这不科学"等字眼。弹幕展示见表 2。表 2 中，在视频的 03:08 min 左右，由

于球员打出了一个旋转弧圈球,弹幕中几乎都是对这个球表示惊叹和赞美,弹幕的情感几乎都是惊的情感。

图 2 乒乓球比赛视频片断

Fig. 2 Videos chip of table tennis match

表 2 弹幕展示

Tab. 2 Danmaku display

时间	弹幕	情感
03:05	哇! 这球太厉害了	惊
03:08	这球惊到我了	惊
03:09	诡异的,弧圈	惊
03:09	这球怎么旋转的	惊
03:09	这不科学	惊

由上述实例可见,视频弹幕情绪蕴含丰富的应用价值,可以用作视频的推荐、舆情监控等多方面的应用。

1.2 数据获取

本文数据来源于 BiliBili 弹幕视频网的弹幕文本,先进行页面的分析,通过 request 请求包里 get 方法获取视频的 url,将返回的 request 传入到 BeautifulSoup 方法中,获得视频信息和视频的弹幕。最终导出弹幕格式为 Excel 格式,再通过数据分析的方式,找到数据的接口,为接下来的数据清洗和去重做准备。

1.3 数据清洗

数据清洗是情感分析中的第一个环节,主要是对原始数据进行处理。比如:重复值的处理、文字噪音的处理等。首先通过 python 中的文件读取操作将数据读取进来,按首字母进行排列,删除噪音字符。例如:"&&&","%%%"等。接下来再进行去重处理,最后将预处理过的数据进行重新存储,共计处理了 1.6 万条弹幕文本数据。

1.4 数据标注

本文数据的标注主要依据大连理工大学《情感词汇本体库标准》[19] 中的 7 大类、16 小类情感,对弹幕进行 7 种情感类别划分,将 1.6 万条弹幕进行情感标注,在人工标注的同时剔除掉不具有情感倾向的弹幕。

2 基于 Bert_BiLSTM_CNN 弹幕情绪分类

2.1 BiLSTM 模型原理

BiLSTM 是一个双通道的 LSTM 结构,模型将一个句子前向和后向的信息相融合,再将 2 个时间序列相反的 LSTM 输出矩阵相结合。对于获取输入的一句话来看,前向 LSTM 可以得到正向的语句信息,后向 LSTM 可以得到反向的语句信息。例如:"长得"、"真有"、"创意"、"活得"、"富有"、"勇气"是 BiLSTM 所获得的前向语义编码信息,同时 Bi-LSTM 还包含反向编码信息,将文本逆序再进行一次编码,将句子倒过来再次编码为:"勇气"、"富有"、"活得"、"创意"、"真有"、"长得",逆序后再生成一次语义向量,最终把这 2 个编码信息合并成一个输出矩阵。LSTM 单元结构见图 3。图 3 中,含有 3 个门结构,分别为遗忘门、输入门、输出门。在 t 时刻,$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_c$ 是权重矩阵,$\boldsymbol{b}_f$、$\boldsymbol{b}_i$、$\boldsymbol{b}_c$、$\boldsymbol{b}_o$ 是偏置矩阵。LSTM 模型单元结构设计可做阐释分述如下。

(1)遗忘门。这里用到的数学公式可写为:

$$\boldsymbol{f}_t = sigmoid\ (\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{H}_{t-1} + \boldsymbol{b}_f) \tag{1}$$

(2)输入门。这里用到的数学公式可写为:

$$\boldsymbol{i}_i = sigmoid\ (\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{H}_{t-1} + \boldsymbol{b}_i) \tag{2}$$

(3)网络内记忆单元。这里用到的数学公式可写为:

$$\boldsymbol{C}_t = \boldsymbol{f}_t \otimes \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \otimes \tanh\ (sigmoid(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{H}_{t-1} + \boldsymbol{b}_c)) \tag{3}$$

(4)输出信息。这里用到的数学公式可写为:

$$\boldsymbol{o}_t = sigmoid\ (\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o H_{t-1} + \boldsymbol{b}_o) \tag{4}$$

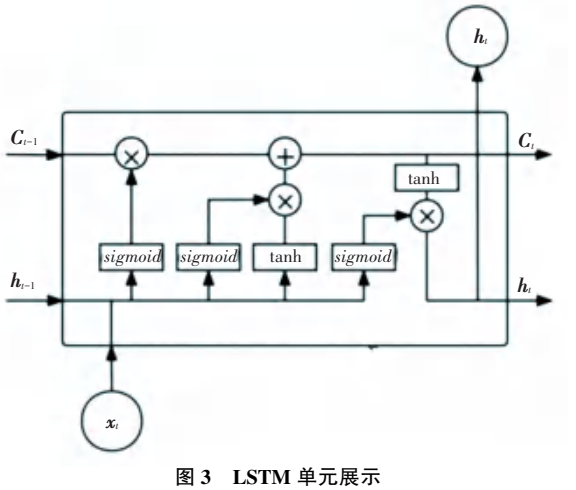

图 3 LSTM 单元展示

Fig. 3 LSTM unit display

2.2 基于 Bert_BiLSTM_CNN 的弹幕文本情绪分析模型

Bert_BiLSTM_CNN 模型由 Bert 层、Bi-LSTM

层、卷积和池化层、全连接层和 *softmax* 层组成。整体模型流程如下：

(1)Bert 层。将弹幕文本输入到模型的 Bert 层中，文本分为字级别，输入形式为 $X_1, X_2, \cdots, X_n$。其中，X_i 表示该条弹幕文本中的第 i 个字。获取文本的位置编码和每个字的字编码，得到每个字的最终编码向量：

$$X_{final\_embedding} = Token\_Embedding + Positional\_Embedding \quad (5)$$

将 Bert 预训练好的 embedding 向量输入到 Bi-LSTM 中。

Bert 采用双向 Transform 组合，能更好地学习语义特征，最终得到动态特征向量，很好地解决了同词不同义的情况。

(2)Bi-LSTM 层。Bi-LSTM 中输入的每个字向量数据维度为 $sequence\_length \times hidden\_size$。其中，$sequence-length$ 是指句子长度，$hidden-size$ 是指每个字的向量维度，传入前向 LSTM 和后向 LSTM 中训练。经过 Bi-LSTM 训练后，将每个字的 2 个文本向量进行叠加，就得到了每个字的向量 $\boldsymbol{h}_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$。其中，$\overrightarrow{h_t}$ 为前向语义特征向量，$\overleftarrow{h_t}$ 代表后向语义特征向量。

(3)卷积和池化层。使用不同尺寸的卷积核进行多个卷积操作。本模型采用尺寸为 2、3、4 的卷积核分别展开动态循环，遍历每一个句子向量进行卷积操作，最大能力获取句子的语义信息，经池化层下采样降维得到该文本向量。池化层主要使用 *maxpooling* 函数，为了获取每一个字向量的最大值，研究将一个句子向量维度压缩到[$sequence\_length \times 1$]维度，下采样主要是为了特征降维、减少参数、压缩数据。

(4)全连接层和 *softmax* 层。将上述所得的向量进行全连接加权计算，最后利用 *softmax* 函数进行归一化，即可得到文本情绪分类的概率分布向量，取最大概率的类别，将弹幕分为好、乐、怒、悲、惧、恶、惊。Bert_LSTM_CNN 模型流程如图 4 所示。图 4 中，X_1，X_2，X_3，…，X_n 代表句子的每个字。

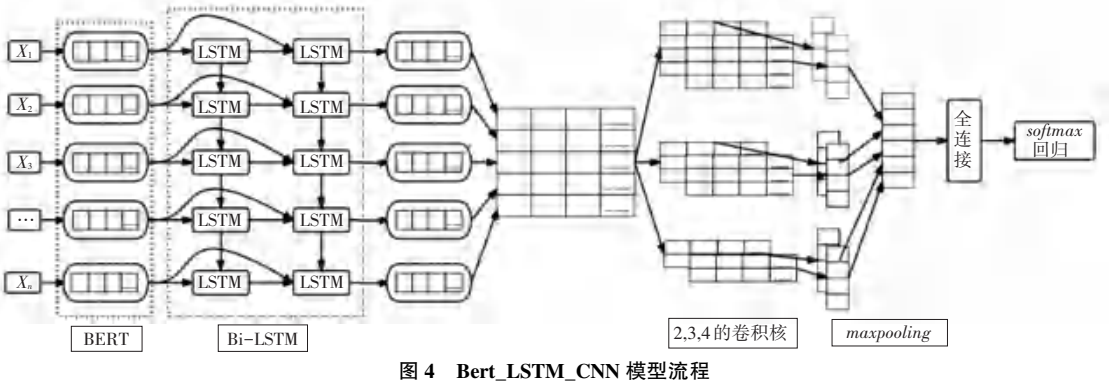

图 4　Bert_LSTM_CNN 模型流程

Fig. 4　Bert_LSTM_CNN model flow demonstration

3　实验结果分析

3.1　实验数据集

本实验所用数据是从 B 站（https://www.bilibili.com/）根据不同主题的视频随机爬取的弹幕。对数据预处理后，最终将 1.6 万条弹幕文本按照 7∶1.5∶1.5 的比例，划分为训练集、测试集、验证集。数据分布见表 3。

表 3　文本情绪分布表

Tab. 3　Text sentiment distribution table　　条

好	乐	怒	悲	惧	恶	惊
4 973	1 843	1 421	2 379	1 243	2 674	1 539

3.2　实验环境及参数设置

本文实验的开发环境为 Pytorch，开发语言为 Python；使用 RTX2080Ti 显卡运行，运行内存为 12 GB。实验所用模型参数见表 4。其中包括：Word2Vec 词向量维度、Bert 词向量维度、BiLSTM 隐藏状态向量维度、CNN 向量维度、滤波器窗口大小、优化函数、损失函数、迭代次数、*dropout* 等。

表 4　模型的参数设置

Tab. 4　Parameters setting of the model

参数名称	参数取值
Word2Vec_dim	300
BERT_dim	768
Word2Vec_pad_size	42
BERT_pad_size	61
Hidden_size	128
Num_filters	128
Filters_sizes	2,3,4
Optimizer	*Adam*
Loss	*Cross_entropy*
Epoch	15
Dropout	0.5

通过为不同的参数设置适应性的学习率,不断地迭代更新参数,提高模型准确率。

3.3 实验结果及分析

为了验证 Bert_BiLSTM_CNN 模型的性能,将该模型与 Bert-CNN、Bert-RNN、朴素贝叶斯(NB)进行对比实验。实验采用准确率、召回率、F_1 值作为评价指标,给出了不同模型在弹幕文本数据集上的准确率对比。实验结果见表 5。

表 5 对比实验效果

Tab. 5 Experimental results %

	精准率(*acc*)	召回率(*recall*)	F_1 值
Bert-CNN	82.6	82.3	81.9
Bert-LSTM	81.6	82.6	82.5
CNN	79.6	80.1	79.8
NB	76.4	78.6	77.6
Bert	83.6	82.9	83.1
Bert_BiLSTM_CNN	**84.6**	85.5	85.2

由表 5 分析可见,相比朴素贝叶斯(NB)、Bert-CNN、Bert-LSTM模型, Bert_BiLSTM_CNN 模型在弹幕文本情绪分析中具有更佳的效果,精准度达到 84.6%。

结合实验效果可以得出,与 Word2Vec 构建的词向量模型相比,使用 Bert 预训练的语言模型有着很明显的优势。由于 Bert 语言模型能够获取的文本特征,可以充分利用语句里的上下文语义信息,从而使文本情绪分析效果得到极大的提升。另外, Bert_BiLSTM_CNN 模型相较于单一的 Bert 模型在文本的情感分析的实际应用中具有更优秀的表现,同时也证明了采用 BiLSTM_CNN 结构,通过双向 LSTM 所得到的向量,再通过卷积操作能更好地融合上下文的语义。

单类别的实验结果见表 6。表 6 中列出了每类数据的精确度、召回率、F_1 值。

表 6 各情绪类别精度

Tab. 6 Accuracy of each sentiment category %

情绪类别	精确度(*Precision*)	召回率(*recall*)	F_1 值
好	89.38	89.38	89.38
乐	93.75	90.91	92.31
悲	96.92	75.00	84.56
怒	71.88	71.88	71.88
惊	91.43	76.19	83.12
恶	64.21	81.33	71.76
惧	79.59	95.12	86.67

4 结束语

本文通过 Bert_BiLSTM_CNN 模型对弹幕文本的情绪进行判别,得到该弹幕的情绪分类。研究了弹幕中蕴含的大量情绪价值,可以应用在很多领域。比如:视频精彩时刻的定位、视频片段的推荐等等。验证了该模型的有效性强于普通的 Word2Vec 和单一的 Bert 或 CNN 模型;使用 BiLSTM 和 CNN 的结合对特征进行训练,相较于单一的 Bert 模型,双向的 LSTM 和卷积操作,能吸收文本的前后信息,在 B 站的弹幕文本里呈现出较好的分类效果。在下一步的研究中,将继续探讨如何将弹幕的情绪用于视频推荐、精彩片段定位以及视频创作等领域。

参考文献

[1] 曾诚,温超东,孙瑜敏,等. 基于 ALBERT-CRNN 的弹幕文本情感分析[J]. 郑州大学学报(理学版),2021, 53(03):1-8.

[2] 洪庆,王思尧,赵钦佩,等. 基于弹幕情感分析和聚类算法的视频用户群体分类[J]. 计算机工程与科学, 2018,40(06):1125-1139.

[3] 庄须强,刘方爱. 基于 AT-LSTM 的弹幕评论情感分析[J]. 数字技术与应用, 2018, 36(02):210-212.

[4] 李平,戴月明,吴定会. 双通道卷积神经网络在文本情感分析中的应用[J]. 计算机应用, 2018, 38(06):1542-1546.

[5] YU Lei, WU Yu, YANG Jie, et al. Bullet subtitle sentiment classification based on affective computing and ensemble learning [J]. Wireless Communications and Mobile Computing, 2021(2): 1-9.

[6] 邱宁佳,丛琳,周思丞,等. 结合改进主动学习的 SVD-CNN 弹幕文本分类算法[J]. 计算机应用, 2019, 39(03):644-650.

[7] 赵庶旭,刘李姣,马秦靖. 基于自注意力机制的弹幕文本情绪分类模型(英文)[J]. Journal of Measurement Science and Instrumentation,2021,12(04):479-488.

[8] XU Jingyun, CAI Yi. Incorporating context-relevant knowledge into Convolutional Neural Networks for short text classification [C]// Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA: AAAI, 2019: 10067-10068.

[9] YU Shujuan, LIU Danlei, ZHU Wenfeng, et al. Attention-based LSTM, GRU and CNN for short text classification[J]. Journal of Intelligent & Fuzzy Systems, 2020, 39(1): 333-340.

[10] WANG Haitao, HE Jie, ZHANG Xiaohong, et al. A short text classification Method Based on N-Gram and CNN[J]. Chinese Journal of Electronics, 2020, 29(2): 248-254.

[11] HAO Ming, XU Bo, LIANG Jingyi, et al. Chinese short text classification with mutual-attention Convolutional Neural Networks [J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2020,19(5):61:1-61:13.

[12] ZHANG Tianyu, YOU Fucheng. Research on short text classification based on TextCNN [J]. Journal of Physics: Conference Series, 2021, 1757(1): 012092.

[13] ZHOU Yajian, DENG Dingpeng, CHI Junhui. A short text classification algorithm based on semantic extension[J]. Chinese Journal of Electronics, 2021, 30(1): 153-159.

(下转第 69 页)