设计研究与应用

基于深度学习的 COVID-19 疫情期间网民情绪分析

刘洪浩

（河南大学 国际教育学院，河南 开封 475000）

摘 要： 微博文本情绪分析技术在舆情监控等领域具有广泛应用。基于传统机器学习模型和情感词典进行情感分析的结果往往不够理想，如何提升性能成为该领域的一个主要挑战。本文中我们使用了基于深度学习的 BERT 以完成语言理解任务并与传统做法性能相比较，结果中 BERT 模型取得了更好的性能。之后我们利用该模型进行三分类以分析 COVID-19 疫情期间的微博评论，总体上正面与中立情绪占主导。此外，我们也针对词频和词云进行相关分析，以期实现全方面了解此次疫情期间社会情感状态的目的。

关键词： 深度学习，词嵌入，BERT 模型，情感分析，微博爬虫，文本处理

中图分类号： TP183 **文献标识码：** A **DOI：** 10.3969/j.issn.1003-6970.2020.09.048

本文著录格式： 刘洪浩. 基于深度学习的 COVID-19 疫情期间网民情绪分析[J]. 软件，2020，41（09）：185–188

Sentiment Analysis of Netizens During the COVID-19 Epidemic Based on Deep Learning

LIU Hong-hao

(*College of International Education, Henan University, Kaifeng* 475000, *China*)

【Abstract】: Sentiment analysis of microblog text is widely used in public opinion monitoring and other fields. The results of sentiment analysis based on traditional machine learning models and sentiment dictionaries are often not ideal. How to improve performance has become a major challenge in this field. In this thesis, we use BERT based on deep learning to complete the language understanding task. Compared with traditional methods, BERT model has achieved better performance. We use the model to analyze microblog comments during the COVID-19 epidemic by conducting a three-category classification and find that positive and neutral emotions are dominant. We also conduct further analysis on word frequency and word cloud to gain more insights into the emotional states during the epidemic.

【Key words】: Deep learning; Word embedding; BERT; Sentiment analysis; Microblog crawler; Text processing

0 引言

文本是用于情感分析的典型数据集。由于情感文本数据的迅速增长和极高应用价值，使得自动识别和分析人们在文本中表达的情感成为一种必要。社交网络文本情感分析被广泛应用于在金融[1]、市场[2]、社会[3]、娱乐[4]等诸多领域之中，关于文本情感识别算法相关的理论研究[5-7]也越发丰富。越来越多基于社交网络的情感分析实践和研究的出现表明其实用性与科学性。微博短文本已成为国内数据的情感表达和舆论走向的代表，它为研究社会发展和人类行为特征提供更多可能性。

新冠肺炎疫情备受社会各界关注。2020 年 1 月 1 日至 2 月 20 日，疫情相关微博话题数超过 200 个。此次疫情为高热度的重大社会热点事件，对疫情期间的情感识别和可视化分析能客观反映出疫情舆情的发展动向，有助于有关机构制定合理科学的决策，具有较高研究价值。

文本分类的精度取决于提取语义特征的方法和分类器的种类。本文关注基于深度学习的中文文本词嵌入方法与传统做法的比较和疫情期间情感分析。我们研究了基于深度学习中词向量技术的情感识别方法，利用 BERT 模型和 Embedding 层预训练方法，分别进行研究，实验对比中 BERT 预训练模型取得更加准确的结果。我们将利用 BERT 模型的分类结果对此次疫情全面分析，并给出疫情期间微博文本的词云表示，以提高情感分析的准确度，达到全面了解此次疫情期间社会舆情的目标。

1 相关工作

本节简要介绍微博数据情感分析的相关研究，以及获得词嵌入的方法。

1.1 微博数据情感分析

现有文献中已有较为丰富的针对微博文本的情感分析策略。王培名等人[8]设计了自适应的并发采集算法优化模拟登录和代理池的构造访客 Cookie 功能，高效获取微博数据，为微博数据采集策略提供了多样性。刘楠[9]针对微博短文本形式的情感分析，归纳新的细粒度情感分析流程，提出 TF 和 TF-IDF 归一化权重计算方法，与传统提取特征的方法相比，能够更准确判断出多种类情感的权重，实现了该方法有效性的评估。

1.2 词嵌入

词嵌入是一种词的数字向量化表示，相似含义的词可用类似的向量表达。词嵌入的研究关键在于获得密集低维的分布式特征向量表示词的不同特征，每一个词与分布式向量相关联，每个词与向量空间中的点相关联，促进与神经网络词的更好拟合和学习更新[10]。

作者简介： 刘洪浩(1998–)，男，本科在读，研究方向:数据挖掘，机器学习。

2013 年 Google 公司的 Mikolov 等人[11]开发出了基于神经网络训练词向量新的模型体系结构 Word2Vec，Word2Vec 核心思想是通过词的上下文窗口得到词的向量化表示得到分布式的词嵌入，其本质是降维操作，将 One-Hot 编码形式的词向量转化为 Word2Vec 形式，Word2vec 包括 CBOW 与 Skip-Gram 两种模型。Pennington 等人[12]在 2014 年提出了继 Word2Vec 以后又一具有较大影响力的词向量训练方法 Glove。Glove 是一种无监督的词嵌入模型，采用共现矩阵并对其降维，将局部信息和整体信息结合，解决了 Word2Vec 的只考虑词与局部窗口信息和忽略了语料库的统计信息的问题。

随着词嵌入模型不断深入研究，词嵌入模型更新速度越来越快，从传统机器学习词袋模型[13]等，发展到如今基于深度学习的预训练方法诸如 Word2Vec[14]、Glove[15]、BERT[16]的词嵌入算法，如今的词嵌入方法通过神经网络模型利用更长的上下文来解决自然语言问题[17]。

2　研究方法

2.1　数据获取

我们采用已标注的 10 万余条微博文本语料库。在数据收集阶段，我们使用微博 API[18]收集微博文本数据，具体包含 create_at（微博发布时间），id（发布用户 id），text（微博文本）属性信息。我们一共收集到 2020 年 1 月 1 日至 2 月 20 日疫情期间的 1 万余条微博文本作为待分析的文本。图 1 展示了研究方法的总体流程。

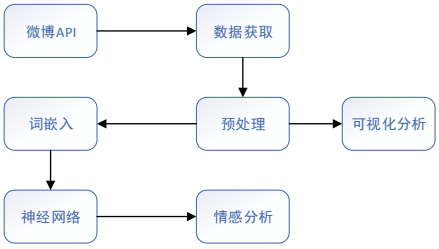

图 1　实验方法
Fig.1　Methodology

2.2　预处理

我们原始的数据集需要经过预处理，过滤掉一些不符合要求的文本以便更契合情感分析任务。我们将对微博数据集中的中英文表情和符号表情进行过滤，然后利用 jieba 库进行分词，利用停用词表进行深度清洗，筛选出停用词，其中包括数字，中英文标点符号，语气词，无实意词等。图 2 显示预处理的步骤。

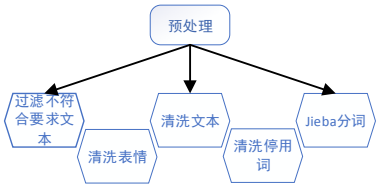

图 2　预处理过程
Fig.2　Pre-processing progress

2.3　模型

我们将使用训练神经网络时拟合词嵌入层方法。深度神经网络工具 Keras，是一个深度学习框架，Keras 的 Embedding 层和 Word2Vec、Glove 本质上是一样的，将词嵌入在浅层神经网络中用密集向量表示同时在更小维度中集合信息，但其特点是可以和神经网络一起训练形成一个端到端的结构，以便高效拟合相应模型任务。Keras 的 Embedding 层输入数据要求为整数编码，我们利用该库中的分词器 Tokenizer API 生成序列化向量作为嵌入层的输入，Embedding 层被定义为神经网络的第一个隐藏层。该层使用方式灵活，我们利用该隐藏层作为深度学习模型的一部分共同进行模型学习和训练，以将整数映射到 Embedding 层向量空间中的获得带有权重的密集向量。

同时我们还将使用最新的 BERT 预训练模型进行研究，BERT 模型由 Devlin 等人[19]在 2018 年提出，BERT 模型被评为目前自然语言处理效果最佳的深度预训练模型。BERT 预训练模型较之于以往预训练模型最大优势在于 BERT 运用了双向转换器如图 3 所示。BERT 的高效能同时体现在其特殊的预训练方法，包含有 Masked Language Model 和 Next Sentence Prediction。BERT 模型的输入由词向量，段向量，位置向量三部分组成，如图 4 所示。在词向量里面有两个特殊标志 CLS，SEP。CLS 作为第一个向量来得到句子向量，SEP 用来区分句子。为了训练深度双向表示模型，需要 BERT 中的 Masked Language Model 和 Next Sentence Prediction。

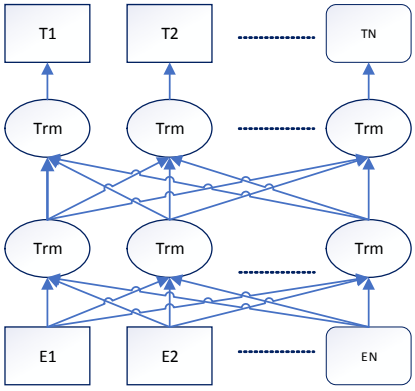

图 3　BERT 模型结构示意图
Fig.3　Schematic diagram of BERT model structure

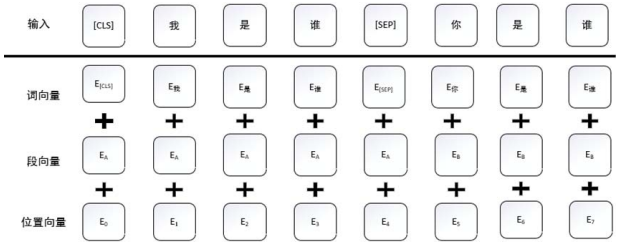

图 4　BERT 输入端
Fig.4　Input of BERT

我们将利用 Embedding 层和 BERT 进行文本预训练进行情感分类效果比较。神经网络模型将用到深度

学习框架 Keras 提供了网络层线性堆叠的 Sequential 顺序模型来搭载 Relu 全连接层和 Softmax 激活函数层。

2.4 可视化分析

我们针对处理后的微博文本数据借助第三方 wordcloud 库和 matplotlib 库对数据进行可视化分析。通过统计出高频词汇、评论量和平均情感数值的时间变化，我们从数字角度定量考察疫情期间舆情发展的情况，以便更加直观了解此次疫情对民众的影响。

3 实验和结果

在这一节中，我们将详细介绍我们在本次实验中利用上述方法完成的具体实验工作和结果。我们将训练集和测试集经过预处理后，如图 2 所示。产生了符合情感分析要求的 10 万条语料集。

3.1 情感分析结果

我们分别利用 Embedding 层和 BERT 模型获得词嵌入，再拟合神经网络构建情感识别模型，我们将前述的 10 万条语料集划分为训练集，验证集，测试集进行训练，训练轮次均为 5 次。在测试集上进行评估，结果对比如表 1、2 所示。

表 1 BERT 训练结果
Tab.1 Results of BERT training

训练次数	F1	准确度	召回度
1	0.7523	0.7475	0.7617
2	0.7557	0.7493	0.7624
3	0.7645	0.7519	0.7741
4	0.7706	0.7675	0.7862
5	0.7832	0.7783	0.7958

表 2 Embedding 层训练结果
Tab.2 Results of Embedding training

训练次数	F1	准确度	召回度
1	0.6519	0.6425	0.6787
2	0.6651	0.6503	0.6820
3	0.6673	0.6561	0.6838
4	0.6728	0.6617	0.6901
5	0.6745	0.6654	0.6919

3.2 2019n-Cov 疫情数据可视化结果

3.2.1 情感分类微博数据分布

情感分析结果为典型的三分类，1 代表积极，0 代表中性，–1 代表消极。我们首先从宏观角度获得了评论在三种情感中所占数量，如图 5 所示。总体分析可得，积极情绪文本比消极情绪文本在数量上较多，表明此次疫情期间网民整体呈现积极心态；中性情绪所占数量最大也代表了多数网民对此次疫情的不信谣不传谣态度，这也证明了相关机构实施的大众居家隔离等防疫措施有效性。

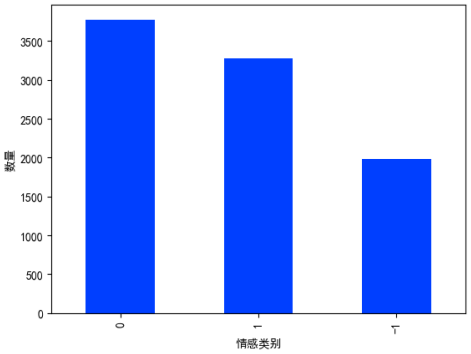

图 5 情感分类结果
Fig.5 Results of Emotion Classification

3.2.2 微博情感时间趋势

我们从动态角度深入研究网民情绪变化，如图 6 所示。消极情绪在 1 月 20 日左右呈现迅速增长趋势，这可能是由于在 1 月 20 日钟南山院士肯定新冠肺炎存在人传人现象，加大网民的重视和恐惧程度。从 2 月 8 日至 10 日消极文本数量达到峰值显示出李文亮医生的逝世可能加重了负面情绪。但国家紧急实施居家隔离和调配全国力量支援湖北等多项措施，这使得 2 月 10 日以后积极情绪占比增大，网民情绪逐渐好转。

3.2.3 平均情感数值时间变换

我们对预测结果进行情感数值平均化，得到平均情绪值随时间的变化趋势，如图 7 所示。网民情绪在

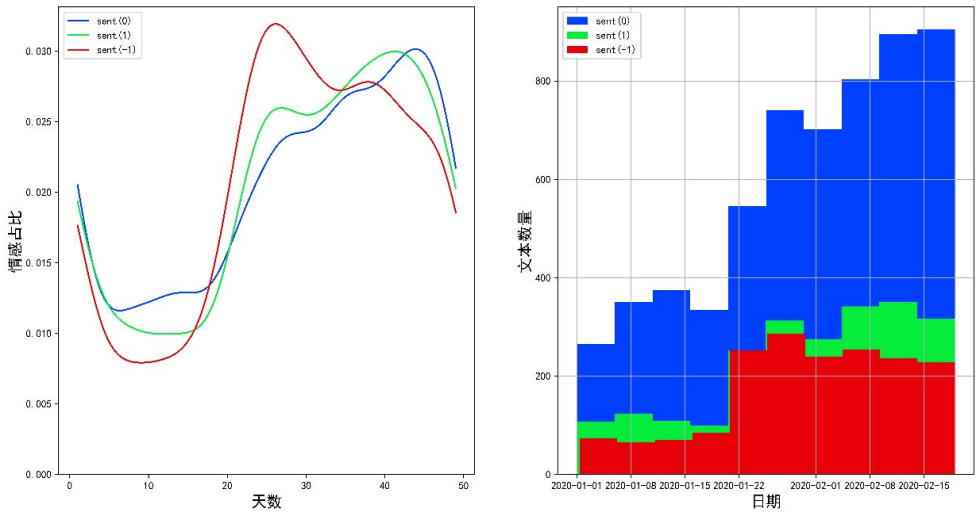

图 6 情感趋势
Fig.6 Trend of Emotion

1 月 20 号左右进入低沉期,验证了图 6 数据所示结论，之后情绪波动起伏，并在 2 月 9 日左右进入网民情绪低谷。之后每日新冠疫情感染人数逐渐下降，网民情绪逐渐好转。从整体情绪幅度观察，情绪波动较大，这说明了疫情期间不同地区的感染人数和死亡人数对网民的情绪造成不同程度的影响；在 2 月 10 日以后情绪逐渐虽有起伏整体仍保持积极心态，平均情绪数值逐渐有上升趋势。

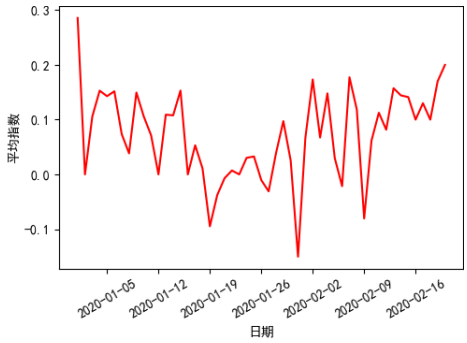

图 7　情感平均值
Fig.7　Average value of emotion

3.2.4　词频统计

关注网民疫情期间的热点话题也是全面了解网民情绪和态度的重要途径之一。我们利用测试集根据词频得到高低排序，得到前 800 词语的词云。如图 8 所示。由词云词频统计可知此次网民热点讨论为“疫情”、“武汉”、“肺炎”、“冠状病毒”，“新型”等，这也表明在 COVID-19 期间，网友对此次新冠肺炎的重视和关注，同时也代表了冠状病毒为此次新型肺炎的起源并对社会造成严重影响，也导致“口罩”、“医院”等资源的缺乏和讨论。此外我们可以从“武汉”、“加油”、“医院”这些高频率词汇中体会到网民对武汉的关心，以及对所有为新冠肺炎抗争的医护天使的感谢。

图 8　词云
Fig.8　Wordcloud

4　结论

本文基于深度学习模型 BERT 比较了其在词嵌入训练的优越性，并获得的较准确的情感分析结果，研究意义总结为如下三个方面：（1）比较了 BERT 模型和 Embedding 层的情感分类效果，解决了预训练模型中单向信息流问题，并大大减少神经网络的复杂度；（2）利用疫情之前的微博数据作为 BERT 模型训练集对 COVID-19 期间测试集进行情感分类；（3）数据化地呈现出此次疫情网民的情绪变化和走向，结果表明网民在 COVID-19 疫情期间整体情绪积极。由于词向量结合神经网络的端到端学习模型具有多样性，后续可比较 Word2Vec，Glove 等热门模型，以获得更高准确度的词嵌入模型，同时利用 Keras 库构建其他形式的神经网络以加强性能。

参考文献

[1] 蒋钰慧. 投资者情绪对我国股票市场收益率的影响研究[D]. 上海外国语大学, 2019.
[2] 张爱华, 陈超雨. 基于文本分析的中国5G产业发展研究——市场主体视角[J]. 北京邮电大学学报(社会科学版), 2019, 21(06): 90-102.
[3] 刘雯, 高峰, 洪凌子. 基于情感分析的灾害网络舆情研究——以雅安地震为例[J]. 图书情报工作, 2013, 57(20): 104-110.
[4] 柳池煜. 票房预测中的社交网络评论情感挖掘技术研究[D]. 南京邮电大学, 2019.
[5] 梁军. 基于深度学习的文本特征表示及分类应用[D]. 郑州大学, 2016.
[6] 陈文. 中文短文本跨领域情感分类算法研究[D]. 重庆大学, 2016.
[7] 彭浩, 朱望鹏, 赵丹丹, 等. 面向多源社交网络舆情的情感分析算法研究[J]. 信息技术, 2019(02): 43-48.
[8] 王培名, 陈兴蜀, 王海舟, 王文贤. 多策略融合的微博数据获取技术研究[J]. 山东大学学报(理学版), 2019, 54(05): 28-36+43.
[9] 刘楠. 面向微博短文本的情感分析研究[D]. 武汉大学, 2013.
[10] BENGIO Y, DUCHA R ME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.
[11] MIKOLOV T, CHEN Kai, COR RADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013, 2(12): 27-35.
[12] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]. Proceedings of the 2014 conference on empirical methods in natural language processing(EMNLP). 2014: 1532-1543.
[13] 黄春梅, 王松磊. 基于词袋模型和TF-IDF的短文本分类研究[J]. 软件工程, 2020, 23(03): 1-3.
[14] 彭晓彬. 基于word2vec的文本情感分析方法研究[J]. 网络安全技术与应用, 2016(07): 58-59.
[15] 陈珍锐, 丁治明. 基于Glove模型的词向量改进方法[J]. 计算机系统应用, 2019, 28(01): 194-199.
[16] 胡春涛, 秦锦康, 陈静梅, 等. 基于BERT模型的舆情分类应用研究[J]. 网络安全技术与应用, 2019(11): 41-44.
[17] 刘胜杰, 许亮. 基于词嵌入技术的文本表示研究现状综述[J]. 现代计算机, 2020(01): 40-43.
[18] 王铁刚. 社交媒体数据的获取分析[J]. 软件, 2015, 36(02): 86-91.
[19] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 1810. 04805, 2018.