基于深度学习的话题流行度预测

徐华韫 龚泽阳 何正杰 朱 政 尚锦浩

（国际关系学院信息科技学院，北京 100091）

摘 要：新浪微博是当下社会中使用最为广泛的网络社交平台，其中的话题模块涵盖了当下社会的关注点。本文通过使用网络爬虫获得的微博话题数据，结合话题主持人、类别及标题内容等进行特征提取，构建两层分类器，使用 FastText 和 GBDT 的架构，对微博话题能否成为热门话题进行预测，预测在实验中可以达到 85.27% 的准确率。

关键词：数据分析；神经网络；自然语言处理

中图分类号：TP393.092 **文献标识码：**A **文章编号：**1003-9767（2018）21-057-03

Deep Learning Based Topic Popularity Prediction

Xu Huayun,Gong Zeyang,He Zhengjie,Zhu Zheng,Shang Jinhao

(School of Information Science and Technology, University of International Relations,Beijing 100091,China)

Abstract: Sina Weibo is the most widely used online social platform in the current society. The topic module covers the current social focus. This article uses the microblogging topic data obtained by the Internet crawler, combines the topic host, category, and title content to extract features, builds a two-layer classifier, and uses the structures of FastText and GBDT to predict whether the microblogging topic can become a hot topic. The accuracy of the prediction can reach 85.27 % in the experiment.

Key words: data analysis; neural network; natural language processing

随着互联网的不断发展，微博热点话题的研究成为了研究社会舆情方向的重要手段。微博热点话题的种类有直接性等特点，其影响巨大。通过对微博热点话题的预测可以了解网民兴趣点的变化态势。

微博热点话题的预测主要采用传统的机器学习方法，如 SVM[1]、K 近邻[2]、贝叶斯网络[3]等，也有采用数据挖掘技术[4]解决此问题的尝试。以上的尝试大都基于话题热度的时间变化进行分析预测，所需周期长。而目前火爆的深度学习方法使用的频率却不高，如 BP 神经网络，而且效果不显著。

为了获得理想的微博热点话题预测结果，笔者提出基于 FastText 和 GBDT 的微博热点话题预测模型，通过该模型对微博话题的某些文本特征进行预测，而不用考虑其随时间演化而形成的特征，具有实时性。结果表明，本文模型提高了微博热点话题的预测精度。

1 数据集获取

测试的数据集来源于新浪微博话题榜单中的话题，包括财经、体育、文娱等板块。采用网络爬虫技术，爬取微博话题，数据时间跨度为 2018 年 2 月至 2018 年 8 月。

1.1 特征爬取

新浪微博提供的热门话题榜单，可以基本涵盖社会中人们所有的关注点。本次实验所获得的数据集均由爬取该榜单获得。为了方便构建特征工程，爬取目标为每个话题的标题、导语、话题主持人、话题类别和标签等数据。与此同时，将爬取话题阅读量作为话题热门度的参考。

1.2 网络爬虫

运用 requests 库等模块编写爬虫程序，从新浪微博 WAP 端侵入获取数据，批量获取话题数据。为了应对可能出现的反爬虫机制，采取设置 headers 以及 ip 代理的伪装方式保证爬虫的顺利进行。

2 特征构建及模型

本文选取 3 个话题属性，即话题主持人认证状况，话题类别，话题标题和导语，作为输入，转换成向量形式嵌入模型中。而使用的模型包含两层分类器，第一层是文本分类器，第二层是话题分类器，最后获得对于该话题热门度的预测。

作者简介：徐华韫（1997-），男，江苏南京人，本科在读。研究方向：数据科学与工程。
龚泽阳（1997-），男，湖北十堰人，本科在读。研究方向：智能信息处理。

2.1　主持人认证状况、类别特征

新浪微博话题主持人主要负责该话题的传播、讨论等方面事项。通常话题主持人影响力与话题的影响力呈正比，而认证情况则是体现话题主持人影响力的一大因素。现有的认证包括微博官方认证、微博个人认证、微博达人和无认证等 4 个类别。而话题的类别同样对于话题热门度有着相当重要的影响。通常而言，文体类别下的话题会更受群众的关注。本次实验的数据共覆盖 214 个类别。

考虑到认证类别和话题类别总数有限，因此分别使用 one-hot 编码形式来表示，将认证情况与话题类别分别转换成向量形式。有时出现一个话题分属多个类别的情况，例如话题“湖南卫视歌手”属于两个类别：“综艺”“内地节目”。此时就将这两个类别的 one-hot 编码相加。

2.2　话题文本特征

提取话题的标题和话题导语组成话题内容特征。使用自然语言处理算法，根据文本对话题未来的流行度进行预测，预测其成为热门话题的可能性，即为话题文本内容这一特征的影响因子，也是文本分类器的输出。借助 Word2Vec 等工具，将文本中的词语向量化，作为词语级别的嵌入，输入到文本分类器中。具体采用的模型如下。

2.2.1　FastText

FastText[5] 是一个构架简单、基于浅层神经网络的高效率分类器，其训练速度相较于 n-gram 等深度学习模型而言，十分优异。FastText 网络结构其模型架构十分类似于 Word2Vec[6] 中的 Cbow 模型，只有三层：输入层、隐藏层、softmax 层，如图 1 所示。

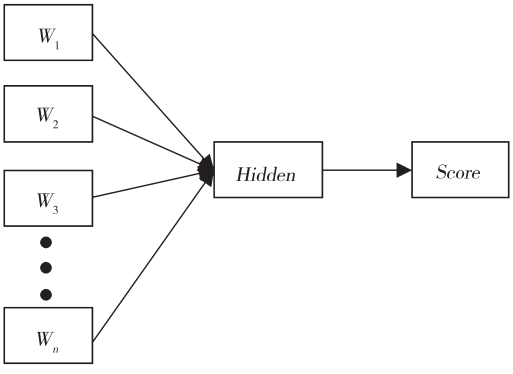

图 1　FastText 架构

在 FastText 的输入端，每一个节点的输入经过词向量矩阵的转换后获得 1 × n 的向量，n 即为设定的词向量维度。在隐藏层经过平均池化后，通过 softmax 层获得各个标签的概率。FastText 独特之处在于，输入端的特征是 n-gram 特征，此外与普通的偏平的 softmax 层不同，FastText 中使用哈夫曼树构建的层次 softmax，对结果进行预测。

2.2.2　CNN 及相关网络

卷积神经网络（CNN）[7] 一般用于图像识别、分类等领域，由于其特有的卷积层和池化层可以对图像中的信息进行归纳、提取，所以 CNN 在目前的分类任务中被广泛运用。在本文中使用的卷积神经网络结构为双层 CNN。对于输入的词向量进行特征提取，通过不断的卷积，最后完成分类任务。此外，为了优化该网络，本文使用了改进模型，包括 TextCnn[8] 和 CNN-BiLSTM 模型，以提升性能。

2.3　热门话题分类器

以主持人认证状况、类别特征、话题文本特征作为输入，判定该话题是否为热门话题的分类器。本文中使用了 BP 神经网络、SVM[9] 和 GBDT[10] 分类器。本段主要介绍 GBDT 分类器。

决策树可以认为是 if-then 规则的集合，易于理解，预测速度快。但是，单独使用决策树算法时，有易过拟合的缺点。GBDT 是一种用于回归、分类和排序的机器学习技术，可将弱学习器提升为强学习器，以构建最终的模型。而且，GBDT 可以通过抑制决策树的复杂性，降低单颗决策树的拟合能力，再通过梯度提升的方法集成多个决策树，从而解决过拟合的问题。另外，GBDT 通过加入正则项等方法能够有效地抵御噪音，具有更好的健壮性，微博文字散漫无序的特点在微博热点话题预测中具有关键作用。

算法可以看作是由 K 棵树组成的加法模型：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (1)$$

其中 F 为所有树组成的函数空间，该模型的参数为 $\theta=\{f_1, f_2, \cdots, f_k\}$。与一般的机器学习算法不同的是，加法模型不是学习多维空间中的权重，而是直接学习决策树集合。

模型的损失函数为：

$$Loss = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

其中 Ω 表示决策树的复杂度，比如树的叶子节点数量、树的深度等。

在笔者提出的模型中，笔者将话题主持人的微博认证情况和话题涉及的类别转化成的向量，W_1 和 W_2 与文本分类器，例如 FastText 输出的评分 Score 进行拼接后，连接形成的向量作为 GBDT 模型的输入，最终输出得到此话题是否可以成为热点话题的预测结果。整体模型如图 2 所示。

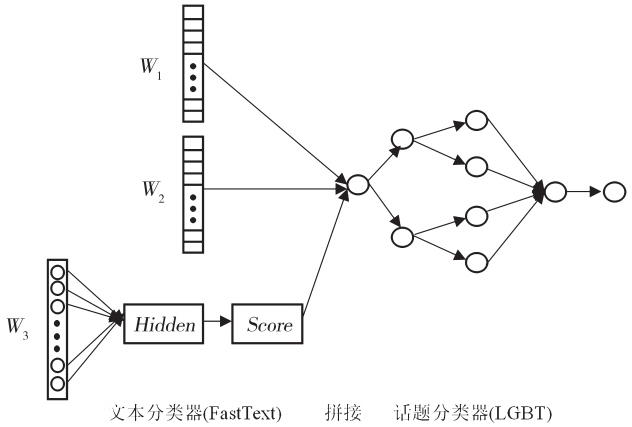

图 2　FastText-GBDT 话题分类器架构图

3 实验

3.1 实验数据集

本文中使用的数据集为使用爬虫获取的新浪微博话题数据集，共 13 457 条。以阅读量 1 亿为分界线，阅读量超过 1 亿的话题被定义为热门话题，反之则为非热门话题。数据集中，共有 3 590 个热门话题，占总数据集的 26.67%。首先对于文本分类器，使用 8 277 个话题训练文本分类器，训练结束后使用剩下 5 180 个话题进行测试，给每一个话题的文本特征给予评分。之后对于 5 180 个话题再进行一次划分，4 148 个话题作为训练集，1 032 个话题作为测试集，训练话题分类器。

3.2 文本分类器

构建文本分类器，将话题标题和话题导语拼接之后，经分词、去停用词等操作后，转换为词向量，作为网络输入。实验中，笔者采用 FastText、CNN、TextCNN 等模型进行测试，结果表明，FastText 在话题文本分类中表现最为出色，而 LSTM 并未对测试结果有产生显著提升。

表 1 不同文本分类器性能比较

文本分类器	准确率
FastText	80.21%
双层 CNN	76.83%
CNN-BiLSTM	76.44%
TextCNN	76.81%

3.3 话题分类器

将 one-hot 表示的认证情况和话题类别，与通过文本分类器得到的文本内容评分，三个部分拼接起来形成一个一维向量，作为话题分类器的输入。在此阶段，主要采用 BP 神经网络、SVM 支持向量机和 LGBT 三种结构分别搭建话题分类器。文本分类器选用之前表现良好的 FastText 和双层 CNN。具体的测试结果如表 2 所示。

4 结 语

对于新浪微博话题的热门度预测，本模型可以做到较准确的判断，并且提供了提升模型准确度的方法。然而话题主持人影响力并不能完全由认证情况反映，且该模型对于之前未出现的陌生节目、人名，并不能及时跟进进行准确预测，所以，该模型还有待改善。

表 2 不同话题分类器和文本分类器组合的性能比较

文本分类器	话题分类器	准确率
FastText	BP 神经网络	75.97%
	SVM	83.27%
	LGBT	85.27%
双层 CNN	BP 神经网络	73.93%
	SVM	80.21%
	LGBT	83.49%

参考文献

[1] 杨俊成 . 基于支持向量机的网络热点话题预测 [J]. 微型电脑应用 ,2017,33(7):30-32,36.

[2] 聂恩伦 , 陈黎 , 王亚强 , 等 . 基于 K 近邻的新话题热度预测算法 [J]. 计算机科学 ,2012,39(6):257-260.

[3] 张一文 , 齐佳音 , 方滨兴 , 等 . 基于贝叶斯网络建模的非常规危机事件网络舆情预警研究 [J]. 图书情报工作 ,2012,56(2):76-80.

[4] 张贵红 , 李中华 . 基于数据挖掘技术的微博热点话题预测 [J]. 现代电子技术 ,2017,40(15):52-55.

[5]A Joulin,E Grave,P Bojanowski,et al. Bag of Tricks for Efficient Text Classification[J].ARXIV,2016.

[6]T Mikolov,K Chen,G Corrado,et al.Efficient Estimation of Word Representations in Vector Space[J].Computer Science,2013.

[7]A Krizhevsky,I Sutskever,GE. Hinton.ImageNet Classification with Deep Convolutional Neural Networks[J].2012,60(2):1097-1105.s

[8]Y Kim.Convolutional Neural Networks for Sentence Classification[J].Eprint Arxiv,2014.

[9]C Corinna,V Vapnik.Support-Vector Networks.Machine Learning,1195,20(3):273-297.

[10]JH Friedman.Greedy Function Approximation: A Gradient Boosting Machine[J].Annals of Statistics,2001,29(5):1189-1232.

（上接第 56 页）

[8] 刘根旺 , 刘永信 , 纪永刚 , 等 . 基于模糊双门限的高频地波雷达与 AIS 目标航迹关联方法 [J]. 系统工程与电子技术 ,2016,38(3):557-562.

[9] 高峰 , 谢小平 , 熊伟 . 基于广义绝对灰关联度的航迹关联算法 [J]. 雷达科学与技术 ,2016,14(6):642-647.

[10] 何友 , 宋强 , 熊伟 . 基于相位相关的目标航迹对准关联技术 [J]. 电子学报 ,2010,38(12):2718-2723.

[11] 何友 , 宋强 , 熊伟 . 基于傅里叶变换的航迹对准关联算法 [J]. 航空学报 ,2010,31(2):356-362.

[12] 宋强 , 熊伟 , 何友 . 基于复数域拓扑描述的航迹对准关联算法 [J]. 宇航学报 ,2011,32(3):560-566.

[13]Tian Wei,Wang Yue,Shan Xiu-ming,et al.Track-to-track Association for Biased Data Based on the Reference Topology Feature[J]. IEEE Signal Processing Letters,2014,21(4):449-453.

[14]Tian Wei,Wang Yue,Shan Xiu-ming,et al.Analytic Performance Prediction of Track-to-Track Association with Biased Data in Multi-sensor Multi-target Tracking Scenarios[J].Sensors,2013,13(9):12244-12265.

[15] 杨哲 , 韩崇昭 , 李晨 , 等 . 基于目标之间拓扑信息的数据关联方法 [J]. 系统仿真学报 ,2008,20(9):2357-2360.

[16] 齐林 , 崔亚奇 , 熊伟 , 等 . 基于距离检测的自动识别系统和对海雷达航迹抗差关联算法 . 电子与信息学报 ,2015,37(8):1855-1861.