目 录

第一章 绪 论

1.1 研究工作的背景与意义

随着信息化进程的不断推进，我国互联网市场空前繁荣。从 WEB2.0 时代开始，到如今的“互联网 +”时代，互联网已经融入了人们的生活中，也对各行各业产生了深远的影响。根据中国互联网络信息中心发布的《第 40 次中国互联网络发展状况统计报告[①]》显示，截止到 2017 年 6 月，我国网民规模达到 7.51 亿，其中手机网民更是达到 7.24 亿，占总体网民的 96.3%，使用率排名前三的互联网应用分别是即时通信（92.1%）、网络新闻（83.1%）、搜索引擎（81.1%），使用率最高的三个 app 应用则是微信（84.3%）、QQ（65.8%）、微博（38.7%）。可以明显看出，人们使用互联网以及获取信息的方式，已经从以前的桌面台式电脑，转变为移动端的手机、IPAD 等掌上设备。

移动设备的盛行，使得人们发布和接收信息都更加方便，据统计，在 2012 年，微博用户已经增长到 3.68 亿，其中 69% 通过移动设备登陆，每天能产生 1.17 亿条微博。手机用户的大量增加，让互联网信息爆炸式增长，并且产生了大量碎片化的信息，如微博、QQ 说说、留言、商品评论等。据有关部门统计，互联网全体文本信息中，80% 以上属于内容较少的短文本信息。

如果能够有效分析这些短文本，对其进行精确的分类，可以方便用户有效的梳理这些浩如烟海的文本并掌握对自身有用的信息，商家也能够根据信息的分类提供更加优质的服务。例如，内容提供商，如新浪微博、知乎等，可以对其提供的信息进行分类，让用户快速获取自己感兴趣的某一类信息，同时商家也可以统计用户浏览过的信息类，精确投放用户感兴趣的广告，减少无关广告对用户体验的伤害；政府相关部门可以根据一段时间内大众发布的微博或朋友圈等信息，掌握当前的热门话题、集中关注点，监控当前的社会舆情；电子商务平台，如淘宝、京东等，可以利用情感分类技术，提取商品的正面评论与负面评论给用户，让用户能够更好的筛选与判断优质商品。

但是和传统文章相比，短文本过于短小（通常在 100 字以内，一般是一句话的长度），不能提供足够的词共现（word co-occurrence）或上下文，以至于很难从中提取出有效的文本特征[1]。因此，常规机器学习技术与文本分类算法很难直接应用在短文本之上。那么如何准确高效的对短文本进行分类，成为了互联网从业者与互联网技术学者所面临的一个重要难题，其突破也会具有重要的商业价值与

① http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201708/P020170807351923262153.pdf

使用价值。

1.2 国内外研究现状

随着互联网中短文本信息的增多，短文本分类领域得到了广泛的关注，越来越多的学者投入到短文本分类的研究之中。但由于短文本短小、信息分散的特性，传统基于词频、词共现的分类方法通常得不到较好的效果，比如贝叶斯方法（Naive Bayes）、最大熵模型（Maximum Entropy Model）、K-邻近算法（K Nearest Neighbors）以及支持向量机（Support Vector Machines，SVMs）。因此学者们开始尝试从其他方面来改进短文本分类算法，比如语义分析（sematic analysis）、半监督（semi-supervised）算法和集中模型（ensemble models）。

纽约大学的 Sarah Zelikovitz 等人[2] 通过隐含语义索引算法（Latent Semantic Index，LSI）对短文本的语义分析，将文本中的词映射到潜在语义空间，来捕获文本单词之间的相关性，提升分类效果。清华大学的 Chen 等人[3] 通过改进的隐含狄利克雷分布（Latent Dirichlet allocation，LDA）模型，将短文本中的单词与多个粒度的话题相关联，从而拓展短文本特征。

Juan Manuel Cabrera 等人[4] 根据分布式词语表示算法（Distributional Term Representations，DTRs），用半监督的方式统计语料中的文档出现特征以及词语共现信息，形成每个词语的上下文信息，最后强化文本表示，以此来克服短文本处理中长度过短、信息高度分散的难点。

中国科学院自动化研究所的冯晓等人[5] 构造了一种集中学习模型，直接确立短文本实例与某一主题直接的相关性，而不是将短文本表示为权重向量，取得了超过基于向量空间模型（Vector Space Model，VSM）的方法的效果。

随着神经网络与深度学习逐渐兴起，并在计算机视觉、语音识别等领域取得了不错的成果，越来越多的学者开始尝试在自然语言处理问题中引入深度学习模型，以此克服之前方法存在的问题。斯坦福大学的 Richard Socher 等人[6] 构造了一个使用矩阵向量的循环神经网络（MV-RNN），来学习长度可变的短语与句子的综合向量表示。牛津大学的 Nal Kalchbrenner 等人[7] 提出了动态卷积神经网络（DCNN），利用动态 K-max 池化的方法，直接获取文本中单词直接的距离关系，避免了对语法分析树的依赖。哈佛大学的 Yoon Kim 等人[8] 将动态词向量与预训练好的静态词向量相结合作为同一段文本的两个表示，输入卷积神经网络的两个通道中进行分类，也取得的较好的效果。

而对于中文文本的分类，国内学者虽然起步较晚，不过依然有很多不错的成功。

中国科学院自动化研究所的来斯惟等人[9]设计了一个循环卷积神经网络，在中文长文本分类任务中取得了较好的效果。纽约大学的张翔等人[10]实现了一个字符级别的卷积网络（ConvNets），将中文语句转化为汉语拼音，对中文语料进行分类。北京大学的李嫣然等人[11]通过在文本表示中引入部首信息，构造出了一个性能优异的字向量模型，在中文新闻标题分类上获得了明显的效果。

近几年，已经有学者将计算机视觉领域中效果显著的 Attention 机制应用与自然语言处理中的循环神经网络或卷积神经网络

1.3 课题主要内容

1.4 本论文的结构安排

参考文献

[1] G. Song, Y. Ye, X. Du, et al. Short text classification: A survey[J]. ISSN 1796-2048 Volume 9, Number 5, May 2014, 2014, 9(5): 635

[2] S. Zelikovitz, F. Marquez. Transductive learning for short-text classification problems using latent semantic indexing[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2005, 19(02): 143-163

[3] M. Chen, X. Jin, D. Shen. Short text classification improved by learning multi-granularity topics[C]. IJCAI, 2011, 1776-1781

[4] J. M. Cabrera, H. J. Escalante, M. Montes-y Gómez. Distributional term representations for short-text categorization[C]. International Conference on Intelligent Text Processing and Computational Linguistics, 2013, 335-346

[5] X. Feng, Y. Shen, C. Liu, et al. Chinese short text classification based on domain knowledge.[C]. IJCNLP, 2013, 859-863

[6] R. Socher, B. Huval, C. D. Manning, et al. Semantic compositionality through recursive matrix-vector spaces[C]. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, 2012, 1201-1211

[7] P. Blunsom, E. Grefenstette, N. Kalchbrenner. A convolutional neural network for modelling sentences[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, 655-665

[8] Y. Chen. Convolutional neural network for sentence classification[C]. Proceedings of the 2014 Conference on EMNLP, 2014, 1746–-1751

[9] S. Lai, L. Xu, K. Liu, et al. Recurrent convolutional neural networks for text classification.[C]. AAAI, 2015, 2267-2273

[10] X. Zhang, J. Zhao, Y. LeCun. Character-level convolutional networks for text classification[C]. Advances in neural information processing systems, 2015, 649-657

[11] Y. Li, W. Li, F. Sun, et al. Component-enhanced chinese character embeddings[C]. Proceedings of the 2015 Conference on EMNLP, 2015, 829–-834