doi:10.3772/j.issn.2095-915x.2022.02.004

基于 BERT+A-Softmax 的多分类模型构建与应用研究

开放科学（资源服务）标识码（OSID）

邵德奇 关培培 石聪

科技日报社技术研发部 北京 100038

摘要：[目的 / 意义] 在信息分类领域中，通过传统的机器学习与深度学习的方法可以对大多数稿件进行分类并取得整体较高的准确率。但是这种方法没有对稿件文体进行区别，而现实生产环境中存在新闻类稿件多，通知报告类少等样本不均衡的情况，如果对文体不加以区分，会产生少样本文体类别准确率低的情况。[方法 / 过程] 本文提出一种可以区别文体的深度学习分类模型方法，该方法先根据稿件文体对稿件进行分类，再根据分类结果分别调用分类模型进行进一步分类，解决样本不均衡、小样本文体类别准确率低等问题。[结果 / 结论] 在公开的数据集上实验结果表示，相对于传统的分类模型，本文提出的多分类模型方法在性能上有了显著提高。

关键词：科技新闻资讯；人工智能；自然语言处理；分类体系；BERT

中图分类号：G35

Research on Construction and Application of Multi-classification Model Based on BERT+A-Softmax

SHAO Deqi GUAN Peipei SHI Cong

Technology R&D Department of Science and Technology Daily, Beijing 100038, China

Abstract: [Objective/Significance] In the field of information classification, the traditional classification methods of machine learning and deep learning can classify most manuscripts and achieve high overall accuracy. However, when using this method, there is no distinction between article's style. In production environment, there are many news manuscripts and few notification reports. If it is mixed trained, it will cause low accuracy of few samples class. [Methods/Process] This paper proposes a deep learning classification model that can distinguish article's style. This method first classifies the style of manuscripts according to the style characteristics, then calls the classification model for further classificationl, so as to solve the problems of unbalanced

基金项目 国家重点研发计划项目“基于可信与共治的全媒体内容社会众创服务平台研发与运营示范”（2020yfb1406900）。

作者简介 邵德奇（1965-），硕士，研究员，研究方向为媒体技术应用、媒体评价、舆情监测分析、电子政务等，E-mail：haidixipan@sina.com；关培培（1979-），博士，研究方向为媒体大数据及人工智能等；石聪（1988-），硕士，研究方向为大数据及人工智能等。

引用格式 邵德奇 , 关培培 , 石聪 . 基于 BERT+A-Softmax 的多分类模型构建与应用研究 [J]. 情报工程 , 2022, 8(2): 51-61.

doi:10.3772/j.issn.2095-915x.2022.02.004

samples and low accuracy of few samples class. [Results/Conclusions] The experimental results on the public data set show that the performance of the multi classification model method proposed in this paper has been significantly improved compared with the traditional classification model.

Keywords: Sci-tech news and information; AI; NLP; classification system; BERT

引言

当今，媒体行业正经历着从新媒体到融媒体的转型阶段，技术进步正不断的给媒体行业赋能。科技日报是面向国内外公开发行的中央主流新闻媒体，它是科技领域的重要舆论前沿，每天处理国内外大量的科技类新闻与资讯。当前语料技术处理手段已经比较落后，在效果与体验上有着很大不足，其中之一体现为无法准确的捕捉到稿件的分类。而在当今媒体融合与移动优先的大背景下，个性化检索与推荐等服务是我们要攻破的重点，而精准分类为这一切的基石。因此本文基于报社自身特点，为了实现高质量的稿件分类，更好的服务科技资讯应用，构建了一种基于 BERT+A-Softmax 的多分类模型，效果上好于传统的分类处理技术，为媒体融合发展的技术探索提供了一些思路。

1 相关工作

近年来，网络用户对于新闻内容的获取方式由传统的报纸、电视等单一平台的获取方式转变为移动互联跨平台的内容获取。新闻以网络为载体，可以更加快速精准的聚焦社会热点。网络新闻文本 [1] 具有概括性、层级性、序列性与包容性 4 个特点。面对大量的新闻数据，如何精准、高效地实现文本分类已成为准确定位用户需求的关键，因此也是各大媒体平台亟待解决的问题之一。

已有的文本特征提取方法在主题挖掘等领域有着广泛的应用。如词频统计的文本分析方法 [2-7]，将文本视为词语的集合，通过统计词语频率对文本进行分析，获取文本所属分类；主题概率模型 [8-12]，区别于词频方法，考虑到了语义之间潜在的联系，因此可以获得更高的准确率；相对于传统的机器学习方法，深度学习模型 [13-17] 可以解决数据稀疏和维度过多、模型泛化能力有限等问题，因此分类效果优于前两者。

预训练的方法最初是在图像领域提出的，并且取得了良好的结果，这几年才被应用到自然语言处理领域。基于预训练模型处理业务问题一般分为两步：首先用某个较大的数据集训练模型，将模型表现最好的参数都保存下来，这个保存好的模型文件即被称为预训练模型；然后再根据不同的业务需求微调改造预训练模型，达到各自业务需求。

自然语言处理领域已有的经典预训练模型有 Word2Vec[18]、ELMo[19] 等。Word2Vec 预训练模型使用 Skipgram（跳字模型）通过预测某词的上下文信息来构建模型，由于 word2vec 挖掘了词语之间的关联属性，因此使得语义信息更加丰富，其在文本分类中也取得了广泛的应用 [20-24]。但由于这种方式使得同一个词的输出为同一行向量，因此存在多义词无法识别等问题。ELMo 采用双向双层的长短期记忆网络

（LSTM，Long Short-Term Memory），能够学习语义信息和句法信息，其中每个词的特征向量是通过计算整个句子计算得到的，因此解决了多义词的问题，ElMo 在文本分析领域也取得了不错的成果 [25-28]。但是其所使用的 LSTM 存在抽取特征能力相对弱且不易于并行计算等问题，因此更好的文本特征提取能力亟待发掘。

2018 年，谷歌发布了预训练模型 BERT (Bidirectional Encoder Representation from Transformers)[29]。BERT 是一个预训练的语言表征模型，网络架构使用的是 Attention is all you need[30] 中提出的多层 Transformer 结构。它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的 Masked Language Model（MLM），以致能生成深度的双向语言表征。BERT 论文发表时提及在 11 个 NLP（Natural Language Processing，自然语言处理）任务中获得了新的突破性的结果，在 NLP 业内引起巨大反响。

预训练模型在垂直领域可以通过 Finetune 方式取得效果上的提升。Finetune 方式是指在已经训练好的语言模型的基础上，加入少量的任务导向的模型参数进行再次训练。在分类任务中，常用方法为在预训练 BERT 模型的基础上加一层 softmax 网络，然后在新的语料上重新训练，得到分类模型。

相对于普通 softmax 网络，A-Softmax[31] 优化了损失函数。A-Softmax 将不同类别的输入映射到了同一个单位超球面的不同区域，同一个类别数据向自己类别中心点汇聚，而类别之间通过角度间隔区分。因此 A-Softmax 将以往的优化损失函数内积问题转换成了优化球面类别的角度间隔问题，由于增加了对角度的惩罚，使得决策边界更加严格和具有区分性。实验证明 A-Softmax 相对于普通 softmax 可获得更高的准确率。

2 基于 BERT+A-Softmax 的多分类模型构建

在构建模型前，我们首先基于报社自身的属性与需要制定了一套分类体系。后续的多分类模型的设计与构建围绕着此分类体系展开。

2.1 分类体系

我们分析了报社内部稿件以及各省市科技厅官网开源数据，设计了本文的分类体系。其中包含科学技术、科技服务、科技管理、区域科技、科技人物 5 个一级分类和对应的 28 个二级分类，其对应关系如表 1 所示。

表 1 分类体系介绍

一级分类	二级分类
科学技术	海洋科学、交通运输、人工智能、科学、农林牧副渔、先进制造、航空航天、资源环境、信息技术、材料技术、能源技术、国防科技、汽车
科技服务	科技中介、科技咨询、科技金融、科技评估、科技推广
科技管理	科技计划、科技经费、科技政策、科技成果、科技诚信
区域科技	国际、经济特区、高新区、地方
科技人物	科技人物

doi:10.3772/j.issn.2095-915x.2022.02.004

此分类体系的构建考虑了报社自身业务发展的需要，搭建成功后可提高编辑、记者的选题效率，为其提供更加准确的行业科技资讯。

2.2 总体框架设计

如图 1 所示，多分类模型总体框架由数据收集与分析、文本预处理、模型构建和模型评估这 4 部分组成。

（1）数据收集与分析：基于已有数据，通过分析与梳理，挖掘数据内部分布情况，为后续模型处理提供数据支撑。

（2）文本预处理：对收集到的数据做数据清洗与数据增强，为后续模型的构建提供训练语料。

（3）模型构建：根据语料数据分布情况构建三模型结合的分类模型结构，形成基于科技新闻资讯语料的分类模型。

（4）模型评估：基于模型评估效果，如果不符合预期需要重新补充语料以及加强数据清洗与数据增强能力。

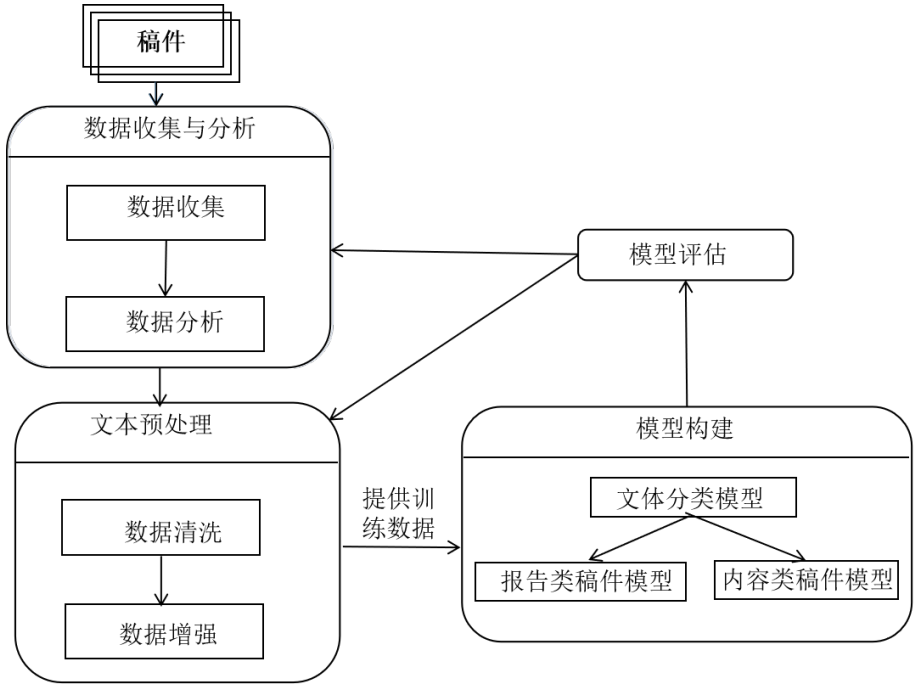

图 1 科技多分类模型构建总框架

2.3 数据收集与分析

如图 2 所示，我们的数据大部分来自报社内部。原则上，在算法领域，数据都起着至关重要的作用。业务数据往往需要来自业务本身，只有这样才会更精准的把握业务数据特性，以便后续模型能够更好的服务于业务。在内部数据不足或者部分内容缺失的情况下，我们爬取了网络中的开源数据以作补充。

我们人工对数据集进行了标注，其中一级分类及二级分类的数据分布如图 3 所示。相对来说，数据多的分类有利于模型训练，相反对于数据少的分类模型无法准确的获取其特征。从图 3 可知我们的数据分布并不均衡，需要对数据进行进一步处理以达到模型训练标准。

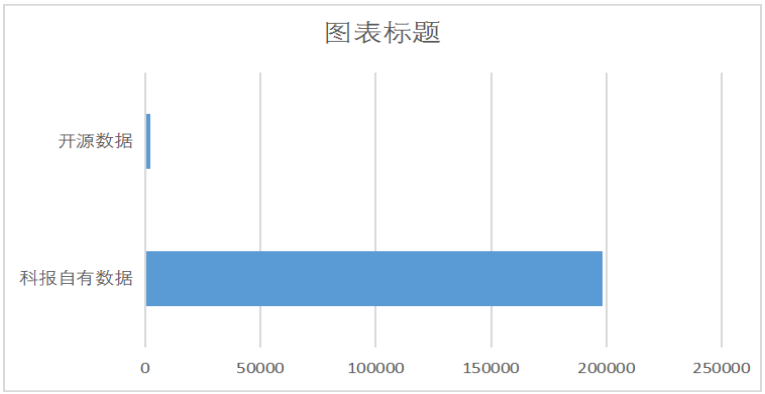

图 2 数据集来源分布

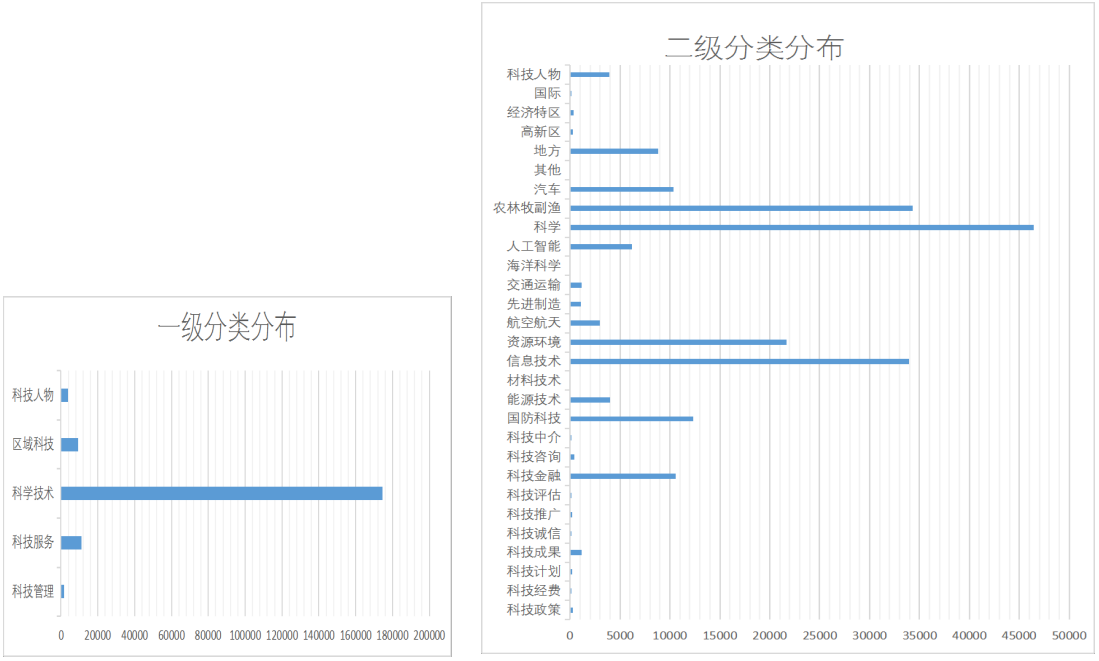

图 3 一、二级分类数据分布

2.4 文本预处理

按照本文设计的分类体系，一个二级分类只对应一个一级分类。因此我们模型训练的目标仅为二级分类，二级分类确定后，一级分类也即确定。文本预处理的主要步骤为文体标注、数据清洗、去重、数据增强以及训练集与测试集切分。

（1）文体标注：在分析稿件时我们发现，科技管理、科技服务中的稿件文体多为公文，如科技计划、科技经费、科技政策、科技评估等，而其他分类文体多为新闻资讯。因此我们需要对稿件进行文体标注以便训练模型对文体进行区分。

（2）数据清洗：数据清洗的主要目的为清理掉文本较短、文本内容不合规的数据。

（3）去重：主要为去掉重复数据，保留发布日期最早的稿件。

（4）数据增强：数据增强为本部分的重点。由于数据不均衡，如果将此批数据直接灌入模

doi:10.3772/j.issn.2095-915x.2022.02.004

型，则会出现“模型失效”等问题。具体原因为：假设我们要做一个二分类模型，正样本占比为99%，负样本占比为1%。因此对于这批数据，只要模型将其全部判断为正样本，则准确率就可以达到99%。而这并不是我们想要的模型，因此我们希望每类数据尽量保持均衡。对于不足的样本，我们除了从网络中爬取有关数据外，还需要对数据进行增强操作。

数据增强的方式有多种，本文所用的方式为翻译处理与同义词替换。具体为（a）翻译：将文本翻译成英文，再翻译回中文，在不改变稿件主题的方式下，改变稿件的表述方式，以达到小样本增强的效果。（b）同义词替换：识别文本中的部分词汇，将其替换成同义词，原理同翻译的方式。

数据增强后保证每个分类有1000篇左右的稿件。

（5）训练集与测试集切分：对经过以上步骤处理后的样本进行随机打乱，选取80%为训练集，剩下20%为测试集。

2.5 模型构建

BERT的网络架构使用的是多层Transformer结构。Transformer是一个编码-解码的结构：编码器由多头Attention和一个全连接组成，用于将输入语料转化成特征向量；解码器输入为编码器的输出以及已经预测的结果，由Masked Multi-Head Attention, Multi-Head Attention以及一个全连接组成，用于输出最后结果的条件概率。

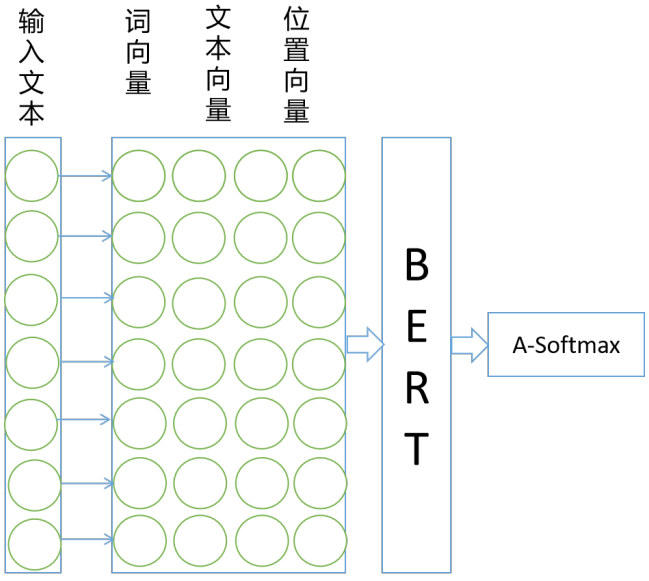

图4 **BERT+A-Softmax**的分类模型

我们将BERT与A-Softmax结合应用于文本分类业务中。从图4中可以看出，BERT模型通过查询向量表，将输入文本中的每个词转换为词向量、文本向量与位置向量，再将这新向量作为模型输入。

（1）词向量：该向量是在模型训练过程中通过学习生成的，表示单个词的语义信息。

（2）文本向量：同样该向量也是在模型训

练过程中通过学习生成的，表示文本的整体语义信息，融合了单个字词的语义。

（3）位置向量：不同位置的词权重存在差异，因此 BERT 使用位置向量表示词的位置信息。

将上述的词向量、文本向量和位置向量作为模型输入，由 BERT 进一步提取语义特征后再进入 A-Softmax 做分类。

最后考虑文体类型，我们需要先设计文体识别模块，通过 BERT+A-Softmax 将稿件进行文体区分，然后根据分类结果输送到不同的分类模型进行进一步的分类划分。

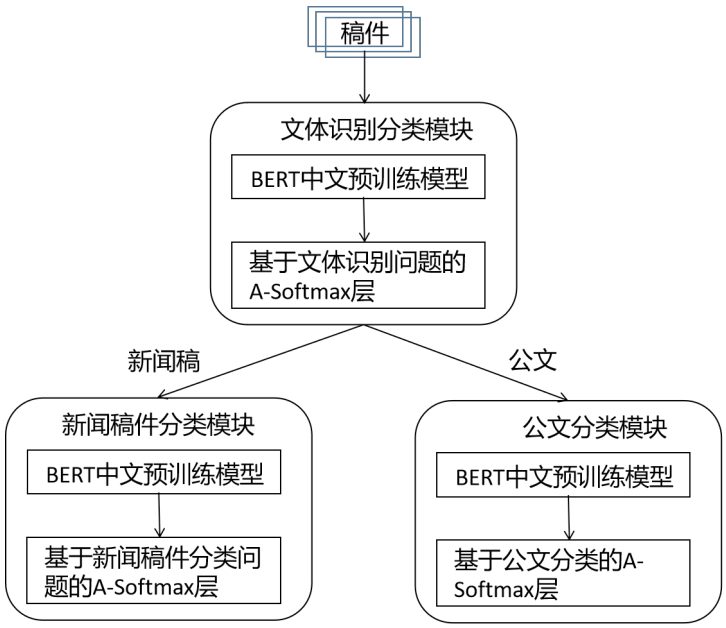

图 5 包含文体识别的 **BERT+A-Softmax** 模型

如图 5，文体识别是一个基于 BERT+A-Softmax 模型训练的二分类模型，目的是识别出稿件的文体类别。而新闻稿件分类和公文分类则单独训练。

2.6 模型评估

评价模型的主要指标有：

（1）准确率（Accuracy）

$$\text{Accuracy} = \frac{TP}{TP + FP} * 100\%$$

其中，*TP* 表示真正例，*FP* 表示伪正例，准确率表示预测为正例的样本中真正例的比例。

（2）召回率（Recall）

$$\text{Re}call = \frac{TP}{TP + FN} * 100\%$$

其中，*TP* 表示真正例，*FN* 表示伪反例，召回率表示预测为正例的样本占所有正样本的比例。

（3）*F*1 值（H-mean 值）

$$F1 = \frac{2 * Accuracy * Recall}{Accuracy + Recall}$$

*F*1 值是对准确率和召回率整体评价的一个指标，只有准确率和召回率都表现的好，*F*1 的值才能高。

doi:10.3772/j.issn.2095-915x.2022.02.004

我们分别对朴素贝叶斯、FastText 和包含文体识别的 BERT+A-Softmax 模型三种训练方式，基于科技新闻资讯语料做分类效果评估，见表 2。

表 2 基于科技新闻资讯语料的模型分类效果对比

模型方法	准确率	召回率	*F*1值
朴素贝叶斯	0.6848	0.6313	0.6283
FastText	0.7226	0.6811	0.6743
包含文体识别的BERT+A-Softmax模型	0.8546	0.8553	0.8547

如表 2 所示，对于同样一批数据，包含文体识别的 BERT+A-Softmax 模型的准确率、召回率及 F1 值均优于其他两个模型。

3 基于 BERT+A-Softmax 的多分类模型在报社科技资讯系统中的应用

科技资讯系统构建的目标为以数据为核心，打造具有鲜明科技领域特色的服务传播平台。科技多分类模型服务在报社系统中的实施流程如图 6：

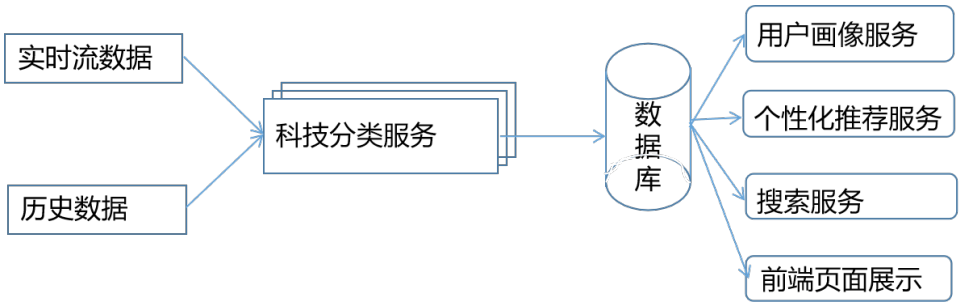

图 6 科技多分类模型服务处理稿件流程

其中，科技多分类模型服务所处理的数据来自两部分：实时流数据与历史数据。实时流数据指的是实时进入系统中的数据，此部分的数据在入库前先经过大数据分类服务打上对应的分类标签；历史数据指的是已经在系统中未打上标签的数据，多来自报社数据资源整合后的留存。

打标后的数据通过数据库保存，后被各个业务系统调用。

在搜索场景下，多分类模型服务的应用如图 7 和图 8 所示。

由图 7 和图 8 可见，左侧导航栏为分类标签，用户可以单选或多选分类以限制检索范围，进而精确的找到目标稿件。图 7，我们选择了“能源技术”作为分类限制条件，而图 8 选择了“汽车”分类。搜索词均为“碳中和”。我们可以看到搜索结果围绕着两个分类范围展开，分别在能源和汽车领域检索到了“碳中和”的有关稿件，达到我们业务预期。

目前，我们每天抓取报社各个端口的数据并对其进行分类打标，并统计打标结果。

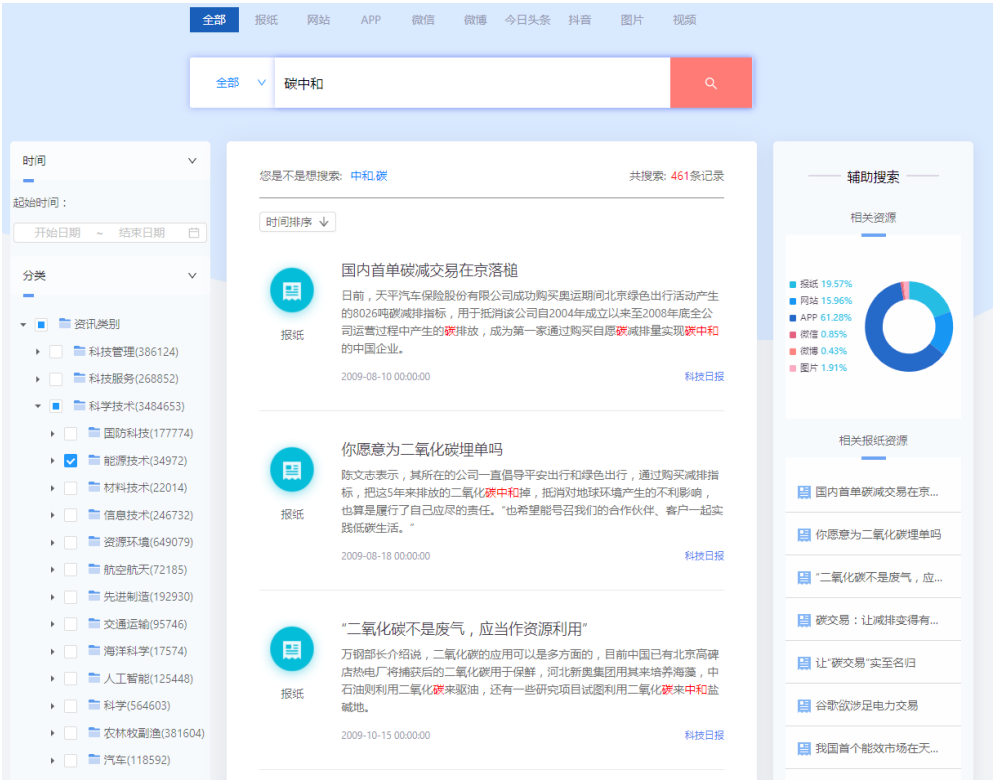

图 7 科技多分类模型服务在报社系统中的应用

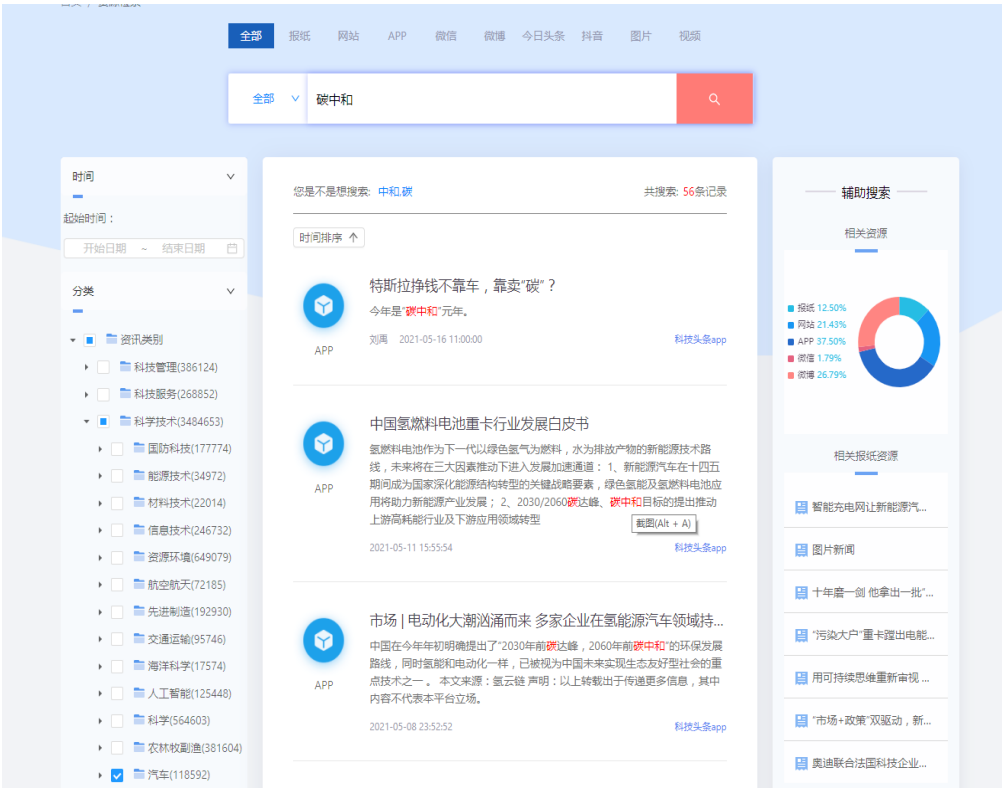

图 8 科技多分类模型服务在报社系统中的应用

doi:10.3772/j.issn.2095-915x.2022.02.004

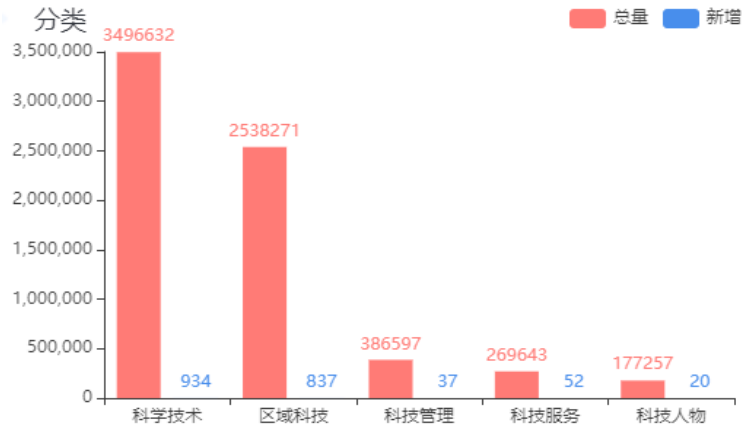

图 9 一级分类存量与日增量统计

如图 9 所示，我们已经实现了对存量 686 万数据以及每天新增近 2 千的数据的分类打标。为进一步媒体融合发展、快速挖掘用户需求，为用户个性化搜索、推荐提供了技术与数据保障。

4 结束语

科技多分类模型服务围绕报社各业务需求建设，以媒体全产业链能更好的在内容领域融合为目标，重点解决大数据时代纸媒技术落后，在分类标签体系构建中存在不足的问题，为打造新的媒体融合目标提供技术能力支撑。在此基础上的针对消费者智能、个性化的推荐服务已经成为未来必然发展的趋势。具有媒体融合特征的个性化新闻服务也需要如此，如何让消费者更轻松、准确的获取科技类的新闻资源将是我们接下来重要的思考和发展的方向。

参 考 文 献

[1] 林纲 . 网络新闻文本结构的语法特征 [J]. 社会科学家 , 2010(7):155-157,161.

[2] 杜芳芳 , 江恒 . 新高考改革政策的价值走向与主题变迁——基于政策文本的词频及共现分析 [J]. 教育与考试 , 2022(1):5-16.

[3] 贺科达 , 朱铮涛 , 程昱 . 基于改进 TF-IDF 算法的文本分类方法研究 [J]. 广东工业大学学报 , 2016, 33(5):49-53.

[4] 公冶小燕 , 林培光 , 任威隆 , 等 . 基于改进的 TF-IDF 算法及共现词的主题词抽取算法 [J]. 南京大学学报 (自然科学), 2017, 53(6): 1072-1080.

[5] 黄承慧 , 印鉴 , 侯昉 . 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报 , 2011, 34(5):856-864.

[6] 覃世安 , 李法运 . 文本分类中 TF-IDF 方法的改进研究 [J]. 现代图书情报技术 , 2013(10):27-30.

[7] 叶雪梅 , 毛雪岷 , 夏锦春 , 等 . 文本分类 TF-IDF 算法的改进研究 [J]. 计算机工程与应用 , 2019, 55(2):104-109+161.

[8] 毕凌燕 , 王腾宇 , 左文明 . 基于概率模型的微博热点主题识别实证研究 [J]. 情报理论与实践 , 2014, 37(2):112-116.

[9] 叶春蕾 , 冷伏海 . 基于概率模型的主题识别方法实证研究 [J]. 情报科学 , 2013(2):135-139

[10] 叶春蕾 , 冷伏海 . 基于引文—主题概率模型的科技文献主题识别方法研究 [J]. 情报理论与实践 , 2013, 36(9):100-103.

[11] 袁柳 , 张龙波 . 基于概率主题模型的标签预测 [J]. 计算机科学 , 2011, 38(7):175-180.

[12] 林洋港 , 陈恩红 . 文本分类中基于概率主题模型的噪声处理方法 [J]. 计算机工程与科学 , 2010, 32(7):89-92+119.

[13] 徐绪堪 , 周泽聿 . 基于多尺度 BiLSTM-CNN 的微信推文的情感分类模型及应用研究 [J]. 情报科学 , 2021, 39(5):130-137.

[14] 陈千 , 车苗苗 , 郭鑫 , 等 . 一种循环卷积注意力模型的文本情感分类方法 [J]. 计算机科学 , 2020, 48(2):245-249

[15] 冯勇 , 屈渤浩 , 徐红艳 , 等 . 融合 TF-IDF 和 LDA 的中文 FastText 短文本分类方法 [J]. 应用科学学报 , 2019, 37(3):378-388.

[16] 何力 , 郑灶贤 , 项凤涛 , 等 . 基于深度学习的文本分类技术研究进展 [J]. 计算机工程 , 2021, 47(2):1-11.

[17] 李洋 , 董红斌 . 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析 [J]. 计算机应用 , 2018, 38(11):3075-3080.

[18] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.

[19] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations [C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics. 2018:2227-2237.

[20] 张谦 , 高章敏 , 刘嘉勇 . 基于 Word2vec 的微博短文本分类研究 [J]. 信息网络安全 , 2017(1):57-62.

[21] 马思丹 , 刘东苏 . 基于加权 Word2vec 的文本分类方法研究 [J]. 情报科学 , 2019, 37(11):38-42.

[22] 程元堃 , 蒋言 , 程光 . 基于 word2vec 的网站主题分类研究 [J]. 计算机与数字工程 , 2019, 47(1):169-173.

[23] 吴德平 , 华钢 . 基于 Word2Vec 词嵌入和聚类模型的安全生产事故文本案例分类 [J]. 计算机系统应用 , 2021, 30(1):141-145.

[24] 邓君 , 孙绍丹 , 王阮 , 等 . 基于 Word2Vec 和 SVM 的微博舆情情感演化分析 [J]. 情报理论与实践 , 2020, 43(8):112-119.

[25] 李铮 , 陈莉 , 张爽 . 基于 ELMo 和 Bi-SAN 的中文文本情感分析 [J]. 计算机应用研究 , 2021, 38(8):2303-2307.

[26] 赵亚欧 , 张家重 , 李贻斌 , 等 . 基于 ELMo 和 Transformer 混合模型的情感分析 [J]. 中文信息学报 , 2021, 35(3):115-124.

[27] 杨书新 , 张楠 . 融合情感词典与上下文语言模型的文本情感分析 [J]. 计算机应用 , 2021, 41(10):2829-2834.

[28] 吴迪 , 王梓宇 , 赵伟超 . ELMo-CNN-BiGRU 双通道文本情感分类方法 [J/OL]. 计算机工程 : 1-10 [2022-03-04]. DOI:10.19678/j.issn.1000-3428.0062047.

[29] Devlin J, Chang M W, Lee K, et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv:1810.04805v1,2018.

[30] Vaswani A, Noam S, et al. Attention Is All You Need[J]. arXiv:1706.03762v5, 2017.

[31] Deep Hypersphere Embedding for Face Recognition. Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, Le Song. ICML 2017