ISSN: 2661-3719(Print); 2661-3727(Online)

基于FastText的电力媒资数据的监督学习优化方法研究

李　嘉　邹海彬　臧艳娇
英大传媒投资集团有限公司　北京　100005

摘　要：本论文旨在研究和探讨基于FastText的监督学习优化方法在电力媒资数据分析中的应用。电力行业积累了大量的媒资数据，包括文本、报告和评论等，这些数据包含了有关电力系统的宝贵信息。然而，有效地分析和利用这些数据仍然是一个挑战。本文介绍了FastText的基本原理和监督学习框架，然后提出了一种优化方法，以提高电力媒资数据的文本分类和情感分析性能。实验结果表明，该方法在电力媒资数据分析中具有显著的优势，有望为电力行业提供更准确和实用的信息。

关键词：电力行业；FastText；监督学习优化

一、基本原理

电力行业是一个信息密集型领域，产生了大量的媒体资源数据，如通讯文稿、报纸、期刊、图书等，为了有效地分析挖掘和应用这些数据，需要对这些数据进行分类。监督学习[1]是一种强大的智能工具，可用于电力媒资数据的文本标引分类、自然语言处理、推荐系统和信息提取等领域，大大提高数据处理和分析的效率。

FastText[2]是一种文本分类和词向量化工具，Facebook AI Research在16年开源的一个用于学习词嵌入和文本分类的模型。在保持高精度的情况下加快了训练速度和测试速度，不需要预训练好的向量词。它建立在连续词袋（CBOW）[3]模型的基础上，通过使用子词级别的n-gram信息来增强性能。

FastText将整篇文本的词及n-gram向量叠加平均得到文档向量，然后使用文档向量做softmax多分类，其具有以下关键特点：

（1）词向量化：词向量化就是把词转化为机器可处理的形式，使用一套统一的标准将词进行向量化，即将词转化为向量形式。FastText将每个词映射到一个向量空间中，以便计算文本的表示。这些词向量捕捉了词汇的语义信息。

（2）子词信息：FastText考虑了词的形态学信息，也就是词的内部构造信息，也就是子词信息，FastText词向量在训练时，会保存每个词的子词信息，因此可以推断出词根。这有助于处理未登录词和复杂的词汇形态。

（3）多类别分类：多类别分类是指在分类任务中，类别数量超过两个的情况。它与二元分类和多元分类有所不同。在二元分类中，我们通常只区分两个类别，而在多元分类中，我们有多个类别，但数量是固定的。然而，在多类别分类中，类别数量是不固定的，它可以有很多个类别。FastText可以用于多类别文本分类任务，如情感分析和主题分类。

二、原有方案

原有的方案首先通过对每个分类节点准备若干条数据，将数据集按照8：2的比例分出训练数据集和测试数据集，使用word2vec基于skip gram模型对预先准备文本数据进行向量化，再进行模型训练。训练完成后，使用测试数据集对训练模型结果进行评估。最后将模型应用于文本分类中。包括以下步骤：

（1）数据预处理：电力媒资数据通常包含通讯文稿、报纸、期刊、图书等。首先，需要进行数据清洗、分词和词干化，以准备文本数据。

（2）模型训练：使用预处理后的数据训练模型。模型将学习文本的表示，并用于后续的分类和分析任务。

（3）文本分类：基于训练好的模型，进行电力媒资数据的文本分类。这可以用于将媒资数据归类到不同的主题或类别中，以便更好地组织和检索信息。

原有的监督学习方法，需要较好的机器学习硬件条件支持，相对成本较高，训练时间较长，整体训练速度

作者简介：
李嘉（1990年1月），女，汉，河北黄骅，硕士研究生，中级工程师，研究方向：人工智能在电力行业媒体的应用。
邹海彬（1986年5月），男，汉族，吉林松原，本科，高级工程师，研究方向：数字化建设及网络安全。
臧艳娇（1989年10月），女，汉，河北保定，硕士研究生，中级工程师，研究方向：融合出版、智能写作和新闻大数据传播等相关领域。

较慢，需要准备大量预训练数据，对分类体系要求较高，对分类节点较多的情况且分类间界限模糊的分类容易混淆，区分较差。如文本属于财经、科技、娱乐、体育等类目清晰明确，且只有一个层级的分类体系结构。电力系统中的分类体系中非常多，并且分类之间的概念界定没有明显的界限。如国家分类下的国家领导与重要人物、国家时间与国家会议，国家电网分类下的营销与宣传、物资与设备，电网建设下的抢修与检修等。

三、基于FastText的电力媒资数据的监督学习优化方法

为了改善电力媒资数据的分类，我们参考了相关文献[5-7]做法，我们提出了一种基于FastText的监督学习优化方法，以优化监督学习在电力媒资数据上的性能，提高电力媒资数据的分析挖掘能力。

1.优化监督学习方法步骤

该优化方法包括以下步骤：

数据预处理：电力媒资数据通常包含通讯文稿、报纸、期刊、图书等。首先，需要收集和准备标注好的电力媒体样本数据。每个分类节点应该准备大约相同数量的分类数据，这些数据应该涵盖该分类节点的各个方面，以便模型能够充分了解该分类的特点。

整合预训练数据：将这些准备好的数据按照节点层级整合起来。一级节点应该放入一个文件中，命名为root。二级节点则应该按照从属的一级节点划分文件，以一级节点的分类名称命名，对应的二级节点数据应该放入文件中。如果有三级、四级节点，也应该按照同样的方式进行划分。需要注意的是，上级节点的数据必须均匀包含标引了子级节点的数据，以确保模型能够全面了解各个分类之间的关系。

文本向量化：使用word2vec等文本向量化方法将准备好的若干个预训练数据集进行向量化并对齐。即将文本数据转化为长度为1024个长度的浮点数数组，长度不足的文本结尾补0。向量是文本分类的重要依据，文本对齐向量化后也就将数据集置于一个坐标系中进行后续的模型训练。

训练分类模型：使用FastText的supervised对每个节点的样本数据进行训练。FastText使用了一个基于线性模型的文本分类算法，使用线性模型（通常是softmax分类器）将文本的特征映射到预定义的类别集合，使用交叉熵损失函数作为优化目标，通过随机梯度下降（SGD）进行模型参数的训练，在训练过程中，通过将文本样本的词向量进行平均或拼接，得到整个文本的表示。然后使用分类模型进行预测，并根据损失函数进行参数更新。

模型评估：使用测试数据集对每个分类节点的模型进行测试。评估结果较差的分类节点模型，对预训练模型中数据样本特征不突出的样本进行剔除，添加特征突出的样本数据，调整后的训练数据集的训练样本数据重复步骤d。评估指标可以使用准确率、召回率等，以综合评估模型的性能。

使用多级分类模型：将分类模型应用到数据分类程序中，对需要进行机器自动分类的文本，首先预测分属于哪个一级分类，对于得分较高的一级分类，再进行二级分类预测，最终将得分较高的二级节点返回，即完成数据的自动分类。在该步骤中的分类程序使用上述步骤中产生的FastText模型进行分类预测，输入文本与训练样本使用相同的词表转化为词向量，将返回结果中得分大于预设阈值的结果分类标签返回。

定期训练模型：模型使用过程中，每月评估标注错误数据（数据标注结果需人工审核，出现错误时人工纠错），标注错误的数据加入到训练数据集中（需要平衡各

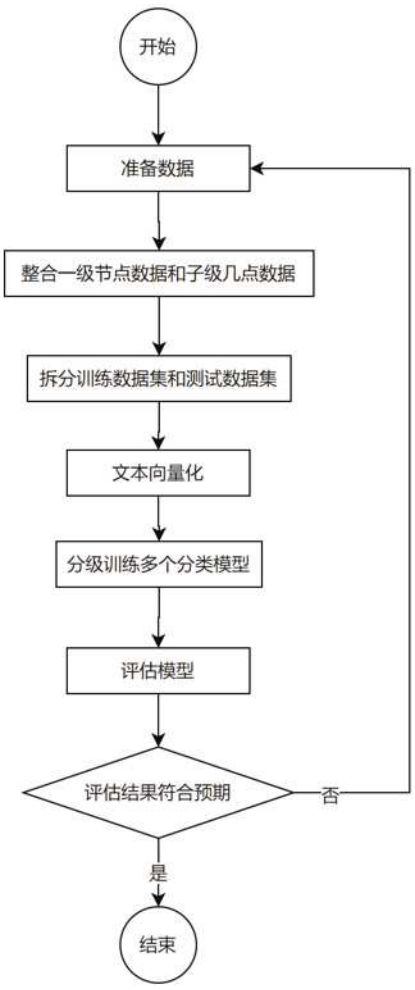

图3-1-1　文本分类模型训练流程图

ISSN: 2661-3719(Print); 2661-3727(Online)

分类的样本数据量，最终结果应保持个分类选取的样本数据条数一致），重新训练分类模型，对模型进行评估，以保持模型的准确性和效果。

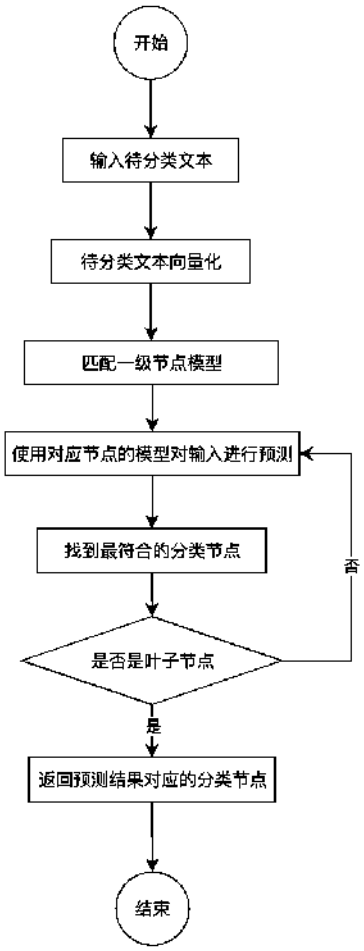

图3-1-2　文本分类预测流程图

2.优化方法关键点

本次优化方法的关键点有：

（1）模型调参：通过调整FastText模型的超参数，如词向量维度、学习率和迭代次数，来优化性能。

（2）数据增强：使用数据增强技术，如文本合成和扩充，以增加训练数据的多样性，从而提高模型的泛化能力。

（3）特征工程：引入领域专业知识，设计新的文本特征或词汇表，以更好地捕捉电力媒资数据的特点。

（4）分层级对多级分类体系的分类模型训练：根据分类体系树节点结构训练多个分类模型，多级分类体系中每个非叶子节点应该训练一个分类模型。

（5）定期加入分类标识错误数据到训练数据集中进行迭代训练：初始训练数据集样本数据不能覆盖所有的情况，在使用过程中积累的错误数据加入到训练数据集中能够增加训练数据集的样本多样性。

3.优化成果

通过在真实的电力媒资数据集上进行实验，我们评估了基于FastText的电力媒资数据的监督学习优化方法的效果。实验结果表明，本次优化方法，能够轻松地适应新的类别或文本类型，采用了层级softmax和负采样技术来加速训练过程，从而能够处理大规模文本数据。也可通过少量数据快速训练出精度较高的实用模型。可以在有限的硬件条件下用少量电力媒体样本数据快速训练出高可用模型，从而高效、快速、精准的对大量电力媒体文本数据进行自动标注分类。

4.成果应用

自动标引：实现内容资源的归类工作。根据标引的数据内容进行模型的学习，之后通过模型进行其他数据内容的自动分类，并在分类过程中根据实际的反馈情况持续调整学习模型，使得分类数据更加准确。

图3-4-1　依据图文内容自动提取文章的分类属性

自动查重：实现资源入库查重计算，根据结算结果生成查重报告，报告整体的重复比例并可查看与其他内容重复的部分。

四、结论

本论文研究了基于FastText的监督学习优化方法在电力媒资数据模型训练中的应用。通过提出优化方法并进行实验验证，我们展示了这一方法在电力行业模型训练中的潜在应用和优势。

在本方法中，针对多层级和定义界限不明确的分类问题，该方法能够通过使用较少量的预训练数据来实现对文本分类的高准确率。其关键点在于：

首先，分级训练模型是一个重要环节。为了减少每个待训练数据集中的分类数量，我们按照节点划分的方式进行分类。通过这种方式，我们能够确保每个待训练数据集中的样本数量相对较少，从而使得模型能够更加专注于每个分类的训练。

其次，定期更新模型也是提高准确率的关键步骤。

当预测结果出现错误时，我们发现这些数据中的文本特征在原始数据集中往往没有得到明显的体现。因此，我们将这部分数据加入到预训练数据集中，这样能够增加这类特征，从而使得模型能够更好地学习并预测这些特征。

通过上述步骤，我们能够实现对文本分类的高准确率，并避免由于多层级和定义界限不明确的问题所带来的分类不准确问题。

这种方法有助于帮助电力公司更好地分析挖掘和利用现有的媒资数据，强化数字赋能，实现转型升级，提升企业核心竞争力，推动建设能源电力领域具有权威传播力和影响力的新型主流媒体。未来的研究可以进一步探索其他领域中的应用和方法改进。

参考文献：

[1]周飞燕，金林鹏，董军.卷积神经网络研究综述[J].计算机学报，2017，40（6）：1229-1251. DOI：10.11897/SP.J.1016.2017.01229.

[2]王江.fastText原理及实践[EB/OL]https：//zhuanlan.zhihu.com/p/32965521.2020-4-25/2023-10-20

[3]周练.Word2vec的工作原理及应用探究[J].科技情报开发与经济，2015（2）：145-148. DOI：10.3969/j.issn.1005-6033.2015.02.061.

[4]谢剑芳，田英明，徐旭，等.基于FastText的专利文本自动分类方法研究[J].仪器仪表标准化与计量，2020（4）：21-24. DOI：10.3969/j.issn.1672-5611.2020.04.008.

[5]代令令，蒋侃.基于fastText的中文文本分类[J].计算机与现代化，2018（5）：35-40，85. DOI：10.3969/j.issn.1006-2475.2018.05.008.

[6]刘测，韩家新.面向新闻文本的分类方法的比较研究[J].智能计算机与应用，2018，8（5）：38-41. DOI：10.3969/j.issn.2095-2163.2018.05.009.

[7]李泽龙.基于FastText的长文本快速精确分类算法研究[D].浙江：浙江大学，2018.