

# 基于文本分类的豆瓣图书分类

韩佳坤<sup>1)</sup>

<sup>1)</sup>(四川大学 计算机学院, 成都 中国 610000)

**摘要** 本项目研究了基于对豆瓣图书分类的文本分类问题。首先，我们人工爬取了豆瓣图书网站上的书本信息及其相关数据，并对其进行预处理后，将图书分类抽象为一个文本分类问题。接下来，我们阅读了文本分类的相关文献，并了解了不同方法的优缺点，选取了BERT模型作为基础模型。我们在其基础上进行改进，同时引入书名短文本和简介长文本，通过融合两者的信息，促进模型对内容特征的提取，从而提高分类性能。最后，我们以两个不同的角度分别开展了实验，验证了我们的改进是有效果的，并且相比基准方法，提高了分类性能。

**关键词** 文本分类 预训练 语言模型 特征融合

## Book Classification Based on Text Classification

Jiakun Han

### Abstract

We study the text classification problem based on the classification of Douban books. First of all, we manually crawled the book information and related data on the Douban book website, and preprocessed it, and abstracted the book classification into a text classification problem. Next, we read the relevant literature on text classification, and understood the advantages and disadvantages of different methods, and selected the BERT model as the basic model. We improve on its basis and introduce the short text of the title and the long text of the introduction at the same time. By fusing the information of the two, we can promote the extraction of content features by the model, thereby improving the classification performance. Finally, we carried out experiments from two different angles to verify that our improvement is effective and improves the classification performance compared to the baseline method.

**Keywords** Text Classification, Pretraining, Language Model, Feature Fusion

---

## 1 引入

近年来，随着互联网的快速发展与普及，数据已渗透到今天的每个行业和业务功能领域，并成为重要的生产要素。随着新一轮的生产力增长和消费者盈余浪潮的到来，海量数据的挖掘和使用预示着“大数据”已经广泛存在于物理学，生物学，环境生态学等领域以及军事，金融，通信等行业[5]。如何对这些海量数据进行有效的挖掘和利用引起了人们的关注。

实际中，按照对数据开发利用程度的不同，一般可将众多的大数据应用分为两个层次。第一层为描述性分析应用，是指从大数据中总结、抽取相关的信息和知识，帮助人们分析发生了什么，并呈现事物的发展历程。如美国的DOMO公司从其企业客户的各个信息系统中抽取、整合数据，再以统计图表等可视化形式，将数据蕴含的信息推送给不同岗位的业务人员和管理者，帮助其更好地了解企业现状，进而做出判断和决策。第二层为预测性分析应用，是指从大数据中分析事物之间的关联关系、发展模式等，并据此对事物发展的趋势进行预测。

然而，某些行业的数据涉及上百个参数，其复杂性不仅体现在数据样本本身，更体现在多源异构、多实体和多空间之间的交互动态性，难以用传统的方法描述与度量。而现有的数据处理方法仅适用于结构化数据，无法将大量的非结构化数据与结构化数据进行统一、整合。如何对复杂条件下的非理想数据进行分析，并全面实时准确地给出分析结果，是大数据技术需要面临的一个重大挑战。

在诸多复杂的数据类型中，文本数据是一种十分常见的数据类型。对于文本数据，它有以下特点：

1、半结构化。文本数据既不是完全无结构的也不是完全结构化的。例如文本可能包含结构字段，如标题、作者、出版日期、长度、分类等，也可能包含大量的非结构化的数据，如摘要和内容。

2、高维。文本向量的维数一般都可以高达上万维，一般的的数据挖掘、数据检索的方法由于计算量过大或代价高昂而不具有可行性。

3、高数据量。一般的文本库中都会存在最少数千个文本样本，对这些文本进行预处理、编码、挖掘等处理的工作量是非常庞大的，因而手工方法一般是不可行的。

4、语义性。文本数据中存在着一词多义、多词一义，在时间和空间上的上下文相关等情况。

经过调研我们发现，在豆瓣图书网站上，陈列的图书多达上百万本。对如此之多的图书信息进行手工处理，将耗费大量的人力物力以及时间精力。同时，除了对历史图书信息的维护，面对未来新增

图书的需求，网站还需要动态进行额外的标注，这无疑将为网站运营维护带来沉重负担。因此，我们受近期AI在文本分类任务[4]中取得的成就激励，尝试设计一种对图书自动分类的系统，并使用深度学习技术来解决这一问题。

总的来说，我们项目的贡献主要总结为以下两点：

(1) 我们基于现有的传统方法，探究了其对于高密度信息分类不理想的问题，并分析其原因。

(2) 我们设计了一种基于预训练语言模型的可结合长文本信息的分类方式，解决了现有方法分类不理想的问题，提高了分类精度。

论文接下来的部分将按如下内容组织：我们将在第二部分介绍我们项目的相关工作，即文本分类任务的相关背景，第三部分介绍我们提出的分类方案以及其具体实现，第四部分介绍两个实验的开展情况，和模型在多个层面上的测试性能，第四部分我们将对本项目进行总结。

## 2 背景知识

### 2.1 文本分类任务

文本分类是指模型将载有信息的一篇文本映射到预先给定的某一类别或某几类别主题的过程[4]，实现这一过程的算法模型叫做分类器。文本分类问题是自然语言处理领域中一个非常经典的问题。文本分类分两种：二分类和多分类。我们组项目解决的是多分类问题。

文本分类方法主要分为传统方法和深度学习方法。传统方法中，词袋模型是一种经典的解决方案，它的基本思想是假定对于一个文本，忽略其词序和语法、句法，仅仅将其看做是一些词汇的集合，而文本中的每个词汇都是独立的，然而，这极其依赖于文本长度和具体内容[9]。SVM算法具有解决高维和非线性问题的强大能力，它最早被[8]用于文本分类。SVM具有较高的泛化能力，但对缺失数据敏感。KNN算法主要靠周围的有限相邻样本，而不是判别类域来确定类别。因此，对于要划分的类域有更多交叉或重叠的数据集，它比其他方法更合适。除此之外，还有DT决策树以及基于图的方法等等。总的来说，传统方法聚焦于一种机器学习算法。它从广泛的文本数据中学习，这些数据是原始数据经过处理产生的预定义特征。然而，这样的特征工程是一项艰巨的工作。对于大型数据集，在计算复杂度的限制下，传统模型通常比深度学习模型表现出更差的性能。

近年来，Word2Vec和深度学习技术的普及，促进了一系列深度学习文本分类模型的发展。

①RNN[3]最早被lai等人用于文本分类。RNN是一种顺序计算的序列模型，不能够被并行计算，这限制了其可扩展性。RNN的该缺点使得在当前模型趋于更深、参数更多的趋势下其分类更为困难。

②CNN[1]. CNN 通过卷积核从文本向量中提取特征。在CNN网络足够深的情况下，理论上它可以捕获远距离的句子特征。但由于深度网络的参数优化方法不足，以及池化层导致的位置信息丢失，更深层的CNN模型并没有为文本分类任务带来明显的提升。与RNN相比，CNN具有并行计算能力，可以为改进后的有效保留位置信息。尽管如此，它的远距离特征捕获能力仍然较弱。

③GNN[6]。GNN的主要思想是为文本构图。当设计出有效的图结构时，GNN学习到的表示可以更好地捕获文本结构信息。

④Transformer[7]。Transformer 将输入文本视为一个全连接图，边上有注意力得分权重。它具有并行计算能力，同时通过自注意力提取不同单词之间的特征非常高效，解决了短期记忆问题。在第二部分我们将进对其行更详细的介绍。

## 2.2 基于深度学习的ransfomer 技术

我们从自然语言处理领域的最新研究情况中，充分进行了调研，并阅读了大量阅读文献，最终确定采用预训练的语言模型Bert。Bert基于目前深度学习领域最为先进的结构之一——Transfomer架构。下面对Transfomer进行简单介绍。

Transfomer可以视作一种编码器-译码器结构。具体来说，Transformer 编码器部分的内部原理为：原始输入inputs 要经过Input Embedding 模块进行向量化，转变为特征向量，然后加上对其的Positional Encoding，形成最终的特征向量向上输入，依次经过Multi-Head Attention，Add & Norm，Feed Forward 模块以及又一个Add & Norm构成的N 个整体模块进行运算，从而得出最终对数据的表征向量。

其中，Multi-Head Attention 的作用最为关键，它主要承担了对数据进行特征提取的功能，多个MLP头使其具有强大的特征提取能力。而Positional Encoding，即位置编码，是用来表示输入句子向量中每个字词所对应的位置的。由于Tranformer 无法像RNN一样获取句子的时序信息，所以我们需要使用Positional Encoding 表示字词在句子中的先后顺序。

而译码器相对来说比较简单，具体解码过程就是decoder端先输入一个起始的token，然后通过self-attention层 和encoder-decoder的attention层，再通过前向层给出输出，从而得到这个token的最后表示。最后通过一个线性层加softmax来预测输

出词典中的哪个词。

## 3 文本分类模型

### 3.1 问题形式化

具体来说，对图书进行分类可以被定义为一个文本分类任务。首先将整个数据集定义为 $D$ ，已有图书产生不同的类别集合定义为 $C = \{\text{文学}, \text{小说}, \dots\}$ 。对于数据集中的任一图书，我们定义其产生的训练数据为： $X_i \in D$ , 其中 $X_i$  的类别标签被定义为 $C_i, C_i \in C$ 。文本分类就是在给定 $X_i$ 的情况下，使用模型对 $X_i$ 附带的相关信息进行建模，从而对 $C_i$ 进行预测。

注意到，虽然对于一个 $X_i$ 在原始数据集中可能对应多个 $C_i$ ，但我们通过对数据集 $D$ 进行预处理，去除了所有有歧义样本标签的样本 $X$ 。

$X_i$	$C_i$
$x_0$ 幽女出没的地方  让所有最最平凡的人也能感受到最最诚挚、温暖的幸福。 不仅读起来很轻松，仔细观察还能发现有很多伏笔！不仅成年人可以读，也非常希望孩子们能看一看。这本书可以成为支撑我们走下去的力量.....	日本文学
$x_1$ 沉默的大多数  终生赋能的思维乐趣。爱智慧、爱自由、反对愚蠢！“上头之选”是走近王小波的三级台阶，《沉默的大多数》是“上头之选”——爱智慧、爱自由、反对愚蠢！收录 40 篇见地不俗、论述精彩、思想深刻的文章，呈现其.....	中国文学

Fig. 1 数据示例

在我们的项目的进行过程中，对模型的设计包含两个阶段，我们在第一个模型中，仅对图书的标题进行分类，即短文本分类任务。在第二阶段中，我们尝试结合图书的标题和图书简介的文本进行预测，结合短文本和长文本进行预测。

### 3.2 文本信息建模

我们考虑使用BERT[2]模型 对文本 进行 分类。BERT的全称为Bidirectional Encoder Representation from Transformers，是一种深度双向的、无监督的语言表示，且仅使用纯文本语料库进行预训练的模型。它具有强大的文本特征提取能力。

之前的部分中，我们介绍了BERT 的整体结构，下面将介绍BERT 模型对文本信息建模的主要过程。首先，bert中将对使用的数据进行预处理。数据预处理对于模型的训练十分重要，关系到模型的训练效率和准确率的提升。BERT 在对数据预处理时，使用了WordPiece 的方法，WordPiece 从字面意思理解就是把字词拆成不同的片。不过，这个方法对我们项目中文的处理是无效的，因为

在中文中每个字都是最小的单位，不像英文使用空格分词，并且许多词还能够进一步拆分，所以对中文使用WordPiece 就相当于按字分割，这也是BERT 的中文预训练模型的一个局限。因此，尽管BERT 中文预训练模型效果很好，但也还存在可以改进的空间。有一些研究者就从这个角度出发对中文BERT 进行了改进，将原BERT 中WordPiece 的分词方法换成了中文分词的方法，然后对词整体添加掩膜，最后进行预训练。在中文数据集测试上，使用这个改进后的预训练模型的测试结果优于使用原版BERT 的中文预训练模型的测试结果。但是在我们的项目中，出于对书名长度的考虑，依然使用基于字分词的base-bert-chinese Tokenizer结构。

然而，若直接使用bert对于一本图书进行分类，可能会出现以下的情况：它的标题能够提供的信息可能很少，或者，模型很难从中挖掘到有效的语义信息。举个例子：有些图书作者为了吸引眼球，可能会起“世界上所有的沙子”这样不明所以的标题，从而即使是人也无法正确分类该书本的类别，除此之外，还有一些书本的名字为了简短起见，使用了例如“闭经记”等的文言文名字，这极有可能使得模型对信息的提取出现困难。

因此，我们考虑融合书本的标题信息和其简介信息，用于bert提取语义特征，以提高分类性能和鲁棒性。具体来说，我们在原有bert基础上做如下的改进：在输入token中引入token-ids 用于区分两个句子，将书名标题的token-ids赋为1，然后引入简介的文本，将书名和简介的token一起输入到bert中，从而融合了短文本和长文本的信息。

### 3.3 模型训练

对于BERT，我们的训练分为两部分，一部分是预训练，一部分是对模型在我们的下游任务上进行微调。

BERT构建的两个预训练任务，分别是Masked Language Model和Next Sentence Prediction。

Masked Language Model (MLM)。该任务能使BERT能够不受单向语言模型所限制。简单来说，我们以15%的概率用mask token ([MASK]) 随机地对每一个训练序列中的token进行替换，然后预测出[MASK]位置原有的单词。然而，由于[MASK]并不会出现在下游任务的微调阶段，因此预训练阶段和微调阶段之间会产生不匹配。因此BERT采用了以下策略来解决这个问题：在每一个训练序列中以15%的概率随机地选中某个token位置用于预测，假如是第i个token被选中，则会被替换成以下三个token之一：1) 80%的时候是[MASK]。如，my dog is hairy——my dog

is [MASK] 2) 10%的时候是随机的其他token。如，my dog is hairy——my dog is apple 3) 10%的时候是原来的token。如，my dog is hairy——my dog is hairy。接下来，再使用该位置去预测出原来的token。该策略令到BERT不再只对[MASK]敏感，而是对所有的token都敏感，以致能抽取出任何token的表征信息。

Next Sentence Prediction (NSP)。一些如问答、自然语言推断等任务需要理解两个句子之间的关系，而MLM任务倾向于抽取token层次的表征，因此不能直接获取句子层次的表征。为了使模型能够有能力理解句子间的关系，BERT使用了NSP任务来预训练，简单来说就是预测两个句子是否连在一起。具体的做法是：对于每一个训练样例，我们在语料库中挑选出句子A和句子B来组成，50%的时候句子B就是句子A的下一句，剩下50%的时候句子B是语料库中的随机句子。接下来把训练样例输入到BERT模型中，用[CLS]对应的C信息去进行二分类的预测。

对于微调，一般是使用较小的学习率在下游数据集上进行训练。在这个阶段，因为我们的任务和原始预训练任务不同，因此采用的预处理方法也有所不同。

在我们的实验中，对于前一部分，由于不具备预训练条件，故我们直接采用了HuggingFace自带的Transformer库中的google中文预训练bert模型。对于后一部分，我们在爬取的数据集按实验设置进行训练。

## 4 实验及结果分析

### 4.1 实验设置

两个实验中，由于我们并未采用具有强烈数据分布偏移的训练数据，因此直接参考了原始bert的设置方案，各项训练超参数与其基本一致。然而，由于我们的文本数据量偏小，我们最终调小了epoch数量，并对模型的性能进行检测，以获取预测准确率最高的模型。现列举部分设置如下：

超参数	值
optimizer	Adam
learning rate	2e - 5
weight decay	0.0002
batchsize	256
learning epoches	50
warmup epoches	5

我们采用交叉熵作为分类损失函数。评价指标，采用总分类准确率，计算公式为：

$$Accuracy = \frac{Correct}{Total}$$

## 4.2 结果分析

我们依据不同的分类方式，将相同的图书数据集分为两个不同的实验。他们分别是：验证模型对不同文体的分类效果的实验一和验证模型对不同文风分类效果的实验二。在实验一中，我们将数据集分为“小说”“散文”“诗歌”“随笔”“杂文”五类。在实验二中，我们将数据集分为“中国文学”“日本文学”“西方文学”三类。对于下述的对比实验与消融实验，我们均分别开展了两次实验。

### 4.2.1 对比实验结果

model	exp1	exp2
Text-CNN	0.230	0.495
Bert	0.390	0.550
Ours	0.420	0.650

为了验证我们改进模型的有效性，在对比实验中，我们选取了Text-CNN[1]和BERT[2]作为baseline，在实验中对比了它们与我们提出方法的性能差异。由上表可以看出，我们方法的准确率高于baseline方法。虽然领先幅度并不大，但这可能受限于我们的训练条件（数据较少，调参时间短）。这说明将长短文本语义结合确实能够促进模型分类。

### 4.2.2 消融实验结果

model	exp1	exp2
Bert(long)	0.330	0.595
Bert(short)	0.390	0.550
Ours(convine)	0.420	0.650

为了更进一步探索模型对长短文本信息结合提取的能力，在消融实验中，我们考虑分别只提取标题短文本或简介长文本的BERT模型，并将他们的结果与我们的模型进行对比。如上表所示，在两个实验数据集上，我们提出的模型的准确率均为最高。

## 4.3 讨论

通过对上述两个实验的实验结果的分析，以及实验过程中的体会，我们发现了以下规律：（1）模型在对图书文体进行分类时，其分类效果不佳，平均表现差于对图书的文风进行分类。这可能是由于

不同文体之间区分度不高，如“散文”和“杂文”之间并无很大概念上的区别，同时，这两者在处理之前的数据部分也有部分重合，因此模型无法有效区分两者。（2）实验过程中，结果较为不稳定。虽然训练epoches设为50，但很多次实验中，训练时间长的模型的准确率不如训练时间较短的模型。这可能归结于实验所爬取数据集的局限性，包括数量少、类别不合适等可能原因。限于技术原因，无法产出更充分有效的实验结果。（3）总体来说，模型分类性能仍然具有较大的提升空间，这可能也需要从数据预处理的角度继续改进，通过减少数据中的噪声来提高模型训练的有效性。

## 5 结论

我们设计了一种基于BERT模型的文本分类模型，它能够提取来源于同一本图书的标题和简介的共同信息，从而在一定程度上解决了项目针对的图书分类问题。具体地，我们在BERT中同时引入书名短文本和简介长文本，使模型在提取文本特征时充分融合两者的信息，从而更好地依据图书的内容对其进行分类。之后，我们分别开展了对比实验和消融实验，验证了我们的改进模型相比基准方法的有效性，以及改进方向的正确性。

未来，我们可能会考虑在分类中引入更多的信息，如与图书内容无关的附加特征（出版时间、作者发行量、发行商），同时研究如何在小规模的数据集下提高模型的表现，减轻数据中噪声的干扰。

## References

- [1] Yahui Chen. “Convolutional neural network for sentence classification”. MA thesis. University of Waterloo, 2015.
- [2] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [3] Siwei Lai et al. “Recurrent convolutional neural networks for text classification”. In: *Twenty-ninth AAAI conference on artificial intelligence*. 2015.
- [4] Qian Li et al. “A survey on text classification: From shallow to deep learning”. In: *arXiv preprint arXiv:2008.00364* (2020).
- [5] James Manyika et al. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.
- [6] Hao Peng et al. “Large-scale hierarchical text classification with recursively regularized deep

- graph-cnn”. In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 1063–1072.
- [7] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [8] Zi-Qiang Wang et al. “An optimal SVM-based text classification algorithm”. In: *2006 International Conference on Machine Learning and Cybernetics*. IEEE. 2006, pp. 1378–1381.
- [9] Zhang Yun-tao, Gong Ling, and Wang Yong-cheng. “An improved TF-IDF approach for text classification”. In: *Journal of Zhejiang University-Science A* 6.1 (2005), pp. 49–55.

## 贡献评述

### 1. 确定选题阶段：

- 调研阶段。韩佳坤提出《共享单车交通流预测》《豆瓣图书分类》两个预备选题。李星提出《根据地理信息预测国家发达程度》预备选题

- 经过讨论，最终选定《豆瓣图书分类》为选题。

- 一分钟堂上报告演讲者：韩佳坤 PPT制作：韩佳坤

### 2. 论文阅读：

- 论文、相关方法调研搜集：韩佳坤

- 论文堂上报告：韩佳坤 汤东儒 李星各负责1/3

PPT制作：汤东儒

(之后李星退出)

### 3. 论文项目实现阶段

- 图书数据爬取：韩佳坤

- 数据预处理、数据集制作：韩佳坤 汤东儒

- 模型搭建：韩佳坤

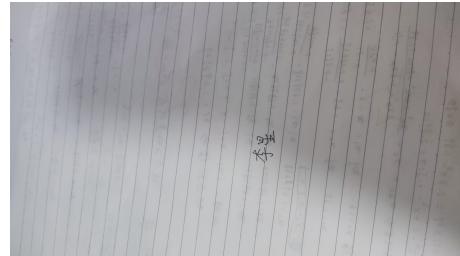
- 实验开展：韩佳坤

- 堂上与老师交流及反馈记录：韩佳坤

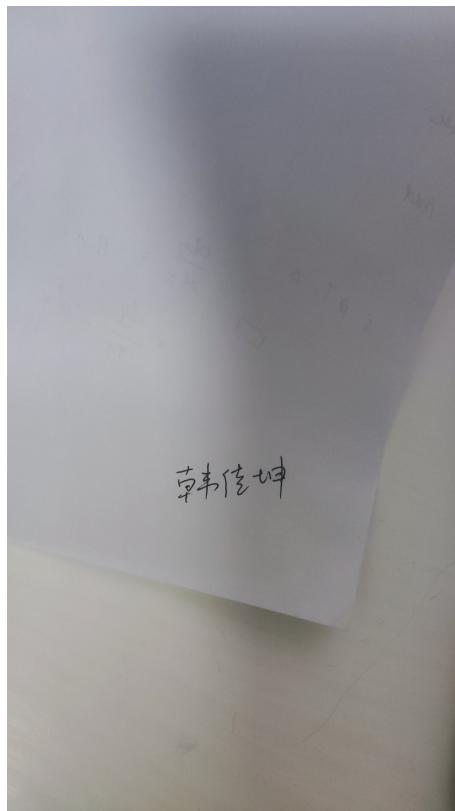
### 4. 结束阶段

- 项目小论文完全由韩佳坤一人撰写完成

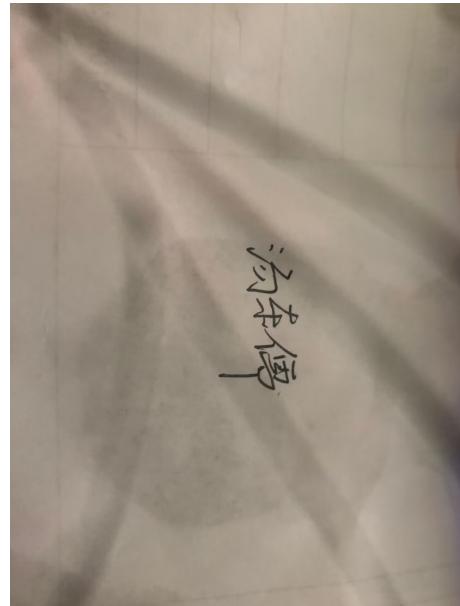
- 最后项目报告演讲者：韩佳坤 PPT制作：韩佳坤 汤东儒



**Fig. 3**



**Fig. 2**



**Fig. 4**