计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700111

基于 BERT 的社交电商文本分类算法

李可悦[1]　陈　轶[2]　牛少彰[1]
1 北京邮电大学计算机学院　北京 100876
2 东南数字经济发展研究院移动大数据中心　浙江 衢州 324000
(likeyue@bupt.edu.cn)

摘　要　随着网络购物的高速发展,网络商家和购物者在网络交易活动中产生了大量的交易数据,其中蕴含着巨大的分析价值。针对社交电商商品文本的文本分类问题,为了更加高效准确地判断文本所描述商品的类别,提出了一种基于 BERT 模型的社交电商文本分类算法。首先,该算法采用 BERT(Bidirectional Encoder Representations from Transformers)预训练语言模型来完成社交电商文本的句子层面的特征向量表示,随后有针对性地将获得的特征向量输入分类器进行分类,最后采用社交电商文本的数据集进行算法验证。实验结果表明,经过训练的模型在测试集上的分类结果 F1 值最高可达 94.61%,高出 BERT 模型针对 MRPC 的分类任务 6%。因此,所提社交电商文本分类算法能够较为高效准确地判断文本所描述商品的类别,有助于进一步分析网络交易数据,从海量数据中提取有价值的信息。

关键词:多标签文本分类;特征提取;模型构建;双向编码器;机器学习
中图法分类号　TP181

Social E-commerce Text Classification Algorithm Based on BERT

LI Ke-yue[1],CHEN Yi[2] and NIU Shao-zhang[1]
1 School of Computer Science,Beijing University of Posts and Telecommunications,Beijing 100876,China
2 Mobile Big Data Center,Southeast Digital Economic Development Institute,Quzhou,Zhejiang 324000,China

Abstract　With the rapid development of online shopping,a large amount of transaction data has been generated in online transaction activities between online merchants and shoppers,which contain great analytical value. Aiming at the text classification problem of social e-commerce product texts,in order to more efficiently and accurately determine the category of products described in the text,this paper proposes a social e-commerce text classification algorithm based on BERT model. The algorithm adopts the BERT pre-trained language model to complete the feature vector representation of social e-commerce text on sentence-level,and then inputs the obtained feature vectors into the targeted classifier for classification. In this paper,we use the social e-commerce text data set for algorithm verification,and the results show that the F1 value of the trained model on the test set can reach up to 94.61%,which is 6% higher than the MRPC classification task based on the BERT model. Therefore,the social e-commerce text classification algorithm proposed in this paper can more efficiently and accurately determine the type of goods described in the text,which is helpful for further analysis of online transaction data and extraction of valuable information from massive data.

Keywords　Multi-label text classification,Feature extraction,Model building,Bidirectional encoder,Machine learning

1　引言

随着移动互联时代的快速发展,互联网逐渐成为人们购物、娱乐、出行的重要平台。2019 年,在经济整体下行的大背景下,电商平台上商品类和服务类交易仍然保持了两位数的增长。社交电商成为了网络消费增长的新动能,其增长势头迅猛,已发展成为了网络消费的新生力量[1]。这一方面显示了我国电商业务促进消费的潜力很大,另一方面也显示了社交电商交易模式在激活消费市场方面的积极作用和重要影响。

社交电商通过将关注、分享、讨论、沟通等社交化元素应用到电子商务的购买服务中[2]来更好地完成交易。相比传统电商以商品为中心的模式,社交电商以人为中心,通过用户评论分享来传播商品信息,形成口碑效应,从而激发消费需求,其最常见的方式就是使用文字评论对商品进行分享和传播[3]。由此产生的大量社交电商文本携带了丰富的数据信息,包括商品的品牌、类型、名称等。为了更好地促进社交电商的发展,对这些社交电商文本进行分类具有较好的实用价值。

针对社交电商文本的特殊背景,本文总结出了社交电商文本的 4 个特点,具体如下:

(1)文本篇幅较短。社交电商文本以商品转发、商品评

到稿日期:2020-07-17　返修日期:2020-12-04
通信作者:牛少彰(szniu@bupt.edu.cn)

论、朋友圈分享等形式传播，通常文本长度较短，不超过 200 字，而且内容相对精简，并不要求上下文之间有很强的逻辑和因果关系，主要以推销和展示商品为目的。

(2)文本的不规范性。社交电商文本出现在人际交流的场合下，文本内容注重交流性，语句倾向于简洁易读，因此文本内容相对口语化，很可能缺少严格的句子主从结构，甚至可能会使用一些新潮的、能博取购物者眼球的网络流行词、表情符号、短语短句等，导致文本数据的噪声较大，具有一定的不规范性，这加大了对文本内容分词的难度。如果训练模型不能正确地对句子进行分词，则提取其中的词语实体就比较困难，从而导致其文本向量表示不够准确。

(3)文本类别的多样性。社交电商文本涵盖的商品类型范围较广，例如本文所统计和标记的商品类型有服饰、首饰、美妆个护、医疗保健、旅游、电子产品、汽车等 20 种商品类型，这增加了社交电商文本的分类难度。

(4)不同类别商品的文本内容特点鲜明。例如，对于服饰这类商品，其商品的文本内容通常会出现“好看”“有气质”“合身”等词汇。对于旅游产品，其文本内容通常会包含“风景优美”“景色宜人”“山川湖泊”“自由行”“自驾游”“几天几晚”等词汇。通过对不同类型商品文本的整理分析可以发现，同一类型商品文本的内容有相似的聚类特点，因此适合使用机器学习的方法来找到其中蕴含的规律，再将其准确分类。

国内外众多学者针对中文文本分类的研究主要体现在以下 3 个方面：中文文本的特征表示、分类算法的选择与改进以及分类方法的效果和评估。

文献[4-8]旨在对中文文本的特征表示提出改进。文献[4]提出一种基于 word2vec 的中文文本表示方法，该方法使用 word2vec 词嵌入技术对文本的分词结果进行向量表示，再使用 TF-IDF 对每个词向量进行加权，最后使用 SVM 算法进行文本分类。实验结果表明该方法可以有效地提高中文文本的分类效果。文献[5]提出使用一种中文字符的注意力表示模型，通过这种方法可以更好地针对中文文本提取中文语句特征表达。文献[6]提出一种 N-of-DOC 的文本表示方法，通过神经网络和信息增益的方法从整个训练集的语句特征中提取有效特征，再经过 word2vec 词向量表示后，使用卷积神经网络的卷积层和池化层提取高层特征，最后使用 Softmax 分类器进行分类。文献[7]提出了一种基于注意力机制的中文文本分类特征增强融合模型，结合一个长短期记忆(Long Short-Term Memory，LSTM)网络、一个卷积神经网络(Convolutional Neural Networks，CNN)和一个特征差异增强注意力算法模型，将中文文本数字化为向量形式，其中将包含某些语义上下文信息的向量放入嵌入层，以便通过预处理来训练和测试神经网络。文献[8]对多标签学习算法中的分词方法进行了比较研究，并使用支持向量机、随机森林、K-最近邻的方法进行实验，结果表明使用支持向量机具有更好的性能。

文献[9-12]主要对文本分类算法的选择与改进进行了研究。文献[9]提出了一种 ELMo(Embeddings from Language Models)方法，该方法是一种双向 LSTM 结构，其训练的语言模型可以学习到句子的上下文信息，在一定程度上解决了模型只能单向学习信息的问题。文献[10]提出了利用 Transformer 的编码器进行预训练的 OpenAI-GPT 模型，OpenAI-GPT 模型使用一种半监督的方式来处理语言理解的任务，使用非监督的预训练和监督方式的微调，目标是学习一个通用的语言表示，经过很小的调整就可以应用在各种任务中。与 LSTM 相比，GPT 语言模型的优点是可以获得句子上下文更远距离的语言信息，并且可以训练一个通用的语言模型用于下游任务。文献[11]采用正则化权值的方式对 K 近邻算法 KNN 进行改进，并结合 PSO(粒子群优化)算法提高了文本分类的效果。文献[12]提出了一种 GloVe 模型，采用共现矩阵，考虑了文本数据的局部信息和整体信息。文献[13]对文本分类发展历程和各阶段的技术进行了概括，总结了文本分类效果的评估指标，如精确率、召回率、均衡点(BEP)和 Fβ(常用 F1)值等。

上述文献针对文本信息的特征表示和算法的提出与改进做了大量工作，对于一般的文本信息内容有一定的效果，但对于特定领域的文本，不能很好地提取文本内容中的特征。由于社交电子商务文本特征稀疏，字符和单词不能完全表达其完整语义，因此社交电商文本的特征表示是提高该领域文本分类性能的关键步骤。

基于以上研究，为提取有效的特征表达，以更加高效准确地对文本所描述商品的类别进行判断，本文提出了一种基于 BERT(Bidirectional Encoder Representations from Transformers)模型的社交电商文本分类算法。该算法使用 BERT 预训练语言模型对社交电商文本进行句子层面的特征向量表示，随后将获得的特征向量输入有针对性的分类器中进行分类。BERT 预训练语言模型基于双向 Transformer 编码器的结构，具有较强的语义表达能力，通过使用本文收集的社交电商领域数据集进行训练，使得模型能够学习到社交电商文本的特征。实验结果表明，针对社交电商文本数据集，本文提出的方法可以有效提升分类任务的分类准确度。

2 基于 BERT 的社交电商文本分类算法

本文提出的基于 BERT 的社交电商文本分类算法主要由社交电商文本数据集的预处理、语言模型预训练、语言模型微调以及分类器分类 4 部分组成。社交电商文本预处理包括对预训练语料库的收集、对社交电商文本数据集的收集、数据集预处理，并将处理好的语料库和数据集进行向量化表示，形成特定格式的数据，然后将这些数据输入到 BERT 语言模型中，对语言模型进行预训练和微调。

2.1 社交电商文本数据集的预处理

(1)数据的收集。对于 BERT 模型预训练过程和模型微调过程，本文收集了两种不同的数据集，一种是包含大量文本内容的大型语料库，另一种是特定的社交电商文本数据集。

本文收集的大型语料库用于 BERT 模型的预训练过程，该语料库包含约 100 万条内容和长度不限的有标签的文本数据。为了收集该语料库，本文下载和使用了两个公开的在线文本数据集，分别是 THUCNews 中文文本分类数据集[14]和搜狐新闻数据库[15]，其中分别包含约 70 万条和 400 万条文本分类数据，且都以分类标签和文本内容的形式存储。本文还在线爬取了一些公开的线上媒体的文本分类数据[16]，同样

以分类标签和文本内容的形式存储。将这些数据收集、清洗、整理后，最终形成数据量大约为 100 万条的用于模型预训练的大型语料库。

本文还收集了社交电商特定领域的文本数据集，用于 BERT 模型的微调过程，因为微调过程将针对社交电商文本数据这类特定领域数据的分类任务进行优化。

社交电商的商家主要通过社交软件从事相关的商业活动，包括发布广告、寻找客源、维护客户关系、进行商品推广等。为了能够采集到规模大、质量高的社交电商文本数据，我们与社交电商辅助软件的提供商开展了合作，借助其一款社交电商软件辅助工具——智能空间，来获取参加实验的近万名社交电商商家的包含商品信息的文本数据。最终，我们采集到了社交电商用户在微信朋友圈、微博等社交软件公开发表的有关商品信息的文本数据，共包含 3.89 万条有标签的社交电商文本，其中标签的类别有 20 种。我们将该社交电商文本数据集用于 BERT 模型微调过程的训练集以及最终检验模型效果的验证集和测试集。

(2)数据集的清洗和整理。该过程主要包括去除文本内容中的特殊符号、表情符号，去除多余空白以及统一文本字体为简体。去除特殊符号、表情符号和多余的空白，主要是针对社交电商文本内容的复杂性，让文本内容的特征表示尽可能只关注文本信息中的有效词汇和语义本身，降低特殊符号和表情符号对模型的特征提取过程的影响。统一文本字体为简体，就是使得数据集中包含的文字和词汇都尽可能地包含于词汇表中。如果文本中使用的词汇不在其建立的词汇表中，就会让当前词汇使用初始化的向量表示方法，从而失去词汇本身的语义。

除此之外，本文在进行数据集的预处理时还去除了一些无用词和口语化的词汇。通常中文文本中存在着高频且无实际意义的词，它们对社交电商文本分类的任务是无用的，如“吧”“啊”“呃”等词汇，去除这些词汇，可以降低文本数据集的噪声，使得模型更加关注具有实际意义的词汇。口语化的词汇，如“早安”“晚安”“开心一刻”等与商品本身无关的词汇，主要用于吸引人的注意力，提高其他人的关注度，但与商品本身的信息无关。经过这样的预处理，在一定程度上降低了输入文本的特征维度，从而提高了文本分类处理的效率和效果。

(3)类别匹配。由于本文所使用的训练算法是有监督的，因此对于数据集中的数据需要标明该样本的分类类别，并将原始文本与其对应的类别一一匹配。

2.2 BERT 模型

BERT 是基于双向的 Transformer 编码器实现的[17]，其 Transformer 编码单元由 6 个 Encoder 堆叠在一起，解码层也一样。Transformer 的总体结构如图 1 所示。

对于编码器来说，一个编码器包含两层，分别是 Self-Attention 层和前馈神经网络，Self-Attention 层能帮助当前节点不仅只关注当前的词，还能获取上下文的语义。一个解码器也包含 Self-Attention 层和前馈神经网络层，但是在这两层中间还有一个 Attention 层，其作用是帮助当前节点获取到当前需要关注的重点内容[18]。编码器和解码器的内部结构如图 2 所示。

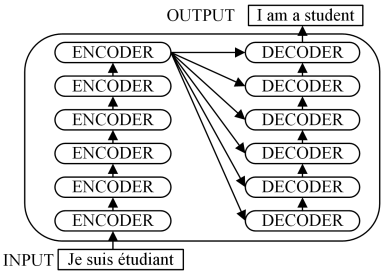

图 1 Transformer 的总体结构

Fig. 1 Overall structure of Transformer

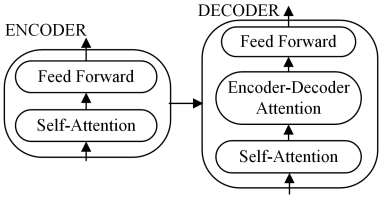

图 2 编码器和解码器的内部结构

Fig. 2 Internal structure of encoder and decoder

首先，模型需要对输入的数据进行词嵌入(Word Embedding)操作，然后将结果输入到 Encoder 层，进行自注意力处理和前馈神经网络的计算，得到的输出会进入到下一个 Encoder 层。

通常，解决这种序列问题的 Encoder-Decoder 结构的核心是基于循环神经网络(Recurrent Neural Network，RNN)实现的，但是 RNN 的结构存在无法并行、运行慢的缺点，为了弥补这一不足，Transformer 使用自注意力(Self-Attention)机制来替代 RNN 的结构。多层自注意力机制代替传统的 RNN 和 CNN，有效地解决了自然语言处理中的长期依赖问题，自注意力机制的核心思想就是计算一个句子中的每个词与这句话中所有词的相互关系，再通过这些相互关系来调整每个词的权重比例，从而使系统能够得到每个词语在这个句子中更高维度的含义，并且这样的表达还蕴含了句子中其他词和这个词的关系，因此它是对于整个句子的全局表达。自注意力机制的运算过程中，首先会计算出 3 个新的向量：***Q-Query***，***K-Key***，***V-Value***，这 3 个向量是词嵌入向量与一个矩阵相乘得到的结果，该矩阵是随机初始化的维度为(64，512)的矩阵。当输入一个句子时，该句子中的每个词都与其他的词进行 Attention 计算，Attention 的计算公式如下：

$$Attention(Q,K,V)=Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

其中，d_k 表示每个字的 *query* 和 *key* 向量的维度，$Softmax()$ 是归一化指数函数。最终得到的 Attention 值是一个矩阵值，矩阵值的每一行代表输入句子中相应字的 Attention 向量，其中包含了句子中该词和其他位置的词的相互关系信息，是一个新的向量表示。例如“The animal didn't cross the street because it was too tired”，对于机器来说很难判断本句中的 it 是指 animal 还是 street，基于自注意力机制和 Attention 值的计算，就可以得知此句中的 it 与 animal 的关联性较强，从而使系统能够正确获知该句子表达的语义内容。

由此，我们可以看到，BERT 模型使用带有自注意力机制

的双向 Transformer 模型获得了句子的前后语义关系，从而更好地获得了一个句子的语义表达。在预训练过程中，BERT 模型还在预训练的目标任务上有所创新，这就是 BERT 模型特有的 MLM 任务(Masked Language Model)和 NSP 任务(Next Sentence Prediction)。MLM 任务可以使语言模型更好地利用上下文的信息对当前词语进行编码，而 NSP 任务可以使语言模型掌握句子与句子甚至章节之间的因果关系。

MLM 可以被理解为完形填空任务，程序会随机遮盖掉每一个句子中 15%的词，然后采用非监督的学习方法预测被遮盖的位置是什么词汇，目的是让 BERT 模型能够实现深度的双向表示。例如，句子“古北水镇的风景很好”中部分词被遮盖掉后变为“古北水镇的[MASK]很好”。为了在微调阶段尽可能降低[MASK]标记带来的负面影响，模型采用的具体策略如下：

(1)80%的情况下采用[MASK]标记，即“古北水镇的风景很好”，变为“古北水镇的[MASK]很好”；

(2)10%的情况下采用一个随机词汇来替代被遮盖的词，即“古北水镇的风景很好”可能变为“古北水镇的味道很好”；

(3)10%的情况下保持原句不变，即保持“古北水镇的风景很好”不变。

NSP 任务是，给定一些句子对(A,B)，其中 50%的数据中的 B 是 A 的下一句子，剩余 50%的数据中的 B 则是语料库中随机选择的。添加这样的任务的目的是让语言模型能够理解两个句子之间的逻辑和因果关系，从而使得模型能够更好地处理需要理解上下文逻辑关系和因果关系的任务，如自然语言处理中的问答任务(QA)和自然语言推理(NLI)。

2.3 基于 BERT 的社交电商文本分类算法的研究

本文提出基于 BERT 的社交电商文本分类算法，具体流程如图 3 所示。

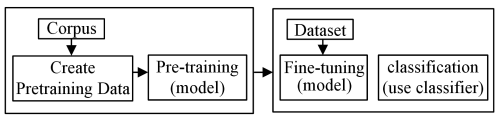

图 3 本文算法流程图

Fig. 3 Flow chart of proposed algorithm

步骤 1 语言模型的预训练过程。首先准备好前文所述的大型语料库，每一条数据按照标签、tab 分隔符、文本内容的方式储存在文本文档中。将大型语料库文档文件按照适当大小分割(本文按 500 MB 一个文件进行分割，以避免文件过大而在处理过程中计算机内存被占满)，分别输入生成特征向量程序，生成以 TFRecord 为格式的训练数据。再将准备好的数据作为预训练数据集，对模型进行预训练。

预训练过程不仅需要大量的数据集作为语料库的输入，而且要消耗大量的计算资源。本文使用 Google 提供的预训练检查点 BERT-Base, Chinese 作为起始检查点[19]，继续预训练过程。在条件充足的情况下，我们也可以从参数的初始状态开始训练。

步骤 2 语言模型的微调过程。将前文准备好的社交电商文本数据集中的每一条数据同样按照标签、tab 分隔符、文本内容的方式储存在文本文档中。同时，将数据集中 1/10 的数据作为测试集保存，不再参加训练过程，用于测试模型的最终效果。由于本文收集的社交电商文本数据集含有 20 个标签种类，因此，对于每一个标签种类，都提取出其 1/10 的数据量作为测试集。

步骤 3 设计一个适用于本文数据的分类器，根据实际情况，将分类标签设为“首饰”“服饰”“美容”食品“香烟”“汽车”“房产”“金融”“书画”“旅游”“培训”“推广”“加粉”“杂文”“卡”“电子产品”“美妆个护”“医药保健”“话费充值”和“其他”共计 20 种。分类器接受一行文本内容的词向量作为输入，通过语言模型的判别，输出该文本内容的预测分类标签，经过与其本身标注的标签类别对比，可以得知预测的标签类别是否正确。

2.4 对 BERT 模型训练过程的调整优化

根据前文总结的社交电商文本内容的特点，即文本内容篇幅短、文本内容不规范和上下文之间难以形成很强的因果关系，本文提出对 BERT 预训练语言模型的改进和调优方法，使其能够更适用于本文研究的社交电商文本语言领域。

(1)降低 NSP 任务的目标占比。社交电商文本内容往往缺少很强的上下文因果关系，内容多以推销和商品介绍为主，因此使用加权的方法对 MLM 任务和 NSP 任务目标进行调整，一定程度上降低了 NSP 任务对模型预训练的影响，使预训练模型更加适合本课题研究的语言文本数据。

(2)降低 Transformer 层级深度。社交电商文本内容与文章、小说和作文不同，不存在大量的语句结构、句式和主谓关系。多层级的网络结构主要用于提取深层次的特征表达，目的是解决语言语句中深层次的含义与联系的问题。但社交电商文本语言具有简单、直观、直接的表达特点，往往不具有深层次含义的特性，使用过度提取深层次的特征表达的预训练语言模型会对简单直观的社交电商文本内容过度解读，进而带来致使错误产生的不利影响，因此，在层级结构上，应适当降低神经网络结构的深度。

3 实验过程及结果

3.1 实验数据

本文实验使用的社交电商文本数据集是本文收集的社交电商用户在微信朋友圈、微博等社交软件公开发表的有关商品信息的文本数据，共包含 3.89 万条有标签的社交电商文本，标签的类别有 20 种，分别为首饰、服饰、美容、食品、香烟、汽车、房产、金融、书画、旅游、培训、推广、加粉、杂文、卡、电子产品、美妆个护、医药保健、话费充值和其他。根据文本的数据量，设置了 3 组实验，每一组实验文本数据量分别设为 8000, 20000, 36000 条，每一组数据集均按照 8:1:1 的比例进行训练集、验证集以及测试集的划分。实验数据集的分布情况如表 1 所列。

表 1 实验数据集的分布情况

Table 1 Distribution of experimental data sets

	首饰	服饰	汽车	…	电子产品
A 组(8000)	400	400	400	…	400
B 组(20000)	1000	1000	1000	…	1000
C 组(36000)	1800	1800	1800	…	1800

3.2 实验环境

本文实验环境如表 2 所列。

表 2 实验环境

Table 2 Experiment environment

实验环境	具体配置
操作系统	CentOS 7.7
CPU	Intel Core(TM) i7-8700 CPU @3.2GHz × 4
GPU	GTX 1080Ti(16GB)
Python	3.6
TensorFlow	1.14
内存容量/GB	16

3.3 实验参数

本文在 Google 的 BERT-Base，Chinese 预训练模型上继续进行训练，模型的网络结构为 12 层、768 维隐藏层、12 头模式，含有 110×10^{12} 个参数。训练过程的参数和超参数如表 3 所列。

表 3 训练过程中的参数

Table 3 Parameters in training process

参数名	参数值
max_seq_length	128
train_batch_size	32
learning_rate	2×10^{-5}
num_train_epochs	3.0

3.4 评价指标

本文研究的分类问题常用的评价指标包括结果的准确率(Accuracy)、精确率(Precision)、召回率(Recall)以及结果的 F1 值(H-mean 值)。通常以关注的类为正类，其他类为负类，分类器在数据集上的预测结果为正确或者不正确，因此有 4 种情况，对应的混淆矩阵如表 4 所列。

表 4 分类结果的混淆矩阵

Table 4 Confusion matrix of classification results

实际值	预测值	
	预测为正例	预测为反例
真实为正例	*TP*	*FN*
真实为反例	*FP*	*TN*

(1)准确率 A 指所有预测正确的数量占总量的比例，计算公式如下：

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

(2)精确率 P 指正确预测为正的占全部预测为正的比例，计算公式如下：

$$P=\frac{TP}{TP+FP} \tag{3}$$

(3)召回率 R 指正确预测为正的占全部实际为正的比例，计算公式如下：

$$R=\frac{TP}{TP+FN} \tag{4}$$

(4)为了评价不同算法的优劣，在 Precision 和 Recall 的基础上提出了 $F1$ 值的概念，来对 Precision 和 Recall 进行整体评价。$F1$ 值是一个综合了 P 和 R 的指标，是基于 P 和 R 的加权调和平均，计算公式如下：

$$\frac{2}{F1}=\frac{1}{Precision}+\frac{1}{Recall} \tag{5}$$

$$F1=\frac{2PR}{P+R}=\frac{2TP}{2TP+FP+FN} \tag{6}$$

可以发现，当 $P=1$，$R=1$ 时，F1 值达到最大值为 1，此时精确率和召回率都达到了 100%。在实际情况中，很难达到二者均为 100%的情况。使用 $F1$ 值来评价分类器性能时，其值越接近于 1，说明分类器的性能越好。

3.5 实验结果

根据数据规模的不同，本文进行了 3 组实验，分别为 A 组、B 组和 C 组，数据规模依次增加。在每一组实验中，训练参数和测试集均保持不变，评价指标主要采用 $F1$ 值，实验结果如图 4 和图 5 所示。

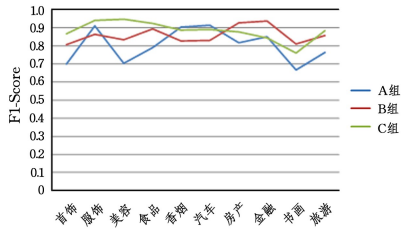

图 4 3 组实验结果的对比

Fig. 4 Comparison of three groups of experimental results

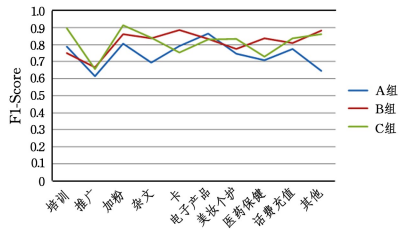

图 5 3 组实验结果的对比

Fig. 5 Comparison of three groups of experimental results

图 4 和图 5 的横坐标为每个类别的具体预测情况，纵坐标为该类别预测结果的 F1 值，3 条折线分别为 A 组、B 组和 C 组实验的情况。可以看出，社交电商文本数据集较小，模型的 $F1$ 值最低，随着数据量的增大，模型的效果有一定提升，但 B 组和 C 组的实验结果相似。

从类别角度来看，对于“服饰”“金融”“旅游”这几个类别的预测准确率较高，原因可能是这些类别的文本内容比较明显，例如，服饰的文本内容大多带有尺码的字样，金融的文本内容经常会涉及到贷款和银行卡，而旅游的文本内容也经常带有“几天几晚”“几人成团”的字样。对于“美容”“美妆个护”“书画”“其他”这几个类别来讲，准确率较低，可能是由于其文本特征不够明显，并且文本长度也较短。

与 Google 提供的 MRPC 语料库的分类结果相比，MRPC 数据集的分类结果为 84%～88%[20]。可以看到，本文提出的针对社交电商文本数据集的分类算法能够对特定领域的数据类别预测有一定提升。

此外，由模型的整体效果可以看出，社交电商文本数据量较小时，模型也有一定的分类准确度，说明 BERT 模型对于提取文本内容的语义有良好的效果。随着实验数据量的增加，模型的 $F1$ 值均有上升，$F1$ 值最高可达 94.61%，可以看出，基于 BERT 的社交电商文本分类算法对于该领域文本分类任务有着良好的效果。

结束语 本文描述了基于 BERT 模型的社交电商文本分类算法的研究，对于社交电商文本这种特定内容的文本数据，在收集了相关数据集的基础上，通过 BERT 语言模型双向 Transformer 结构获取句子的语义表示，再通过分类器对文本数据的类别进行分类。实验结果表明，本文方法对于社交电商文本分类任务有良好的效果。另一方面，本文算法存在的问题是对数据集的质量要求较高，需要收集足够的特定领域的数据集，进行良好的数据预处理，并且需要为数据集添加正确的标签，因此提高了使用成本。此外，本文提出的标签类别是多种多样的。在实际使用中，标签类别可以根据实际情况减少，如类似的标签可以组合成一个，以降低分类的复杂性。

文本分类问题只是提取海量交易数据所蕴藏价值的一个方面，在未来的工作中，还可以根据文本内容进行价值判断、实体识别、自动生成广告等研究。在信息科技全面融入人们生活的时代，数据中所蕴藏的价值非常值得我们去发掘和研究。

参 考 文 献

[1] CNNIC. The 45th "Statistical Report on Internet Development in China" (Full Text) [OL]. (2020-04-24)[2020-11-01]. http://www.cac.gov.cn/2020-04/27/c_1589535470378587.htm.

[2] WANG B. The Essence, Causes and Practical Trends of "New Retail" [J]. China Business and Market, 2017(7): 3-11.

[3] YU H. The Development Status, Trends and Countermeasures of New E-commerce Business Types in China [J]. China Business and Market, 2016, 30(12): 47-56.

[4] LI Z, DUAN M. Research of Chinese Short Text Classification Based on Word2vec [J]. Computer Life (CPL), 2019, 7(2): 90-96.

[5] QIAO X, PENG C, LIU Z, et al. Word-character attention model for Chinese text classification[J]. International Journal of Machine Learning and Cybernetics, 2019, 10: 3521-3537.

[6] WANG L. Research on Chinese short text classification method based on hybrid neural network [D]. Hanzhou: Zhejiang Sci-Tech University, 2019.

[7] XIE J, HOU Y, WANG Y, et al. Chinese text classification based on attention mechanism and feature-enhanced fusion neural network [J]. Computing, 2020, 102: 683-700.

[8] HE J, WANG C, WU H, et al. Multi-label chinese comments categorization: comparison of multi-label learning algorithms [J]. Journal of New Media, 2019, 1(2): 51-61.

[9] 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. 2018: 1-6.

[10] ALEC R, KARTHIK N, TIM S, et al. Improving Language Understanding by Generative Pre-Training [EB/OL]. [2020-07-01]. https://s3-us-west2. amazonawa. com/openai-assets/research-covers/language-unsupervised/language_unders-tanding_paper. pdf.

[11] WU F, ZHENG Y. Adaptive normalized weighted KNN text classification based on PSO [J]. Scientific Bulletin of National Mining University, 2016(1): 109-115.

[12] JEFFREY P, RICHARD S, CHRISTOPHER M. Glove: Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014.

[13] FABRIZIO S. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[14] SUN M, LI J, GUO Z, et al. THUCTC: An Efficient Chinese Text Classifier[EB/OL]. [2020-07-01]. http://thuctc.thunlp.org/.

[15] Sohu News Data [EB/OL]. [2020-03-01]. https://www.jianshu.com/p/370d3e67a18f.

[16] Netease News Data[EB/OL]. [2020-03-01]. https://news.163.com/.

[17] JACOB D, CHANG M, KENTON L, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv: 1810.04805v2, 2018.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need [J]. arXiv: 1706.03762v5, 2017.

[19] Google. Pre-trained models, google-research, bert[EB/OL]. [2020-05-10]. https://github.com/google-research/bert#pre-trained-models.

[20] Google. Sentence (and sentence-pair) classification tasks, google-research, bert[EB/OL]. [2020-05-10]. https://github.com/google-research/bert # sentence-and-sentence-pair-classification-tasks.

LI Ke-yue, born in 1995, postgraduate. His main research interests include big data processing and machine learning.

NIU Shao-zhang, born in 1963, Ph.D supervisor, is a member of China Computer Federation. His main research interests include digital image forensics and information security.