计算机和电气工程 100 (2022) 107852

内容列表可在ScienceDirect上找到

计算机和电气工程

杂志主页: www.elsevier.com/locate/compeleceng

使用机器学习技术的社交媒体的人格预测模型☆。

Murari Devakannan Kamalesh [a,*], Bharathi B [b]

[a] 印度钦奈Sathyabama科技学院计算机科学与工程系研究学者。
[b] 印度钦奈萨蒂亚巴马科技学院计算机科学与工程系。

我的朋友们，你们知道吗？

关键词。
人格预测 大五人格特征
二元分割变压器(BPT) Term Frequency & Inverse Gravity Moment (TF-IGM)

A B S T R A C T

从Facebook、Twitter和Instagram等社交媒体应用中预测人类的行为和性格，正在获得研究人员的极大关注。关于人类通过社交媒体上的状态所表达的想法的统计信息是预测各种人类行为和个性的研究的重要资产。目前的工作主要集中在根据五大人格特征来猜测用户的性格。我们建立了一个智能句子分析模型来提取人格特征。在这篇文章中，提出了一个新的二元分割变换器（BPT）与术语频率和反重力矩（TF-IGM），该变换器可以识别数据集中的特征集和特质之间的关系。所提出的工作在多个社会数据集上的表现优于所有特征提取平均基线集。在Facebook数据集上取得了最大的F1分数0.762和准确率78.34%；在Twitter数据集上取得了0.783和79.67%；在Instagram数据集上取得了0.821；86.84%。

1. 简介

在过去的十年中，像Facebook、Twitter、Instagram等社交媒体在用户中获得了最高的人气。用户在社交媒体上的活动被专家系统研究，以识别用户的行为和个性特征[1]。对个性的分析是基于社交媒体上规模庞大的形式之间互动时使用的特征。流行的五种人格特征被用来分析行为模式。用NLP提取特征的其他方法是语言学查询和字数（LIWC）和，语言学线索提取的结构化方案（SPLICE）二进制分区变换器（BPT），术语频率和反重力矩（TF-IGM）。现有的关于预测推特平台上的人格特征的研究使用监督机器学习算法的基准问题。与现有工作相关的主要问题是，对个性特征进行不必要的数据集分类处理。它降低了系统的个性预测性能。维度的诅咒是一个与词的排名有关的问题，而数据的稀疏性增加了数据的维度[1]。这项工作使用了二元分割变换器（BPT）与术语频率和反重力矩（TF-IGM）特征提取技术的融合来克服这些问题。它通过使用皮尔逊相关技术从社交媒体文本数据中选择特征，以获得更好的结果，通过机器学习算法预测用户的个性。

这项研究工作讨论了六个方面的贡献。

☆ 本文为VSI-hci2特别章节。评论由特邀编辑Carlos Montenegro博士处理并推荐发表。

* 通讯作者。
*电子邮件地址：*kamal3gd@gmail.com（M.D. Kamalesh）。

https://doi.org/10.1016/j.compeleceng.2022.107852

2021年9月13日收到；2022年2月18日收到修订版；2022年2月22日接受

2022年3月11日可上网查询
0045-7906/

- 展示了用户的个性与他们在社交媒体平台上的互动行为之间的关系。通过使用二元分割变换器（BPT）、术语频率和反重力矩（TF-IGM）、语言学查询和字数、语言线索提取结构化方案（SPLICE）以及二元分割变换器（BPT）术语频率和反重力矩TF-IGM的融合，实施特征提取技术，得出个体用户的高潜力特征。
- 对于高维度的特征选择，我们使用皮尔逊相关技术对社交媒体数据集的用户之间的特征进行关联。
- 使用最大熵分类器的机器学习方法，预测社交媒体平台上的人格特征，进行探索。
- 用现有的算法评估预测系统的性能，并利用特征分析结果。

该研究文章以文献调查的方式组织了相关工作，以及如何实现所提出的算法，即二元分割变换器（BPT）与（TFIGM）术语频率和反重力矩的融合。并讨论了在不同的数据集上使用该算法所获得的结果。

2. 相关工作

最近，许多与预测社交媒体网络中的人格特征有关的研究论文在研究界获得了关注。之前关于预测人格的研究使用了Twitter、Instagram、Facebook的一些特征技术，如语言学查询和单词计数（LIWC）、语言线索提取的结构化方案（SPLICE）、SNA（社会网络分析）和基于时间的特征[2].Facebook数据集选择了250个数据集，在研究工作中使用了语言学查询和单词计数（LIWC）和SNA（社会网络分析）的特征。论文[3]讨论了通过使用开放性词汇差分语言分析（DLA）和语言学查询和词数（LIWC）特征来预测Facebook状态的个性。在Facebook的数据集中处理了带有单字的词袋（BOW）算法。Twitter的特征是Lin- guistic Inquiries and Word Counts (LIWC)，MRC被用于一篇文章中[4]，用于人格评估。在巴哈萨，Twitter的数据集是用五种人格模型[5]处理的。[6]提出了机器学习算法来对Facebook的人格模型进行分类。术语频率/反向文档频率（TF-IDF）实现技术有助于从数据集中识别出最接近的人格关键词的词汇[7]。人类心理学的理论被用来作为选择五个因素的预测方案。更多的研究者接受这种技术来描述社交媒体上人类个体的基本预测结构。一些不必要的个人形容词被删除，以实现高效预测。文章[8]通过使用机器学习算法分析人们在社交应用程序上的互动/推文，提出了一个创新的想法。这有助于预测他们的想法和猜测他们的个性。文章[9]提出的特征选择方法使用TF/IDF、术语频率/反向文档频率、语言学查询和词数（LIWC）等来构建向量。构建的向量被用于分类算法的训练和测试模块，如Naive Bayes、神经网络和SVM（支持向量机）分类器。这里，SVM（支持向量机
机器）显示出良好的预测和分类结果，准确度很高。数据集MBTI（Myers-Briggs类型指标）是
在[10]中介绍的性格预测，从Reddit社交媒体中得出的特征。性格分析需要一个富裕的数据集

表1
关于预测人格特征的调查。

作者	目标	技术	优势	劣势
Mehta, Y.; Majumder, N.(2020) [13]	性格预测	深度学习	性能高，速度快执行	需要 复杂的结构
Fiok, K. et al. (2020) [14] 。	句子		分类MLM用于高级	NLPL较低的准确性
苏赫巴托尔等人（2019年）[15] 。	文本	分类 适应跨度 变换器	减少高度减少内存	复杂的连接性质。在处理大型数据集方面很复杂使用成本；长期未实施文本
刘和拉帕塔等人（2019年）[16] 。	文本	分类分层变换器		
Bharadwaj等人（2018）[17] 。	在线	文本SVM、TF-IDF、LIWC	等对MBTI的所有维度都有 最好的准确性。	对于词的分析，给出了一个较小的权重系数值
Chaudhary等人（2018）[18] 。	使用MBTI	模型在线文本NaïveBayes, SVM, LR和随机森林	减少了内存	使用量降低了精度
Kaur和Gosain（2018）[19] 。	Buraya等人（2017） [21]	采样技术的不平衡数据集。Reddit数据集的人格预测。		使用超采样的SVM和MLP 决策树更好的性能在语言学特征
Gjurković和Šnajder (2018) [20]		使用大型多源数据集，NUS-MSS进行人格预测		

	需要更多的重新取样方法 取 。 中 不考虑年龄和性别 表 现 处理大型数据集的性能差 表 佳 监督了特征矢量的最佳性能			
Ong等人（2017）[22]。 Ngatirin等人（2016 ）[23]。	Big5 Model for Bahasa 基于Twitter数据集的ML分类 器个性预测	tweetsXGBoost, 随机森林、随机树、 SVM	SVMFaster 更好的	ExecutionLimiteddataset of only. 精确度复杂数据集

和基准模型，以便有效地检测个别字符。SVM（支持向量机）、逻辑回归等，是情感分析过程中使用的一些分类模型。 MBTI（Myers-Briggs Type Indicator）具有基于语言的特征，性能良好。论文[11]使用MBTI数据集进行性别和性格预测。他们使用了几乎120万条推文，分析MBTI（迈尔斯-布里格斯类型指标）数据集的四个维度。N-gram logistic regression被用来选择合适的特征。从通过印尼语和英语发布的推文中识别人格因素[12]。为了获得更好的结果，在Facebook数据集中使用了两个或更多的分类器。 表1显示了基于预测人格特征的文献回顾。

3. 使用二元分割变压器(BPT)实现所提出的方法，并采用期频和反重力矩(TF-IGM)。

社交媒体是一个平台，世界各地的人们可以在这个平台上相互交流。个性化的公共档案通过文字、图片、音频和视频表达用户的想法。最受欢迎的媒体，如Twitter、Facebook和Instagram，在过去几年中迅速增加。各种语言被用于对话，根据用户状态和档案进行判断[24]。目前的工作包括五个阶段：数据收集、预处理、特征提取和选择，然后是用于性状预测的机器学习分类。每个阶段的描述见图1。

3.1. 数据收集

来自Twitter、Facebook和Instagram的数据集被用来观察社交网络的个性行为。Instagram的数据集是从Instagram的API（应用程序接口）收集的。Facebook的数据集被称为 "我的个性"，而Twitter的数据集是由心理学专家手动收集的；注解被定义为Twitter用户的个性特征[25]。这些数据集在本研究工作中被用来分析五大人格特征的心理测试。五大人格特征在图2中给出。

表2显示了大五人格特征在OCEAN方面的特点。

3.2. 数据预处理

社交媒体文本数据与标准英语不同，这是执行NLP任务的最大挑战。所有来自Twitter、Instagram API、My Personality的社交媒体数据集将在特征提取之前进行预处理。预处理使文本数据集正常化，因为Facebook的数据集是用英语写的，而Twitter的数据集是用印尼语写的[27]。预处理社交媒体数据需要采取一些步骤，以消除污点、URL、标签、符号、表情符号、空格、小写字母、删除停顿词和词根。为了对数据集进行预处理，我们采用了spaCy。spaCy是一个开源库，提供了丰富的处理功能，NLP（自然语言处理）协助清理社交媒体数据集。图3描述了社交媒体的预处理阶段。

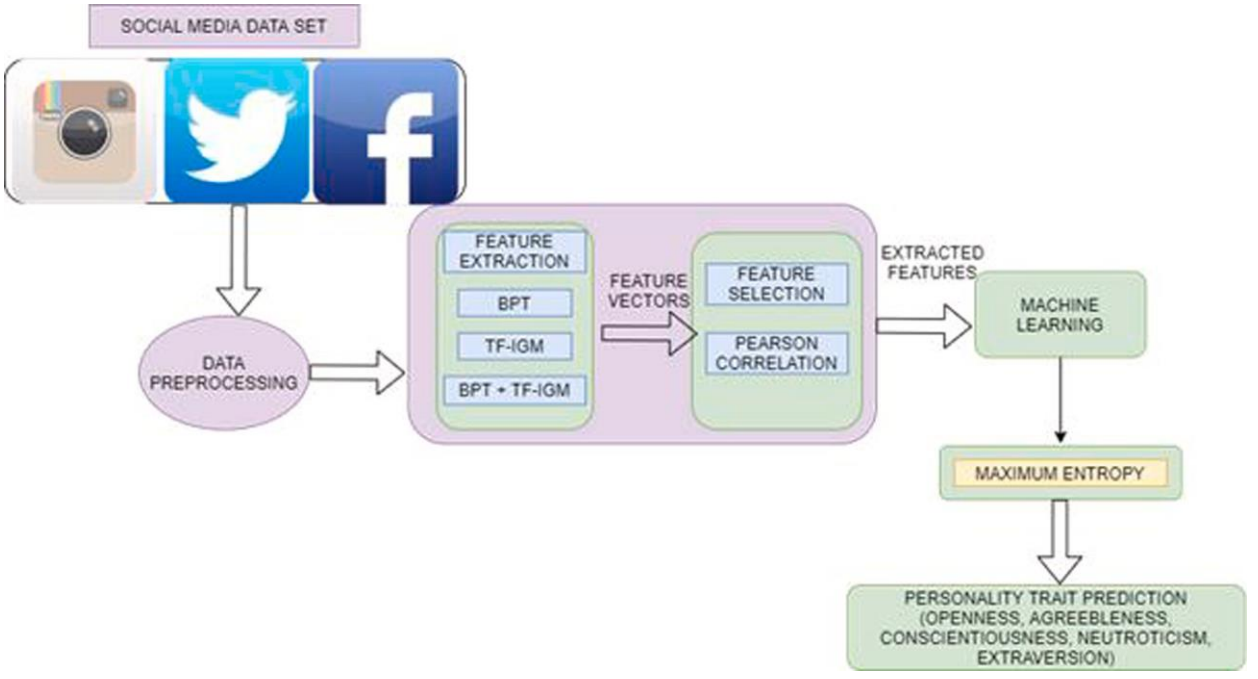

图1.工作流程模型。

图2. 五大人格特征。

表2
大五人格特征的特点[26]。

人格	特征　性格特征
开放性	(O)审美敏感，更关注内在感受，创造力，权威挑战
	自觉性 (C)细致性,完美主义, 判断力
外向性	(X)大胆，大嗓门，自尊心强
赞同性	(A)友好、无私、灵活、宽恕
神经质	（N）人类的　　情绪波动（愤怒、抑郁、焦虑）。

数据。以下是预处理的摘要。

第1步：通过删除URL、符号、标签来清理社会数据。
第2步：删除制表符、行距、标点符号。
第3步：删除在语句序列中重复的元音。
第4步：删除表情符号
第5步：将文本转换为小写，删除多余的空白。

3.3. 特征提取

不相关的人类行为普遍滥用了人类在社交媒体平台上的活动，这种影响对用户可能过境的新思想或他们的智慧反映在群体中 为了分析社交媒体文本数据状态的内容，采用了传统的两种字典的方法，即语言学查询和单词计数（LIWC）和语言学线索提取的结构化方案（SPLICE）。在这项拟议的工作中，从社会数据集收集的所有信息都可以使用二元分割变换器（BPT）与TFIGMTerm频率和反重力矩的融合来进行预训练。

3.3.1. 二元分割变换器（BPT）

通过二进制分割将输入序列分割成多尺度的跨度，图的每个节点都被认为是一个输入标记。神经网络变换器在二进制分割变换器中使用，以获得更好的自我注意系统的效率。自我注意过程有助于预测识别单句中的单词之间的关系[28]。图4显示了使用二进制分割变换器（BPT）的预训练模型。

为了获得从细到粗的注意效果，我们通过二元分割（BP）将一个序列分割成多颗粒的跨度，以获得从细到粗的注意效应。它是一个将输入序列一分为二直至满足要求的递归过程。在这项工作中，它需要一个长度为n的输入序列；有$^{2n-1}$个分区。图5说明了BPT（）序列的例子。

上面的例子二进制分割变换器（BPT）序列显示了一个完美的二进制树的建立，所有的内部节点都有两个孩子。输入序列标记被认为是一个叶子节点。图6显示了自我关注机制。自我注意过程采用递归方法。一个超级句子被分成两个子句，然后这两个子句被当作超级句子，进一步分成两个子句。这个过程一直持续到每个词都在独立的子句中。

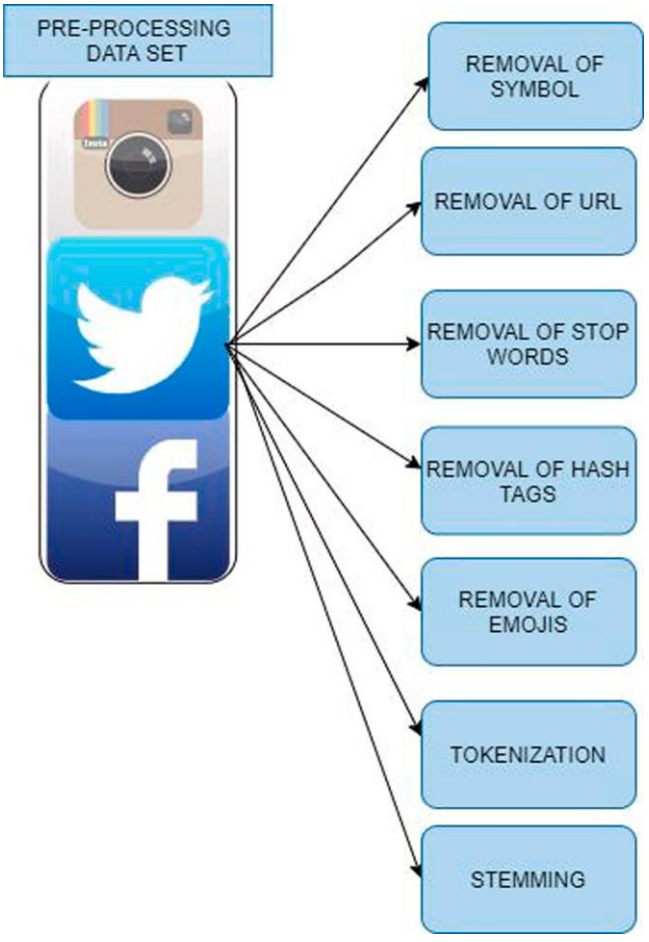

图3. 社交媒体数据的预处理阶段。

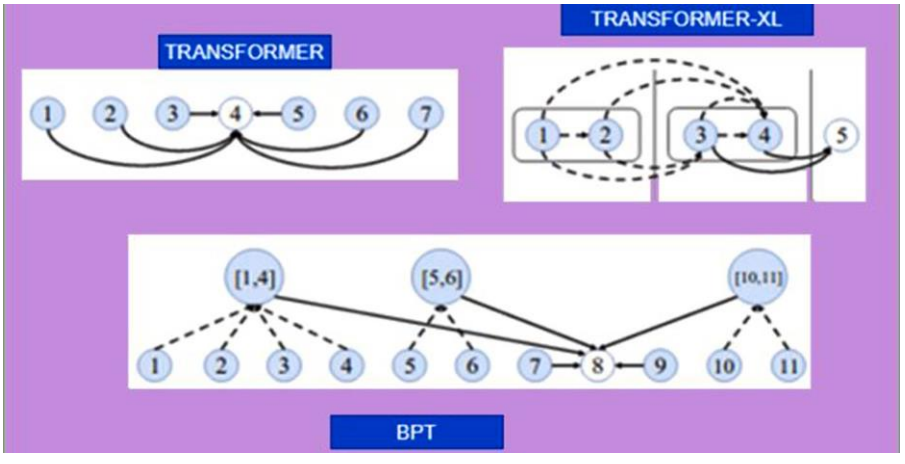

图4. 使用BPT的预训练模型。

句子。

3.3.2. 术语频率和反重力矩(TF-IGM)

它将术语频率（TF）与IGM测量相结合，形成术语频率和反重力矩（TF-IGM）。它被用来

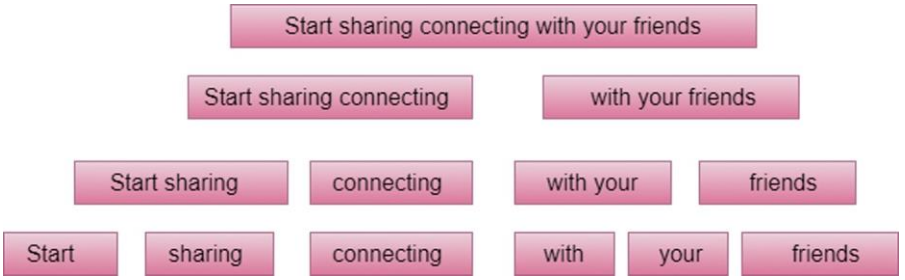

图5. BPT序列示例。

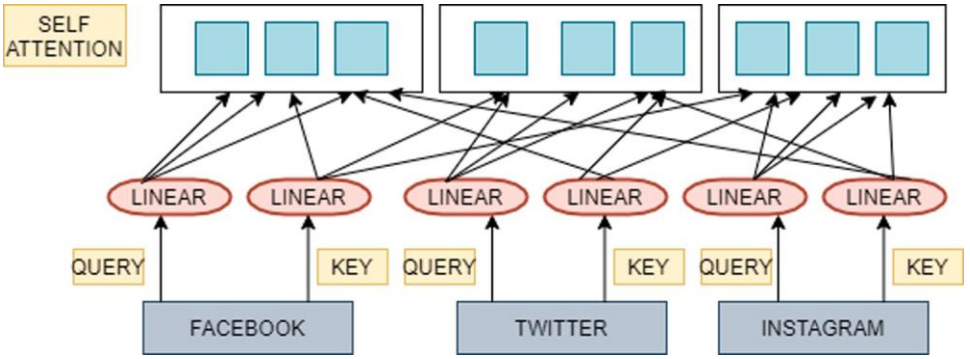

图6.自我注意。

计算文本中所有类别文件的权重。这种技术能够在一个类别的文本处理中拆分标签。假设数据中的标签不止一个。在这种情况下，术语频率和反引力矩（TF-IGM）适合于预测具有一个以上人格属性的人。术语频率和反引力矩（TF-IGM）的值是通过观察上述术语频率（TF）和反引力矩IGM的值计算出来的。TF值通过计算文档中出现的词的数量来表示词的权重。IGM的值用于计算该词在区分一个类和另一个类方面的强度。TF和IGM的计算表示如下。

$$\frac{TFT\,文件中的单词总出现次数}{总数}\ 文件中的单词 \tag{1}$$

$$= 1 + IGM_i\left(\frac{单词在文档中的总出现率\ 各类单}{词的总出现率}\right) \tag{2}$$

其中代表一个可调整的系数，TF-IGM值范围为0至1。

3.3.3. 二元分割变压器（BPT）与期频和反重力矩（TF-IGM）的融合（拟议工作）。

这项拟议的工作是基于一种监督训练技术。它结合了二进制分割变换器（BPT）和术语频率与反重力矩（TF-IGM）的特点，在快速访问中产生了更好的结果。该算法解释如下。

第1步：递归地将句子分成两部分，直到它不能被分割。这个过程被称为二进制分区。**第2步：**每个分区被认为是一个图谱神经网络节点。边缘倾向于将每个节点连接到父节点和其子节点。同样，边会将叶子节点连接到其他节点。

第三步：在图中的每个节点对其邻居节点执行自我关注。

第四步：对句子本身执行注意操作，称为自我注意。为了识别一个句子中的单词之间的关系，需要自我注意。自我注意机制的目的是为了消除隐藏的表面。

第5步：二进制分割变换器（BPT）的输出是将单个单词从句子中分离出来。

第6步：在文本分类的新统计计算中处理术语频率和反重力矩（TF-IGM）。

第7步：局部和全局加权因素应决定文本中的一个词在术语加权中的权重，使用的公式是

$$w(t_m) = 1 + \lambda . igm(t_m) \quad (3)$$

而字，其中$_m$ 和 λ是可调整的权重因子系数值。系数值λ用于保持局部和全局权重之间的平衡。

第8步：为了应用公式（1和2），通过使用公式

$$w(t_m, s) = tf_{ms} \left(1 + \lambda \sum^{fml} \frac{m}{i=1} f_{ki} \cdot \right) \quad (4)$$

$$i\, w(t_m, s) = \sqrt{tf_m} \left(1 + \frac{\lambda}{\sum^{fml} i} f_{ki} \cdot \right) \quad (5)$$

其中t_m 的TF在d,$tf_{ms} > 0$。否则，$w(t_{m,s}) = 0 if tf_{ms} = 0$。频率，$f_{ki}$，和i=1,2,...,m

**第$_m$9步：**术语*频率*和反重力矩（TF-IGM）在句子中的权重是基于术语频率的局部加权系数和基于IGM的全局加权系数，即$w(t_m, s)$ $w_1(t_m, s).w_g(t_m)$，在公式（4）和（5）中表示。

第十步:术语频率和反重力矩（TF-IGM）是一个有监督的训练，因为它取决于训练文本的已知类别信息，有一个权重系数。它应该为多类文本分类术语计算，以产生更好的输出[28]。

3.4. 特征选择

特征选择对于预测社交媒体平台的个性特征至关重要。在特征提取过程中，它可以降低社交媒体数据集的高维度。为了提高人格特征模型的预测效果，减少训练时间，为了更好地理解特征以及它们与被访者特征的关系，我们实施了特征选择过程[29,30]。在这项研究工作中，我们使用了皮尔逊相关分析。皮尔逊相关被定义为两个变量之间的线性相关措施之一。这些变量被用来计算人格特征的分数和提取的特征之间的关系。对于一个变量对p，q；线性相关系数lc是由公式（6）定义的）。

$$lc = \frac{\sum_n (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{2} (p)} \sqrt{\sum_{i=1}^{2} (q)}} \quad (6)$$

其中m和n是，和样本平均值，由。

$$\bar{p} = \frac{\sum_{i=1}^{n} pp}{n} \quad (7)$$

在公式（6）和（7）中，n代表样本量。

p_i 和 q_i 代表以i为索引的单个样本；lc的值在-1和1之间，包括1。如果p和q是完全相关的，lc取1为正相关，取-1为负相关。如果p和q完全独立，则lc为零[30]。

3.5. 机器学习算法

在这项研究工作中，通过结合二元分割变换器（BPT）与术语频率和反重力矩（TF-IGM）文本特征的融合，引入了机器学习算法，以提高预测一个人的五大人格特征时的性能。每个社交媒体输入数据集都是从预先训练好的二元分割变换器（BPT）模型中取出的。这个数据集将在自我关注机制下被处理。自我关注允许模型将每个词与输入结合起来，同时还有术语频率和反重力矩（TF-IGM）模型。然后，机器学习算法提供结果，以预测人格特征并展示数据库。在这项工作中，我们使用了一个最大熵分类器。

3.5.1. 最大熵分类器(MEC)

这个MEC被用于分组和预测。此外，它还被用于部分语音标记、歧义解决和解析选择。MEC的优点是对四舍五入的错误敏感，灵活。但它的成本效益非常高。MEC可以被定义为

$$P_{ME}(c \mid d, \lambda) = exp\left[\sum_{i} \lambda_i f_i(c, d)\right] \quad (8)$$

$$\sum_{ei} exp\left[\sum_{i} \lambda_i f_i(c, d)\right] \tag{9}$$

其中，c是一个类，d是社交媒体数据，λ是矢量权重。

4. 结果和讨论

本节描述和解析了所提出的特征提取和机器学习分类（MEC）最大熵分类器的实验结果和讨论。该分析使用了来自Facebook、Twitter和Instagram的社交媒体数据集，解析了评价和使用python spaCy开源库实现。

所有的特征提取结果都是通过使用传统的算法，如语言询问和词数（LIWC）、语言线索提取的结构化方案（SPLICE）、二元分割变换器（BPT）、术语频率和反重力矩（TF-IGM），以及这项拟议的工作，即二元分割变换器（BPT）与术语频率和反重力矩（TF-IGM）的融合。特征提取的结果将使用以下几种度量技术进行评估。

1个F1分数测量

F1分数结合了平均精度和召回值，这些测量指标表示假负值和假正值。在预测人格特征时，假正值和假负值被认为是为了减少预测错误。如果预测错误，也许某人可以被安置在与他们的个性不相符的地方。

$$精度 \quad \frac{TP}{TP + FP} \tag{10}$$

$$召回 \quad \frac{TP}{TP + FN} \tag{11}$$

$$1 - \quad FScore = 2 * \frac{precision * recall}{精度 + 召回率} \tag{12}$$

1 准确度

这种准确性的测量侧重于对总的社会媒体数据集进行精确的预测，即真正的正面和真正的负面。

$$准确度 \quad \frac{TP + TN}{tp + tn + fp + fn} \tag{13}$$

在特征提取技术实现后，我们使用皮尔逊相关进行了特征选择，通过使用最大熵的机器学习算法作为Facebook、Twitter和Instagram数据集的主要分类器来预测五大人格特征模型。

4.1. 使用Facebook数据集对结果分析进行评估

表3显示了使用Facebook数据集对人格特质进行预测的结果。 表4显示了使用Facebook数据集对人格特质进行预测的结果。从这两张表格的数据集来看，Facebook数据集的结果评估显示，每个人格特征预测所形成的最高F1分数在所提出的模型中得到了较高的表现，并提高了其效率。最高的F1分数产生于开放性人格预测，F1分数为0.921，准确率为88.17%，排在第二位的是外向性特征，F1分数为0.758，准确率为78.92%。

图7显示，平均F1分数和准确率适用于所有人格特征，并适用于所有算法。在观察图7中Facebook数据集的每个人格特征的平均F1分数和准确率时，发现提议的工作（BPT+TF-IGM）的平均F1分数和准确率最高，分别为0.762和78.34%。

表3

使用Facebook数据集预测F1分数的个性特征。

算法	F1-SCORE 开放性	自觉性	外向性	认同度	神经质
LIWC	0.893	0.532	0.732	0.612	0.632
分割	0.872	0.551	0.715	0.621	0.625
二元分割变换器（BPT）	0.878	0.572	0.727	0.644	0.643

TF-IGM	0.892	0.621	0.746	0.681	0.668
拟议的BPTþTF-IGM	**0.921**	**0.692**	**0.758**	**0.706**	**0.729**

表 15
使用Facebook数据集预测人格特征的准确性。

算法	ACCURACY 开放性	自觉性	外向性	认同度	神经质
LIWC	84.87%	67.96%	73.34%	64.51%	71.34%
分割	82.11%	69.21%	72.13%	65.12%	71.15%
二元分割变换器（BPT）	82.51%	69.34%	72.67%	66.34%	70.12%
TF-IGM	85.68%	71.04%	74.96%	70.51%	73.45%
拟议的BPT+TF-IGM	88.17%	75.85%	78.92%	73.33%	77.56%

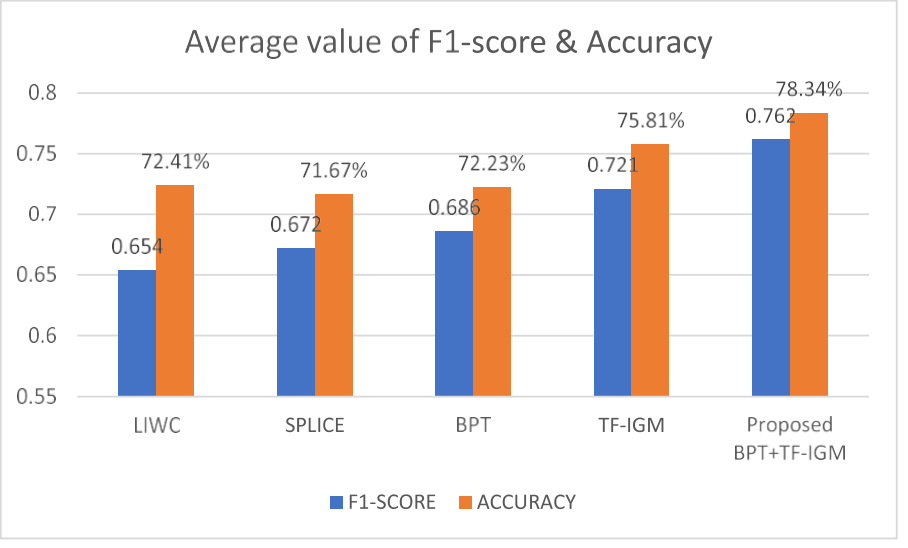

图7. Facebook数据集的平均得分。

4.2. 使用Twitter数据集对结果分析进行评估

表5显示了使用Twitter数据集对人格特征进行预测的结果。 表6显示了使用Twitter数据集对人格特征进行预测的准确性。从这两张表格的数据集来看，传统的算法和目前提出的工作（BPT + TG-IGM）的结果评估显示出最高的性能。在神经质人格特征方面，最高的F1分数为0.894，准确率为88.94%。第二高分是0.868的F1分数和81.71%的合意性人格特征的准确性。

图8显示，平均F1分数和准确率适用于所有人格特征，并适用于所有算法。在观察图8中Twitter数据集的平均F1分数和准确率时，发现提议的工作（BPT+TF-IGM）的平均F1分数和准确率最高，分别为0.783和79.67%。

4.3. 使用Instagram数据集对结果分析进行评估

表7显示了使用Instagram数据集对个性特征进行预测的结果。 表8显示了使用Instagram数据集对个性特征进行预测的准确性。从这两张表的数据集来看，传统算法和本建议工作的二元分割变换器（BPT + TG-IGM）的结果评估显示出最高的性能。在开放性人格特征方面，最高的F1-得分是0.872，准确率是87.34%。第二高分是0.783的F1分数和78.56%的神经质人格特征的准确性。

图9显示了F1分数和准确率的平均值适用于所有的人格特征并适用于所有的算法。在对图9中Instagram数据集的观察中，我们发现每个人格特征的F1分数和准确率的平均值都是最高的，即0.821和86.84%，这也是拟议的工作（BPT + TF-IGM）。

从以上分析来看，所提出的二元分割变换器（BPT）与术语频率和反重力矩（TF-IGM）的工作融合，在F1分数和准确性方面优于其他现有算法。

5. 总结

所提出的特征提取方法使用二元分割变换器(BPT)与TF-IGM方法更好地预测了人格特征。这个系统在大多数人格技术上的平均表现是，在Facebook数据集上产生了最高的平均F1分数和准确率，即0.762和78.34%，平均F1分数，和准确率，平均F1分数，和准确率是

表 16

使用Twitter数据集预测F1分数的个性特征。

算法	F1-SCORE 开放性	自觉性	外向性	认同度	神经质
LIWC	0.602	0.641	0.679	0.732	0.731
分割	0.623	0.652	0.611	0.657	0.738
二元分割变换器（BPT）	0.663	0.659	0.662	0.671	0.686
TF-IGM	0.714	0.675	0.691	0.751	0.738
拟议的BPTþTF-IGM	**0.831**	**0.821**	**0.852**	**0.868**	**0.894**

表6

使用Twitter数据集预测人格特征的准确性。

算法	ACCURACY 开放性	自觉性	外向性	认同度	神经质
LIWC	65.41%	70.47%	65.51%	78.13%	81.67%
分割	66.45%	70.65%	65.41%	77.82%	81.49%
二元分割变换器（BPT）	67.98%	71.15%	66.43%	78.62%	81.42%
TF-IGM	69.58%	73.54%	65.51%	79.15%	85.46%
拟议的BPT + TF-IGM	70.89%	75.89%	70.91%	81.71%	88.94%

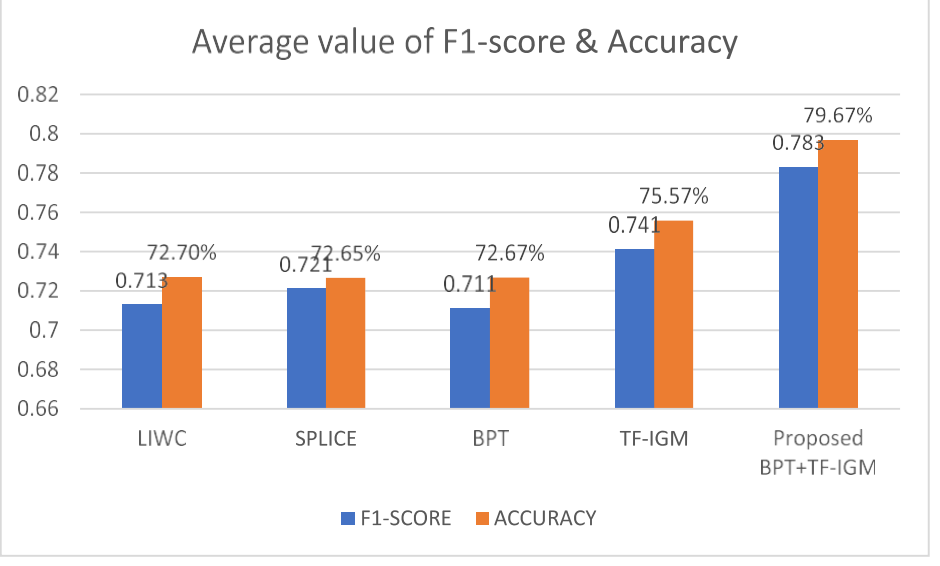

图8. Twitter数据集的平均得分。

表7

使用Instagram数据集预测F1分数的个性特征。

算法	F1-SCORE 开放性	自觉性	外向性	认同度	神经质
LIWC	0.858	0.621	0.612	0.732	0.741
分割	0.829	0.692	0.621	0.715	0.748
BPT	0.863	0.695	0.644	0.727	0.721
TF-IGM	0.839	0.657	0.681	0.746	0.738
拟议的BPTþTF-IGM	**0.872**	**0.712**	**0.695**	**0.768**	**0.783**

0.783和79.67%。在Twitter数据集上，平均F1分数和准确率，是0.821和86.84%.在Instagram数据集上。该预测系统的未来研究发展可能会利用预训练模型与高级NLP的使用。它也可能提高预测人格特征的准确性。其局限性在于，由于文本可能是转发的推文，所以文本基础不能预测其个性。基于当前的情况或趋势，文本内容可能是有影响力的。实时流媒体数据可以被使用，并与面部表情相结合

表 17
来预测个性，以完善它。这个特殊的项目可以

表8

使用Instagram数据集预测人格特征的准确性。

算法ACCURACY	开放性	自觉性	外向性	认同度	神经质
LIWC	81.67%	67.96%	73.34%	64.51%	71.34%
分割	80.49%	69.21%	72.13%	65.12%	71.15%
BPT	79.42%	69.34%	72.67%	66.34%	70.12%
TF-IGM	85.46%	71.04%	74.96%	70.51%	73.45%
拟议的BPT+TF-IGM	87.34%	75.85%	77.92%	71.33%	78.56%

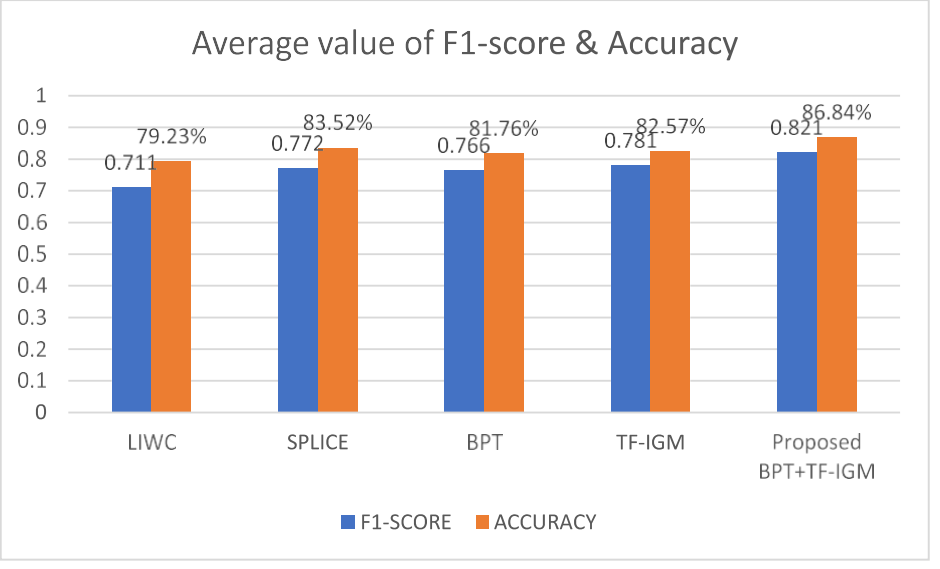

图9. Instagram数据集的平均得分。

在面试过程中有所帮助。当一个公司雇用一个人从事需要耐心的管理工作时，至少有一种性格会在推文中得到一定程度的反映（经理、看守员等），这可能有助于找到一个合适的人。

竞争性利益声明

作者声明，没有利益冲突。

参考文献

[1] Liu K, Bellet A, Sha F. Similarity learning for high-dimensional sparse data. Artificial intelligence and statistics. PMLR; 2015. p. 653-62.

[2] Farnadi G, Zoghbi S, Moens MF, De Cock M. How well do your Facebook status updates express your personality?In: belgian-dutch机器学习年会（BENELEARN）第22期会议记录。BNVKI-AIABN; 2013. p. 88.

[3] Rajendran S, Mathivanan SK, Jayagopal P, Venkatasen M, Pandi T, Somanathan MS, et al. Language dialect based speech emotion recognition through deep learning techniques. Int J Speech Technol 2021; (3/2021): 1-11.

[4] Ong V, Rahmanto AD, Williem, Suhartono D. 探索从社交媒体上的文本中预测个性：文献回顾。Internetwork Indonesia 2017; 9(1):65-70.

[5] Rajendran S, Mathivanan SK, Jayagopal P, Janaki KP, Bernard BAMM, Pandy S, et al. Emphasizing privacy and security of edge intelligence with machine learning for healthcare.Int J Intell Comput Cybern 2021. https://doi.org/10.1108/IJICC-05-2021-0099.

[6] Majumder N, Poria S, Gelbukh A, Cambria E. 基于深度学习的文档建模用于从文本中检测个性。 IEEE Intell Syst 2017;32(2):74-9.

[7] Christian H, Agus MP, Suhartono D. Single document automatic text summaryization using Term Frequency-Inverse Document Frequency (TF-IDF).ComTech Comput Math Eng Appl 2016;7(4):285-94.

[8] Kumar MS, Prabhu J. Recent development in big data analytics: research perspective. Res Anthol Artif Intell Appl Secur 2021:1640-63. https://doi.org/10.4018/978-1-7998-7705-9.ch072.

[9] Bharadwaj S, Sridhar S, Choudhary R, Srinath R. 基于Myers-Briggs类型指标（MBTI）的角色特征识别--一种文本分类方法。In:2018年计算、通信和信息学进展国际会议（ICACCI）论文集。 IEEE; 2018. p. 1076-82.

[10] Gjurković M, Šnajder J. Reddit：个性预测的金矿。在。关于人们意见的计算建模的第二次研讨会论文集。个性，以及社交媒体中的情感；2018年，第87-97页。

[11] Plank B, Hovy D. Twitter上的人格特征--或如何在一周内获得1500个人格测试。In:第六届主观性、情感和社交媒体分析的计算 方法研讨会论文集；2015年。第92-8页。
[12] Pratama BY, Sarno R. 使用Naive Bayes、KNN和SVM进行基于Twitter文本的人格分类。In:2015年 数据和软件工程国际会议（ICoDSE）的论文集。 IEEE; 2015. p. 170-4.
[13] Mehta Y, Majumder N, Gelbukh A, Cambria E. 基于深度学习的个性检测的最新趋势。 Artif Intell Rev 2020; 53(4):2313-39.

[14] Fiok K, Karwowski W, Gutierrez E, Reza-Davahli M. 用现代语言模型比较句子分类的质量和速度。Appl Sci 2020;10 (10):3386.
[15] Sukhbaatar, S., Grave, E., Bojanowski, P., & Joulin, A. (2019).arXiv preprint arXiv:1905.07799.DOI10.48550/ arXiv.1905.07799.
[16] Wang M, Yu L, Zheng D, Gan Q, Gai Y, Ye Z, et al. Deep graph library: towards efficient and scalable deep learning on graphs. arXiv preprint; 2019, https://doi.org/10.48550/arXiv.1909.01315.
[17] Bharadwaj S, Sridhar S, Choudhary R, Srinath R. 基于Myers-Briggs类型指标（MBTI）的角色特征识别--一种文本分类方法。In:2018年计算、通信和信息学进展国际会议（ICACCI）论文集。 IEEE; 2018. p. 1076-82.
[18] Chaudhary S, Singh R, Hasan ST, Kaur MI.不同分类器对Myers-Brigg性格预测模型的比较研究。Linguist anal 2013:21. https://doi.org/10.1109/ACCESS.2021.3121137.
[19] Kaur P, Gosain A. 通过结合类不平衡问题，比较类不平衡学习的过度采样和欠采样方法的行为。 基于ICT的创新. 新加坡: Springer; 2018. p. 23-30.
[20] Gjurković M, Šnajder J. Reddit：个性预测的金矿。在。关于人们意见的计算建模的第二次研讨会论文集。个性，以及社交媒体中的情感；2018年，第87-97页。
[21] Buraya K, Farseev A, Filchenkov A, Chua TS.争取从多个社交网络中获得用户个性分析。In:第三十一届AAAI 人工智能会议论文集；2017。
[22] Ong V, Rahmanto AD, Suhartono D, Nugroho AE, Andangsari EW, Suprayogi MN.基于印尼语Twitter信息的性格预测。In:2017年计算机科学与信息系统联盟会议（FedCSIS）论文集。 IEEE; 2017. p. 367-72.
[23] Ngatirin NR, Zainol Z, & Yoong TLC.不同分类器在自动人格预测方面的比较研究。In:2016年第六届IEEE 控制系统、计算和工程国际会议（ICSCE）论文集。IEEE; 2016. p. 435-40.
[24] Pennebaker JW, King LA.语言风格：语言使用是一种个体差异。J Pers Soc Psychol 1999; 77(6):1296.
[25] Christian H, Suhartono D, Chowanda A, Zamli KZ.使用预训练的语言模型 和模型平均化，从多个社交媒体数据源中进行基于文本的个性预测。 J Big Data 2021; 8(1):1-20.
[26] Abood N. Big five traits: a critical review. GadjahMada Int J Bus 2019; 21（2）：159-86.
[27] Ye, Z., Guo, Q., Gan, Q., Qiu, X., & Zhang, Z. (2019).Bp-transformer：通过二进制分区对长距离背景进行建模。arXiv预印本arXiv:1911.04070。https://doi.org/10.48550/arXiv.1911.04070。
[28] Chen K, Zhang Z, Long J, Zhang H. Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst Appl 2016; 66:245-60.
[29] Farnadi G, Zoghbi S, Moens MF, De Cock M. 使用Facebook状态更新识别个性特征。在。国际AAAI会议论文集 on web and social media. 7. PKP出版服务网；2013年。第14-8页。
[30] Yu L, Liu H.高维数据的特征选择：一个快速的基于相关的过滤器解决方案。In:第20届国际机器学习会议（ICML-03）论文集；2003.p.856-63.

Murari Devakannan Kamalesh先生是计算机科学与工程系的助理教授，在印度钦奈的Sathyabama大学工作。他有15年的教学经验。他的主要研究领域包括软件工程、软件质量保障和大数据。他在参考期刊上发表过一些文章。

B.Bharathi博士是在印度钦奈萨蒂亚巴马大学工作的计算机科学与工程系教授。她有15年的教学经验。她的主要研究领域包括软件工程、软件架构和大数据。她在参考杂志上发表了40多篇文章。她是许多参考期刊的审稿人，也是各种会议的顾问成员。