**基于机器学习的网络媒体热点话题预测**

**摘要**

**在当今数字时代，网络媒体已成为信息传播的关键渠道，显著影响着公众意见、社会话语和市场趋势。这些平台上的大量热门话题反映了公众兴趣的变化，迫切需要先进的方法来准确地预测。传统的内容分析方法依赖人工努力，已无法满足这些需求。本研究基于机器学习的方法，用于预测网络媒体中的新兴热点话题，通过分析大量数据，识别可能吸引广泛关注的话题的特性，进而预测话题的热门程度。该实验包括数据清洗、特征工程和模型构建与评价，通过交叉验证和网格搜索进行参数调整，以提高预测准确性。本文对比多种模型在性能以及耗时两个维度的优劣势，确定最终模型对网络媒体热点话题检测的良好效果。**

**关键词**

**机器学习，热点话题预测，数据分析，特征工程**

1. **引言**

在当今数字化时代，网络媒体的发展使得话题信息的生成和传播速度及其体量达到了前所未有的水平。这些海量的信息不仅丰富了人们的信息来源，也为分析公共情绪、预测社会趋势提供了宝贵的资源。然而，面对快速涌现的数据，传统的内容分析技术显得力不从心。这些传统方法往往依赖于人工操作，不仅效率低下，而且在处理大规模数据时易出错。

本研究提出了一种基于机器学习的方法，目标是通过技术手段通过大量数据来识别并预测网络媒体中可能引发广泛关注的热点话题，这些热点话题通常能够反映出公众的关注点和社会的关切，因此，通过机器学习模型来预测这些话题，不仅可以提高预测的准确性和效率，还可以为决策者和媒体工作者提供实时的数据支持，帮助他们更好地理解和引导公众舆论。

1. **相关工作**

“网络热点话题预测”是一个结合了计算机科学、数据科学和社会科学的方法和理论。随着互联网和社交媒体的迅猛发展，人们在网络上生成和分享的信息量剧增，这使得能够及时捕捉和分析这些信息的需求变得更加迫切。网络热点话题预测不仅有助于媒体和内容创建者更好地理解公众兴趣，还可以对市场动向、政治事件甚至灾害响应等方面产生深远的影响。该领域的核心任务包括但不限于：话题检测、趋势分析以及影响力建模等。

网络媒体的数据特点之一是海量，该特点要求研究方法要能处理大规模数据。为了实现这些任务，研究者通常会利用各种机器学习模型和算法，如分类算法、聚类技术和深度学习网络，来自动化分析和处理大规模的数据集。这些模型能够从大量数据中提取有用的特征，识别模式，并预测未来的趋势。本文旨在通过特征工程手段以及研究不同的模型，尝试stacking手段，对大量网络媒体信息进行较为快速准确的预测。

1. **实验过程**

**3.1实验数据集描述**

原始数据为一个38541\*60维度的csv文件，每个样例共有59个特征描述文本的情况，最后一个‘shares’为数据标签。特征涵盖了文章发布日期与数据集中最早文章日期之间的时间差、标题中的词语数量、文章正文中的词语数量、正文中不重复词语的比例、文章中非停用词的比例等，部分特征情况如下图所示:

|  |  |
| --- | --- |
| 特征名称 | 特征含义 |
| timedelta  n\_tokens\_title  n\_tokens\_content  n\_unique\_tokens  n\_non\_stop\_words  n\_non\_stop\_unique\_tokens  num\_hrefs  num\_self\_hrefs  num\_imgs  num\_videos  average\_token\_length  num\_keywords  data\_channel\_is\_lifestyle  data\_channel\_is\_entertainment  data\_channel\_is\_bus  ··· | 文章发布日期与数据集中最早文章日期之间的时间差  标题中的词语数量  文章正文中的词语数量  正文中不重复词语的比例  文章中非停用词的比例  文章中非停用词的不重复比例  文章中链接的数量  文章中指向同一网站内其他页面的链接数量  文章中包含的图片数量  文章中包含的视频数量  文章中所有词的平均长度  在文章的元数据中标记的关键词数量  文章是否属于生活方式类别  文章是否属于娱乐类别  文章是否属于商业类别  ··· |

**3.2数据预处理**

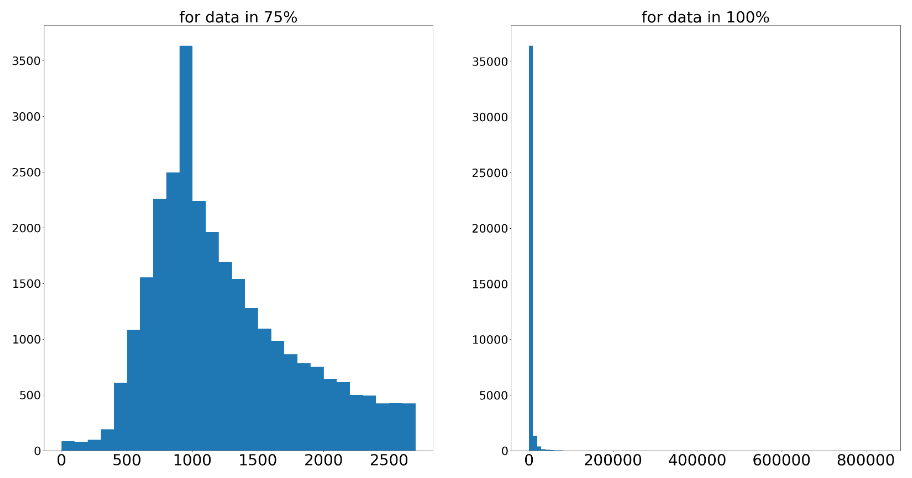
数据预处理包含去除不相关特征、缺失值处理、去除重复数据、标签转化以及特征数据转化五部分。

**去除不相关特征：**通过特征含义可知，特征‘timedelta’与本任务无关，判断其在之后的分类过程中无重要意义，所以将其去除。

**缺失值处理：**缺失值处理对于保证数据分析和机器学习模型的准确性和完整性至关重要。通过确保数据的一致性和完整性，帮助避免分析偏差，从而提高决策和预测的质量。我们通过统计数据缺失值发现数据缺失为0，所以该数据集质量较好，无需再做缺失值处理的后续步骤。

**去除重复数据：**去除重复数据有助于保持数据集的准确性和可靠性，提高数据处理效率，避免分析结果出现偏误，确保分析和模型训练的效果更符合实际情况。我们通过统计数据重复值发现数据重复为0，无需再做重复值处理的后续步骤。

**标签转化：**统计数据，发现‘shares’标签大部分处在0~2500之间，如下图所示，而1400可以作为分类标准来表示“热门”与否。



**特征数据转化：**特征数据转化是为了确保不同尺度的特征在模型训练中具有相同的影响力，从而提高机器学习算法的性能和收敛速度。此操作还有助于防止模型在训练过程中对某些特征过度敏感，实现更稳定和公平的预测结果。观察数据得知，一些特征数据数值相比于其他数据较大，需要进行特征数据转化。所以我们通过对数函数将数值缩小到与其他大致相同范围，需要转化的特征如下表：

|  |
| --- |
| 特征名称 |
| n\_tokens\_content  kw\_max\_min  kw\_avg\_min  kw\_min\_max  kw\_max\_max  kw\_avg\_max  kw\_max\_avg  kw\_avg\_avg  self\_reference\_min\_shares  self\_reference\_max\_shares  self\_reference\_avg\_sharess |

**One-hot 编码**：One-hot编码是将分类变量转换为机器学习模型可以理解的数值格式的一种方法，它通过为每个类别创建一个独立的二进制变量来避免数值大小带来的误导。这种编码方式有助于模型正确解释无序的分类数据，提高模型的预测准确性。但同时观察到部分本需要进行one-hot向量编码的特征已经在原数据集中分开，所以无需后续处理步骤。

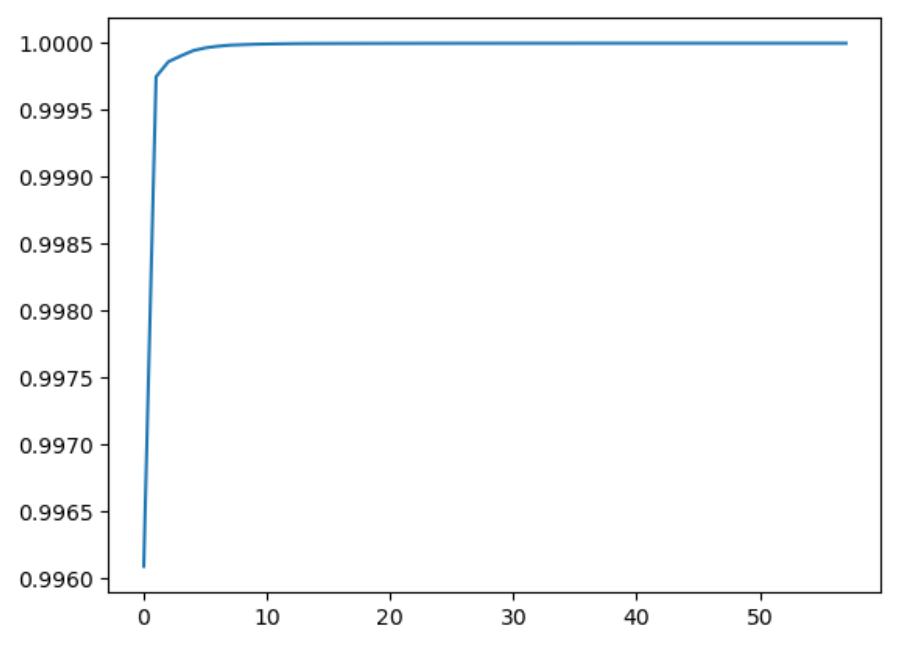
**3.3 特征工程**

特征工程是机器学习中提升模型表现的关键步骤，它通过选择、修改和创建新特征来增强数据的表达能力。这一过程有助于更好地捕捉数据间的复杂关系，从而显著提高模型的准确性和效果。我们本实验的特征工程主要包括特征生成以及特征选择。

在进行特征生成前，我们先把‘shares’标签分出，再将数据集和标签以8:2的方式分为训练集和测试集，以避免在后续过程中泄露测试集信息。

**1）特征生成**

PCA（主成分分析）是一种统计技术，用于降低数据的维度同时尽量保留原始数据的变异性。它通过找出数据中的主成分（即方向上的最大方差），并将数据投影到这些成分上，从而实现维度的减少。这使得PCA非常适用于去除数据中的噪声和冗余，同时突出最重要的特征。除了维度减少，PCA还可以进行特征生成，通过融合原数据以及处理过后数据来增加后续模型对数据的预测能力。我们首先通过PCA方法进行新特征的生成，并且添加到原数据中，并且通过逻辑回归进行初步判断生成特征的质量。下图是sklearn.PCA中explained\_variance\_ratio\_属性的图像，它代表降维后的各主成分的方差值占总方差值的比例（其和等于1），这个比例越大，则越是重要的主成分。



**2）特征选择**

由于生成后的总特征维数达到了100以上，对于分类器的时间损耗以及分类准确度都有负面影响，所以通过随机森林模型来选择重要特征。在随机森林中有feature\_importance\_属性来输出各个特征的重要程度，所以我们将输出的大于0.01的特征按照重要程度排序，并且选取前50个作为最终模型的输入特征。最终训练集和测试集的维度如下表所示：

|  |  |
| --- | --- |
|  | 维度 |
| 训练集输入特征 | (30832, 50) |
| 测试集输入特征 | (7709, 50) |
| 训练集标签 | (30832,) |
| 测试集标签 | (7709,) |

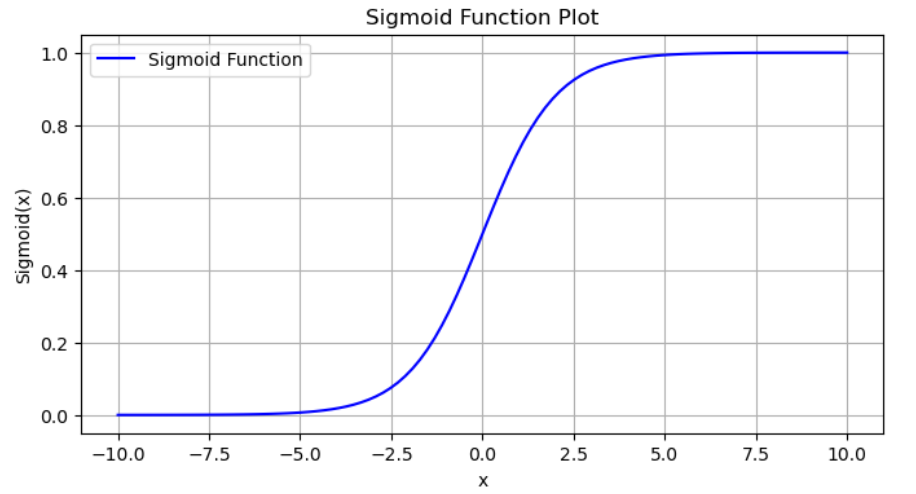
**3.3网络媒体热点话题预测模型**

1）逻辑回归模型

逻辑回归是一种多变量分析方法，主要研究二分类或多分类的因变量（观察结果）与一个或多个影响因素（自变量）之间的关系。这种模型属于概率型非线性回归，通过对传统线性回归模型的输出应用Sigmoid函数进行转换，将其映射到[0,1]区间，从而用于分类。线性回归模型的一般形式如下：

其中，是要学习的参数，是对应的是特征。

Sigmoid函数定义为，其图像是S形曲线。在这种映射下，Sigmoid函数的输出可以被解释为概率，例如当的值超过0时，，迅速接近1，从而可以通过阈值（通常是0.5）来判定分类。



2）决策树模型

决策树是一种简单但强大的分类和回归方法。每棵决策树都是一个由多个决策节点组成的树形结构，每个节点都涉及对某个特征的测试。决策树的构建过程如下：

* **特征选择：**选择一个最适合分类的特征，基于信息增益、信息增益比或基尼指数等标准。
* **树分裂：**基于选定的特征将数据集分割成两个或多个子集，这一过程在每个节点重复进行。
* **递归构建：**对每个生成的子集使用相同的方法递归构建决策树，直到满足停止条件（如节点纯度达到一定程度、达到预设的最大深度、或节点中的样本数量低于阈值）。
* **剪枝：**为防止过拟合，对构建好的树进行剪枝，移除对最终预测不具有统计意义的节点。

3）随机森林模型

随机森林是由多棵决策树构成的一个集成学习模型，主要用于分类和回归任务。其核心思想是通过构建多个决策树并将它们的结果综合起来（通常是通过投票机制或平均）来提高预测的准确性和模型的稳定性。具体来说，随机森林的构建过程包括以下几个步骤：

* **样本随机抽样：**对原始数据集进行多次抽样，每次抽样可能包括重复选择某些样本（自助抽样），为每棵树生成不同的训练数据。
* **特征随机选择：**在每个决策点，不是考虑所有可能的特征来分裂节点，而是随机选择一部分特征。这种随机性的引入减少了模型的方差，提高了泛化能力。
* **构建决策树：**使用上述随机选出的样本和特征，构建一棵决策树。树的构建一直进行，直到达到某个停止条件（如节点下的样本数量少于最小分裂数或达到最大深度）。
* **聚合结果：**将所有决策树的预测结果综合起来。在分类任务中，通常采用投票机制；在回归任务中，采用平均预测结果。

4）stacking模型

Stacking（堆叠泛化）是一种集成多个不同预测模型的方法，目的是通过一个二级模型来综合不同模型的预测结果，从而得到更准确的预测。Stacking通常涉及以下步骤：

* **基模型训练：**训练多个不同的模型（如随机森林、支持向量机、逻辑回归等）。
* **创建新特征集：**使用上述模型的预测结果作为新的特征集，这些预测结果来自对整个训练集或通过交叉验证得到的部分训练集的预测。
* **训练元模型：**在新的特征集上训练一个元模型。这个模型的任务是学习如何最优地结合各个基模型的预测结果。
* **最终预测：**使用元模型进行最终的预测。这种方法的优势在于它可以捕捉到基模型间不同的、有用的信息，并通过元模型进行有效整合。

**3.4模型评价标准**

本文将采用精度（accuracy）、精确率（precision）、召回率（recall）、F1分数（F1 score）以及受试者工作特征曲线下面积（AUC）作为模型评价的主要指标。

**精度（Accuracy）：**精度是分类任务中最常用的性能度量标准，定义为分类正确的样本数与总样本数的比例。然而，当正负样本极度不均衡时，单纯的精度并不是一个很好的度量指标。例如，在点击率预测的场景中，如果未点击用户与点击用户的比例为1:10，则即使模型将所有用户都预测为未点击，精度也能达到90%。为了解决这一问题，本文在数据预处理阶段已进行了正负样本均衡处理，使得精度能够作为一个有效的评价标准。

**精确率（Precision）：**精确率是指在所有被模型预测为正类的样本中，真正为正类的样本所占的比例。精确率反映了模型在预测正类时的准确性。

**召回率（Recall）：**召回率是指在所有真实正类样本中，被模型正确预测为正类的样本所占的比例。召回率衡量了模型捕捉正类样本的能力。

**F1分数（F1 Score）：**F1分数是精确率和召回率的调和平均，是一个综合反映模型精确率和召回率的指标。当模型的精确率和召回率都高时，F1分数也会高。

**受试者工作特征曲线下面积（AUC）：**AUC 是评价二分类模型性能的一个重要指标，通过测量模型将正类预测为正类（真正例率）和将负类错误预测为正类（假正例率）的能力。真正例率是竖直方向的度量，假正例率是水平方向的度量。ROC 曲线是以假正例率为横轴、真正例率为纵轴绘制的曲线，AUC 表示此曲线下的面积。AUC 值越大，表示模型性能越好，理想值为1，而0.5表示效果无异于随机猜测。

以上评价指标共同构成了一个全面的模型性能评价体系，能够综合反映模型在不同方面的表现。

1. **实验结果**

**4.1不同特征组合对热点话题检测的影响**

选择合适的特征对于模型预测结果有重要影响。本文将热点话题数据集中的特征分为原始特征，原始特征，原始特征加上PCA处理组合特征，随机森林筛选过后特征并依次对比不同特征下模型的效果。

首先将原始特征应用于训练模型，接着加入PCA处理组合特征，再将经过随机森林筛选的特征输入模型进行预测。实验模型选择工业界最常见的逻辑回归，计算模型在训练集与验证集上的准确率指标表现结果。实验结果如下表所示，结果保留四位有效数字。

|  |  |  |
| --- | --- | --- |
| 特征集 | 训练集ACC | 测试集ACC |
| 原始特征 | **0.6104** | **0.6055** |
| 原始特征+PCA处理组合特征 | **0.6139** | **0.6070** |
| 随机森林筛选过后特征 | 0.6225 | 0.6157 |

从实验结果中可知，单独使用原始特征时模型的能取到良好的效果，ACC能够达到0.61。当加入PCA处理后的特征并融合之后，模型的性能有所提升。然后使用随机森林筛选过后ACC进一步提升。

**4.2不同机器学习模型对比分析**

对比了不同机器学习模型应用于热点话题预测时的性能如下所示，对于逻辑回归、决策树与随机森林三个模型采用了网格搜索对模型的超参数进行调节，并且调节过程中使用五折交叉验证来寻找最优参数组合使得模型ACC表现结果最佳。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型类别 | 训练集准确率 | 训练集AUC | 测试集准确率 | 测试集AUC |
| 逻辑回归 | 0.6226 | 0.6637 | 0.6155 | 0.6588 |
| 决策树 | 0.6567 | 0.7142 | 0.6444 | 0.6944 |
| 随机森林 | 0.7752 | 0.8629 | 0.6601 | 0.7283 |

从表可以看出来三个模型性能从高到低依次为：随机森林、决策树、逻辑回归。同时实验还添加了随机模型来进行对比，AUC为0.5116，ACC为0.5088，低于任何一个模型。

同时实验还测得了模型auc、精确率和召回率的数值：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模型类别 | 训练集精确率 | 训练集召回率 | 训练集F1分数 | 测试集精确率 | 测试集召回率 | 测试集F1分数 |
| 逻辑回归 | 0.6284 | 0.7198 | 0.6710 | 0.6171 | 0.7163 | 0.6631 |
| 决策树 | 0.6793 | 0.6782 | 0.6788 | 0.6600 | 0.6728 | 0.6664 |
| 随机森林 | 0.7697 | 0.8270 | 0.7973 | 0.6591 | 0.7375 | 0.6961 |

三种机器学习模型在运行耗时上也有差别，将各个模型在训练集的总耗时以相对于逻辑回归模型耗时的比值展示，结果如下表所示：

|  |  |  |
| --- | --- | --- |
| 机器学习模型 | 10折平均训练耗时（秒） | 相对耗时 |
| 逻辑回归 | 0.359 | 1.00 |
| 决策树 | 1.17 | 3.26 |
| 随机森林 | 56.36 | 155.98 |

从表中可以看出，随机森林模型耗时远远超出所有模型，但尚且在可接受范围内，而逻辑回归和决策树的训练时间都在2秒以内，响应速度快。

接着我们使用stacking方法将三个模型结合起来，并且添加元分类器来输出最终预测结果。最终性能如下表所示：

|  |  |  |
| --- | --- | --- |
|  | 结果 | 较目前最好模型的提升度 |
| 训练集准确率 | 0.6671 | -0.1081 |
| 训练集AUC | 0.6914 | -0.1715 |
| 测试集准确率 | 0.6667 | 0.0076 |
| 测试集AUC | 0.7133 | -0.0242 |
| 测试集精确率 | 0.6657 | 0.0066 |
| 测试集召回率 | 0.7404 | 0.0029 |
| 测试集F1分数 | 0.7011 | 0.0050 |

根据数据对比，尽管当前的随机森林模型在训练集上的表现优于Stacking模型，显示出较高的训练准确率和AUC，它在测试集上的表现却略显不足，反映出可能的过拟合问题，即模型在训练数据上过于优化而在新数据上表现不佳。相比之下，Stacking模型虽然在训练集的表现略逊一筹，但在测试集上的准确率、精确率、召回率和F1分数略有提升，暗示其更好的泛化能力和对新数据的适应性。综合考虑模型的泛化能力和稳定性，选择Stacking模型可能是更优的决策，尤其是在面对未知数据时，它可能提供更为可靠的预测性能。

1. **结语**

本文从两个方面研究了热点话题预测的问题。首先，对于该研究进行合适的特征选择。特征数据能很好地反映话题的类别以及热门程度，所以特征处理尤为重要其次，针对热点话题预测任务，可首先考虑逻辑回归、决策树、随机森林三个模型并依据实际环境条件，再进行stacking融合，选择效果最好的模型。

三个模型中随机森林的预测效果最好但是模型耗时较长，存在延迟预测的风险造成预测误差，而决策树的运行耗时虽然比较短，但是模型本身存在准确度问题，逻辑回归模型虽然性能稍逊于随机森林模型及决策树模型，但是其耗时最短可解释性较强，也是个不错的模型。同时将三个模型通过stacking方法结合到一起之后的模型，对于预测效果不错，同时能够中和三者的缺点。但是总体来说时间较长，可能会对实时预测产生影响。