

**项目开发报告**

**题目：互联网虚假新闻检测**

**专 业： 计算机科学与技术**

**班 级： CS1901**

**学 号： U201914899**

**姓 名： 柏威良**

**成 绩：**

**指导教师： 李玉华**

**计算机科学与技术学院**

# 项目开发报告

## 1.1项目目的

随着互联网的飞速发展，家家户户都逐渐牵上了网线，凭借着网络的便利性，外界信息得以轻松的得知，对于百姓们熟知的新闻信息的获取也能在须臾之间完成。然而任何事物的发展都具有两面性，互联网虽说使得信息的产生和获取更加便捷，也滋生了不少虚假新闻扰乱民众的视听，倘若放任不管甚至可能导致谣言泛滥成灾，最终影响到现实社会秩序。

本项目旨在通过机器学习训练出具有一定可靠性的互联网虚假新闻检测模型，更具体一点是通过监督学习训练出一个判断新闻是否虚假或者无需判断的分类模型，从而为网络环境的精华献出一份力量。

## 1.2问题分析

互联网虚假新闻检测可以视为一个三分类监督学习问题，主要难点在于如何将新闻材料中的信息主体和评论两种文本转化为特征向量，以及筛选、过滤掉对模型训练无用的信息。为方便后续分析，可以构建如下模型：记一条新闻的新闻内容为content，全部评论为comment\_all，新闻真假为label，这样的三元组(content, comment\_all, label) 为一个训练样本，二元组的集合{(content, comment), …}为待预测的数据。

于是问题转化为：

1. 数据清理（清理无用样本，对content、comment\_all分词，停用词过滤）
2. 特征向量提取（首先训练出词向量模型然后组成文本向量）
3. 选择训练模型

④ 预测和评价

## 1.3设计与分析

### 1.3.1 数据分析

首先通过粗略地观察数据集可以发现存在部分N/A空值，这些不完整地信息肯会干扰到后续数据清理，因此要事先调用pandas库进行空值填充；文本信息中通常会夹杂着一些表情代码（通常以中括号包裹）和@字段，这些对于新闻真假的判断来说作用不大可以去掉。

### 1.3.2 算法流程设计

本项目主要分为以下几个步骤：

1. 预处理proccess；这个阶段要进行数据导入，数据清理和分词工作，此外还要生成词向量模型，再根据得到的词向量模型构建样本的特征向量。
2. 模型参数自动调优 optimize；
3. 模型训练 fit；本项目将采用多层感知机的机器学习模型，该模型是一个多分类的线性分类模型。
4. 预测 predict；
5. 分析优化 analyze；

由于中文分词、训练词向量、模型参数选择、模型训练这四个部分耗时较大，为了避免不必要的时间浪费，我将分词结果、词向量模型、模型参数、训练好的MLP模型分别保存，方便下一阶段直接使用，以实现模型的持久化。

总体的算法流程图描述如下：

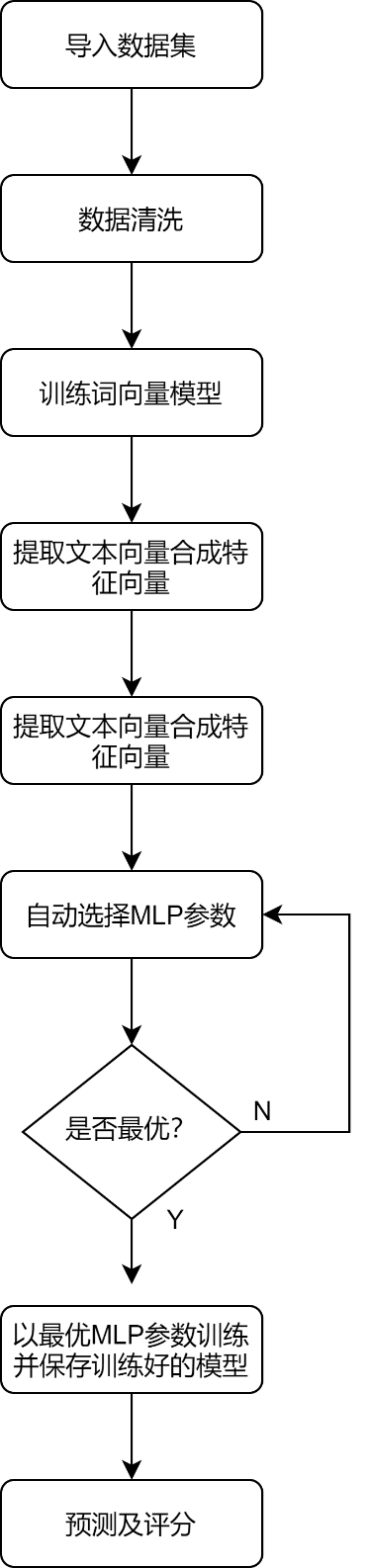


图 1 算法流程图

### 1.3.3 机器学习算法设计

新闻数据属于自然语言数据，其中的评论带有很多现实特性，往往格式混乱包含较多的脏数据，于是要想处理好它们就必须谨慎对待数据的预处理阶段，在这个阶段的数据集的导入和数据清理我将采用pandas模块，分词采用jieba中文分词模块，生成词向量模型将采用word2vec，该模型的优点在于Word2vec会考虑上下文，跟Embedding方法相比，效果要更好，[维度](https://www.zhihu.com/search?q=%E7%BB%B4%E5%BA%A6&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22article%22%2C%22sourceId%22%3A98645675%7D)更少，速度更快，用在包含评论的新闻语境中非常合适。 由词向量转化为文本向量的方法我将采用取均值的方式，即文本中每个词的词向量的各个维度取均值产生的和词向量相同维度大小的文本向量。由于一则新闻样本包括了content和comment\_all两条文本，因此特征向量可以取这两条文本向量的连接。

机器学习的模型我将采用带标签的神经网络多层感知机MLP，调用sklearn的mlp分类器进行模型训练。MLP准确率较高而且可以通过改变隐层数目来改进模型，由于数据实现采用了word2vec向量化，因此对于文本的上下文环境也有考虑。然而MLP的学习时间较长，为了节约时间成本，在自动选择参数调优的过程中可以选择较小的迭代最大次数来防止过长时间得不到最优参数。自动选择参数阶段将调用sklearn的GridSearchCV模块。

## 1.4结果分析

自动参数的结果如图2所示，预测准确度如图3所示：



图 2 自动参数



图 3 预测准确度

## 1.5思考与总结

这个项目是我巩固了机器学习的基本流程，切身体会了数据的预处理、特征提取、模型选择、模型调优、训练和预测的流程，掌握了使用python机器学习库sklearn、数据处理和表示库pandas、数学库numpy、中文分词库jieba、词向量模型word2vec的基本使用方法，为我以后的机器学习之路打下了坚实的基础。

数据预处理的部分花了很大一部分精力，主要是要考虑如何去除不必要的样本以及清理文本中的脏数据，以及如何借由词向量提取样本特征向量的算法设计。

项目还有很大的优化空间，例如说压缩特征向量、进一步过滤原始数据、选择其他机器学习模型，但是出于时间限制暂时无法实现。希望以后能够将本次项目的经验运用到工作学习生活当中去。