**小组项目报告**



**2022 至2023 学年第 一 学期**

课 程 名 深度学习与大数据智能

学生学号 20204133 20204143 20204100 20204101

20204117

学生姓名 杨一帆 涂藤耀 张胤 马睿 陈思伦

任课教师 文静

报告得分

假新闻分类

目录

[一. 分工： 2](#_Toc136293232)

[二．摘要： 2](#_Toc136293237)

[三．项目所用数据集和各个字段： 3](#_Toc136293246)

[四．项目相关研究 5](#_Toc136293247)

[五．实验方法： 6](#_Toc136293248)

[六．实验过程： 13](#_Toc136293249)

[七．结论： 24](#_Toc136293250)

[八．总结和不足: 25](#_Toc136293251)

[九．参考文献： 26](#_Toc136293252)

### 一、分工：

涂藤耀：课程答辩，PPT写作，数据预处理，深度学习模型的开发与训练(RNN、LSTM、残差网络)，指导修改项目书，汇总内容。

杨一帆：机器学习模型的开发与训练、将模型改写适配到云上训练、总结的对比各类模型的优缺点，作为组长组织调度project推进

张胤：web前端开发，各模型对比实验

马睿：前后端连接调试，机器学习可行性分析

陈思伦：撰写项目书，PPT写作，查阅文献

### 二、摘要：

本项目旨在利用循环神经网络 (RNN) 实现假新闻分类。我们将使用Kaggle的Fake and real news dataset数据集，该数据集包含了23481条“假”推文和21417条“真实”文章。我们将训练一个深度学习模型来自动判断一篇文章是否为真实新闻或假新闻。

建议的基本方法是使用RNN模型，并使用该数据集进行模型训练和评估。进阶的方法是使用LSTM或GRU这样的长短时记忆循环神经网络模型来完成假新闻分类，并且对结果进行准确度评估。能够自动分类假新闻和真实新闻，以提高公众对新闻的准确性和信任度。

### abstract：

In recent years, the rise of fake news and misinformation has become an increasingly prevalent issue, leading to significant social, political, and economic consequences. Automated systems that can accurately classify news articles as real or fake news are essential for combatting this problem. The proposed RNN model provides a powerful tool for identifying and classifying fake news, with the potential to significantly improve the accuracy and reliability of news reporting.

One of the key advantages of RNN models is their ability to analyze sequences of data, making them well-suited for processing news articles that often contain complex text and structures. Additionally, advanced variants such as LSTM or GRU models can overcome the limitations of traditional RNNs by enabling the network to selectively remember important and relevant information.

By training a deep learning model on the Fake and real news dataset, we can create a model that can accurately distinguish between real and fake news articles. The model can then be used to classify new articles that are not part of the training dataset, identifying instances where the information provided is potentially unreliable or inaccurate.

Overall, the ability to automatically classify news articles as real or fake is vital for maintaining trust in journalism, public institutions, and democracy. As the scale and impact of fake news continues to grow, the use of RNN models offers a practical solution for identifying and mitigating its impact.

**关键词：深度学习，神经网络，假新闻分类**

### 三、项目所用数据集和各个字段：

数据集各个字段与阐述：

|  |  |
| --- | --- |
| title | 新闻标题 |
| text | 新闻正文内容 |
| Subject | 新闻所属主题 |
| Date | 新闻发布日期 |

Fake数据集有23503个数据

True数据集有21418个数据

数据集类别较平衡，不存在不平衡的问题，预处理部分我们放在了实验过程中

我们观察数据集挑出两个典型新闻翻译：

True：

华盛顿（路透社）-美国国会一个保守的共和党派系的领导人本月投票支持大幅扩大国债以支付减税费用，他周日称自己是“财政保守派”，并敦促在2018年限制预算。为了与共和党人正在进行的急剧转变保持一致，美国众议员马克·梅多斯在哥伦比亚广播公司的《面对全国》节目中对联邦支出采取了强硬立场，议员们正准备在1月份就这一问题展开斗争。当他们周三度假归来时，议员们将开始试图通过一项联邦预算，这场斗争可能与移民政策等其他问题有关，尽管11月的国会竞选活动即将到来，共和党人将寻求控制国会。唐纳德·特朗普总统和他的共和党人希望大幅增加军费预算，而民主党人也希望在支持教育、科学研究、基础设施、公共卫生和环境保护的项目上按比例增加非国防“可自由支配”支出。“（特朗普）政府已经愿意说：‘我们将增加非国防可自由支配支出……约7%，’”规模较小但有影响力的众议院自由核心小组主席梅多斯在该项目上说。“现在，民主党人说这还不够，我们需要给政府加薪10%到11%。对于一个财政保守派来说，我不知道理由在哪里……最终你会耗尽别人的钱，”他说。梅多斯是12月底投票支持本党债务融资税收改革的共和党人之一，该改革预计将扩大联邦预算赤字，并在10年内为20万亿美元的国债增加约1.5万亿美元。美国民主党众议员约瑟夫·克劳利在哥伦比亚广播公司表示：“听到马克谈论财政责任很有趣。”。克劳利表示，共和党的税收法案将要求美国借款1.5万亿美元，由子孙后代偿还，为企业和富人的减税提供资金。克劳利说：“这是众议院历史上通过的最不负财政责任的法案之一。我认为我们将在未来很多年为此付出代价。”。共和党人坚持认为，这是美国30多年来最大的税收改革，将促进经济和就业增长。众议院议长保罗·瑞安也支持该税收法案，他最近比梅多斯走得更远，在一次电台采访中明确表示，福利或该党经常称之为“福利改革”，将是共和党2018年的首要任务。用共和党人的话说，“福利”计划是指为老年人、穷人和残疾人提供的食品券、住房援助、医疗保险和医疗补助医疗保险，以及华盛顿为帮助穷人而制定的其他计划。民主党人抓住了瑞安12月初的言论，称这表明共和党人将试图通过削减社会项目的支出来为他们的税收改革买单。但众议院共和党人的目标可能不得不退居参议院，参议院需要一些民主党人的投票才能批准预算并防止政府关门。民主党人将利用他们在参议院的影响力，捍卫可自由支配的非国防计划和社会支出，同时解决“梦想家”的问题，即儿童时期被非法带到美国的人。9月，特朗普将儿童入境暂缓行动（DACA）计划的截止日期定为2018年3月，该计划保护年轻移民不被驱逐出境，并为他们提供工作许可。总统在最近的推特消息中表示，他希望为拟议的墨西哥边境墙和其他移民法修改提供资金，以换取同意帮助梦想家。众议员黛比·丁格尔告诉哥伦比亚广播公司，她不赞成将这个问题与其他政策目标联系起来，比如隔离墙资金。“我们需要清理DACA，”她说。周三，特朗普的助手将与国会领导人会面，讨论这些问题。白宫表示，随后将于1月6日和7日为特朗普和共和党领导人举行周末战略会议。特朗普还计划于周日会见佛罗里达州共和党州长里克·斯科特，后者希望获得更多紧急援助。在佛罗里达州、得克萨斯州和波多黎各发生飓风，加利福尼亚州发生野火后，众议院通过了810亿美元的援助计划。该方案远远超过了特朗普政府要求的440亿美元。参议院尚未就援助进行投票。

Fake:

唐纳德·特朗普就是不能只祝所有美国人新年快乐。相反，他不得不向他的敌人、仇恨者和非常不诚实的假新闻媒体大声疾呼。这位前真人秀明星只有一份工作要做，但他做不到。“随着我们的国家迅速变得更强大、更聪明，我想祝我所有的朋友、支持者、敌人、仇恨者，甚至是非常不诚实的假新闻媒体，新年快乐、健康，”愤怒的裤子总统在推特上写道。2018年对美国来说将是伟大的一年！随着我们的国家迅速变得更强大、更聪明，我想祝我所有的朋友、支持者、敌人、仇恨者，甚至是非常不诚实的假新闻媒体，新年快乐、健康。2018年对美国来说将是伟大的一年！唐纳德·J·特朗普（@realDonaldTrump）2017年12月31日，特朗普的推特如你所料地火爆。什么样的总统会像这种卑鄙、琐碎、幼稚的胡言乱语一样表达新年问候？只有特朗普！他缺乏体面，甚至不允许他在阴沟里呆上足够长的时间来祝美国公民新年快乐！塔尔伯特·斯旺主教（@TalbertS旺）2017年12月31日没有人喜欢你加尔文（@calvinstowell）2017年11月31日你的弹劾将使2018年成为美国伟大的一年，但我也会接受重新控制国会。Miranda Yaver（@mirandayaver）2017年12月31日你听到自己说话了吗？当你不得不包括那么多讨厌你的人时，你不得不想知道？为什么他们都恨我？Alan Sandoval（@AlanSandoval13）2017年12月31日谁在新年愿望中使用了“仇恨者”一词？？Marlene（@marlene399）2017年12月31日你不能只说新年快乐吗？Koren pollitt（@Korencarpenter）2017年12月31日这是特朗普2016年新年前夕的推文。祝所有人新年快乐，包括我的许多敌人和那些与我作战并输得如此惨重的人。爱！唐纳德·J·特朗普（@realDonaldTrump）2016年12月31日这对特朗普来说并不是什么新鲜事。多年来，他一直在这样做。特朗普在新年、复活节、感恩节和9/11周年纪念日向他的敌人和仇恨者发出信息。pic.twitter.com/4FPAe2KypA Daniel Dale（@ddale8）2017年12月31日特朗普的节日推文显然不是总统推文。在成为总统之前，他在霍尔马克工作了多久？史蒂文·古丁（@SGoodine）2017年12月31日他一直都是这样。唯一不同的是，在过去的几年里，他的滤镜一直在崩溃。Roy Schulze（@thbthttt）2017年12月31日除了青少年之外，谁还会使用“仇恨者”一词？Wendy（@WendyWhistles）2017年12月31日他是一个他妈的5岁的孩子谁知道（@rainyday80）2017年11月31日所以，对于所有投票支持这个洞的人来说，他们认为一旦他掌权，他就会改变，你错了！70岁的男人不会改变，现在他大了一岁。

通过观察可以发现：真新闻的客观事实多，用词得当，而假新闻的夸张用词多，而且短句和情感类句子多。紧接着我们再进行对文本的预处理。

### 四、项目相关研究：

假新闻分类是一个热门的研究领域，在过去的几年里已经有很多相关的研究工作被发表。以下是一些与该项目相关的已发表工作：

1. “Detecting and Tracking Fake News” (2018)：该论文提出了一种用于检测和跟踪假新闻的方法。该方法基于逻辑回归模型，将特定的文本特征输入到模型中。然后，该模型会对文本进行分类，以确定它们是否是假新闻。
2. “A Survey on Detection Methods of News Spoofing Based on Natural Language Processing Techniques” (2019): 该论文提供了一个综合的调查，对基于自然语言处理技术的新闻欺诈检测方法进行了总结。该文章列举了一些常用的方法，包括逻辑回归、支持向量机、决策树和深度学习等。
3. “A Multi-Channel Convolutional Neural Network for Fake News Detection” (2020): 该论文描述了一种多通道卷积神经网络（CNN）的方法，用于假新闻检测。该模型包括3个卷积层和1个全连接层，同时将文本和元数据视为不同的通道，进行融合，以实现更准确的分类。

假新闻分类的意义非常重大。随着社交媒体及互联网的普及，假新闻、虚假信息的传播日益猖獗，这对社会、政治、经济等方面都产生了重要影响。特别在重大事件发生时，假新闻的传播甚至可能导致公众产生恐慌，对社会稳定产生威胁，而且还会严重影响公众对媒体的信任度。

对于媒体机构，能够自动对新闻进行分类意味着可以更好地提供准确、可靠的新闻报道，从而加强对公众的信任。对于政府机构，假新闻关乎国家的形象和安全，因此能够识别和对抗假新闻对于维护国家利益非常重要。

此外，在社交媒体及互联网平台上，假新闻的传播速度非常快，因此快速准确地识别和分类假新闻至关重要，可防止假新闻通过互联网传播，降低假新闻的传播影响和造成的社会危害。

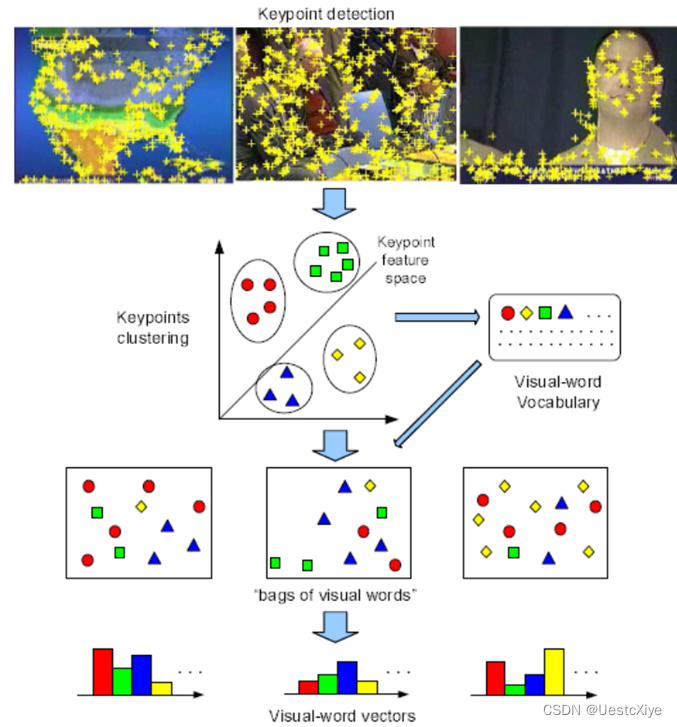
相比这些现有的工作，本项目将重点放在使用循环神经网络(RNN)来解决假新闻分类的问题。而且，项目还探索了使用LSTM或GRU这样的长短时记忆循环神经网络模型来实现假新闻分类任务。与其他方法相比，这些模型是高度可拓展的，并且可以更好地处理序列数据。因此，项目将深入研究这些模型，并将尝试在识别假新闻方面取得更好的表现。

### 五、实验方法：

1：词袋模型（bag-of-words model）：

是一种文本表示方法。它将一段文本表示为一个包含所有单词的向量。词袋模型不考虑单词之间的顺序和语法规则，只考虑不同单词在该段文本中的出现频率。因此，它非常适合用于文本分类、情感分析、信息检索等自然语言处理任务。

在词袋模型中，每个单词被看作是向量中的一个维度。向量的每个维度表示一个单词，当该单词在文本中出现时，该维度的值加上1。例如，对于一段文本"this is a sample text"，它可以表示为一个向量[1, 1, 1, 1, 0, …, 0]，其中第一维对应单词"this"，第二维对应单词"is"，以此类推。该向量的长度等于词汇表的大小，即包含了所有可能出现的单词。



词袋模型的优点是简单易用，能够快速获取文本中的关键信息，且不受文本长度的限制。但是，它也有一些缺点，比如无法处理单词的语义信息、无法考虑单词的顺序等。因此，在实际应用中，词袋模型通常需要和其他技术一起使用，以获得更准确的文本表示。

#### 2：TF-IDF模型（term frequency-inverse document frequency model）：

是一种用于表示文本中单词重要性的技术，常被应用于文本挖掘、信息检索等领域。TF-IDF模型基于单词的词频和在文本集中的出现情况来计算单词的重要性。

TF-IDF模型中的TF指的是单词在某篇文本中出现的频率。IDF指的是逆文档频率，即对于一个单词，它在整个文本集中出现的频率越低，则其IDF值越大，表示这个单词的重要性越高。

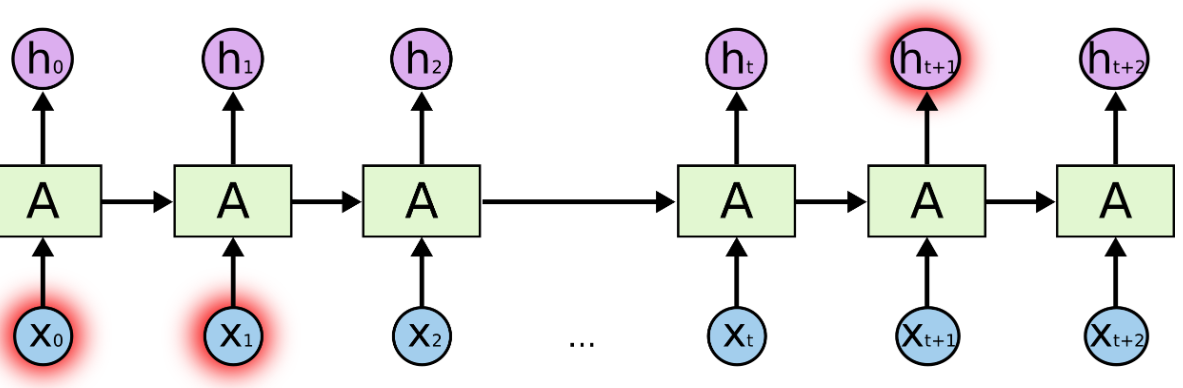
TF-IDF模型的计算步骤分为两步：第一步是计算每个单词在每篇文本中的TF值，即该单词在该篇文本中出现的次数除以文本中单词的总数；第二步是计算每个单词在整个文本集中的IDF值，即用总文本数除以包含该单词的文本数，再取对数。最终，将TF和IDF相乘，得到该单词在该篇文本中的TF-IDF值。

TF-IDF模型可以用于提取文本中的重要词汇，并基于这些重要词汇进行文本分类、聚类、推荐等操作。由于它能够计算每个单词在文本集中的重要性，因此在信息检索等领域被广泛应用。

#### 3：循环神经网络（Recurrent Neural Network，RNN）：

是一种常用的神经网络结构，主要用于处理序列数据，如自然语言、音频、视频等。

RNN中的每个节点都有一个内部状态（记忆单元），用来保存之前的信息，并将其传递到下一个节点。在处理序列数据时，RNN每次接收一个输入，并根据输入内容和前一个节点的状态来计算输出值和新的状态，将输出值传递到下一个节点。这种递归计算能够考虑上下文信息，使得RNN在处理序列数据时有良好的表现。



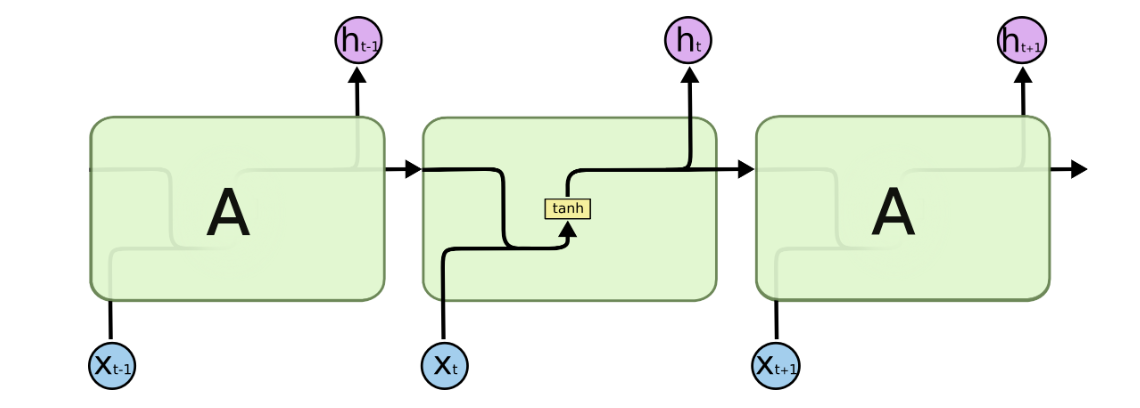
#### 4：长短时记忆网络（Long Short-Term Memory，LSTM）：

是一种循环神经网络（Recurrent Neural Network，RNN）的变体。与传统的RNN不同，LSTM在记忆单元的设计上采用了门控机制，能够更好地避免长期依赖问题，从而有效地捕捉序列中的长期依赖关系。

LSTM的基本结构包括三个门：输入门、遗忘门和输出门，以及一个记忆单元。其中，输入门控制哪些信息可以进入记忆单元，遗忘门控制哪些信息可以在记忆单元中保留，而输出门控制哪些信息可以输出。LSTM的记忆单元能够有效地处理序列中长期依赖的问题，使得LSTM在处理序列数据时具有很好的表现。

LSTM在自然语言处理、时间序列预测、语音识别等领域得到了广泛的应用。例如，在语音识别中，LSTM可以将音频序列作为输入，逐步地解码成文本输出；在机器翻译中，LSTM可以将源语言序列转换为目标语言序列。

LSTM的一个变体是门控循环单元（Gated Recurrent Unit，GRU），它是一种简化版的LSTM，并尝试减少LSTM中的一些计算量，但仍然能够有效地解决序列数据中的长期依赖问题，同时在一些应用场景中也有优异的表现。

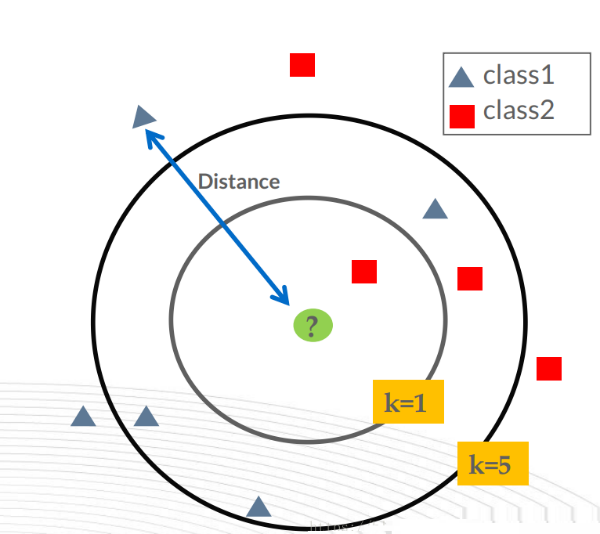


#### 5：k最近邻算法（K-Nearest Neighbor，KNN）：

是一种常见的机器学习算法，主要用于分类和回归问题。KNN算法的核心思想是找到最近的k个邻居来进行分类或回归。

在分类问题中，KNN算法会根据样本之间的距离来确定最近的k个邻居，并根据它们所属的类别来确定待分类样本所属的类别。当K=1时，就是最简单的最近邻算法。

在回归问题中，KNN算法会根据样本之间的距离来确定最近的k个邻居，并使用它们的值来预测待预测样本的值。



KNN算法的优点在于简单易用，不需要进行模型的建立和训练，同时也具有较好的分类效果。然而，KNN算法也有一些缺点，比如需要大量的计算，当数据集较大时，计算代价较高；并且对于高维数据，KNN算法的效果往往不如其他算法。

在实际应用中，KNN算法可以与其他算法结合使用，以提高分类或回归的准确率，并且在一些特定的场景下，如图像分类、自然语言处理等方面，KNN算法仍然是一种有效的技术手段。

#### 6：朴素贝叶斯（Naive Bayes）：

是一种基于概率论和贝叶斯定理的分类算法。它的核心思想是根据已知类别（标签）的特征和频率，来计算未知样本分别属于各个类别的概率，进而选择概率最大的类别作为预测结果。

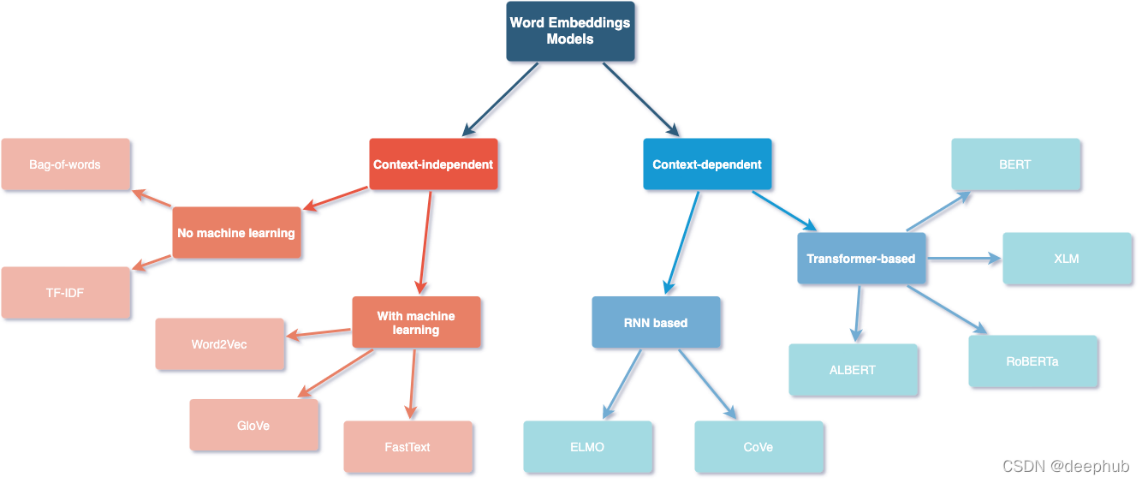
朴素贝叶斯的“朴素”指的是它的条件独立假设，即假设每个特征在分类标签确定的情况下是独立的，这样可以简化计算，并使其能够在小样本情况下也能取得较好的效果。在具体实现中，朴素贝叶斯可以分为三种类型：高斯型（Gaussian Naive Bayes）、多项式型（Multinomial Naive Bayes）和伯努利型（Bernoulli Naive Bayes），分别适用于不同类型的特征。

#### 

朴素贝叶斯的优点在于简单易懂、快速、可扩展性强、对小规模数据表现优秀等，同时还能够进行分类概率估计，并能够处理多分类问题。在文本分类、垃圾邮件过滤、个性化推荐等领域，朴素贝叶斯都具有一定的应用价值。不过，朴素贝叶斯的缺点是对于输入数据的质量和先验概率的选择较为敏感，而且它所假设的特征之间是相互独立的，在某些情况下可能会导致分类效果不佳。

#### 7：词嵌入模型：

是一种将自然语言转换为连续向量表示的技术，在自然语言处理中得到了广泛的应用。它的主要目的是将单词或短语映射到一个低维向量空间中，以捕捉词语之间的语义关系和上下文的信息，从而提高自然语言处理模型的性能。



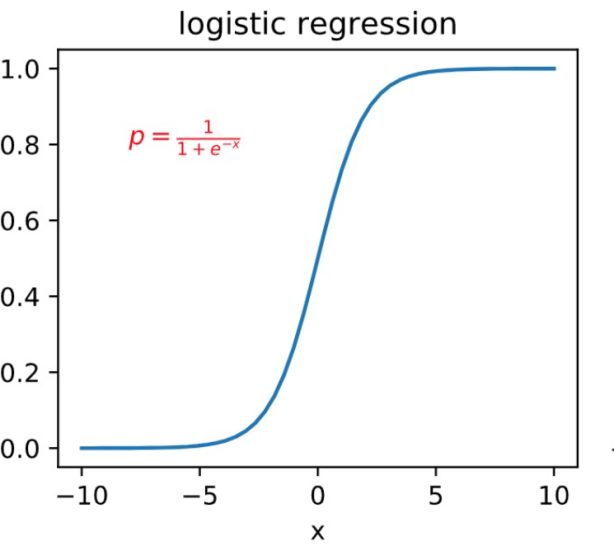
在实践中，比较常用的词嵌入模型是基于神经网络的模型，如Word2Vec、GloVe和FastText等。其中，Word2Vec是一种基于神经网络的词向量模型，它通过学习单词在上下文中的分布来捕捉单词的语义信息。GloVe（Global Vectors）是一种利用全局词汇统计信息来学习单词向量的模型，它将Word2Vec中的局部上下文信息和全局词汇信息结合起来进行训练。而FastText则是一种基于字符级别的词向量模型，它将单词表示为其子单词的向量平均值，可以更好地处理稀有单词和拼写错误的情况。

词嵌入模型在自然语言处理的许多任务中都得到了广泛的应用，如文本分类、情感分析、命名实体识别、语言模型等。通过对文本数据进行词嵌入，能够将较为稀疏的离散式的单词表示转换为稠密的连续式的向量表示，从而便于基于此进一步进行特征提取和模型训练。同时，词嵌入还可用于计算词语之间的相似度、实现单词推荐等应用，具有很强的实用价值。

#### 8：逻辑回归（Logistic Regression）：

是一种机器学习算法，用于进行二分类或多分类问题预测。它使用一条直线或曲线来分割不同类别的数据，而这条直线或曲线被称为“决策边界”。

逻辑回归的原理是基于sigmoid函数，将原始的特征值输入给该函数后，可以将特征值映射到0到1之间，便于计算概率进行分类。该函数的数学表达式为：



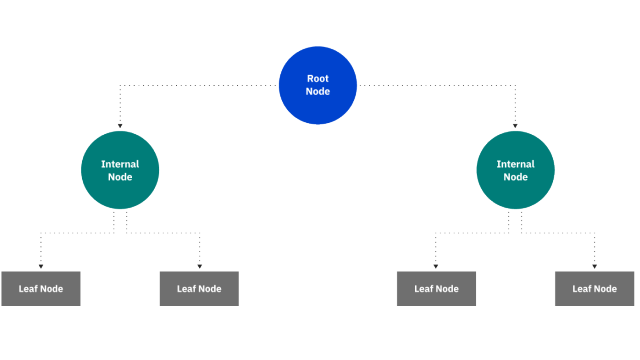
其中，z 为输入的特征值和参数的线性组合，g(z)为sigmoid函数的输出值，g(z)越接近1代表该样本属于正例的可能性越大，越接近0代表该样本属于负例的可能性越大。对于二分类问题，逻辑回归的决策边界可以表示为一个阈值，比如当 g(z)=0.5 时，该样本判定为正例，否则为负例。

逻辑回归常用的优化算法是梯度下降算法，通过不断迭代更新模型参数，直至目标函数达到最小值。梯度下降的目标是求出使误差（即预测值与真实值之间的差距）最小的参数值，以便使模型预测更加准确。

逻辑回归在实践中被广泛应用于二分类或多分类问题，如情感分析、垃圾邮件过滤、医学诊断和金融风险预测等领域。它具有模型简单、易于解释、易于实现、计算速度快等优点。但是，逻辑回归的表达能力有限，很难对非线性问题进行分类。

#### 9:决策树（Decision Tree）：

是一种基于树结构的机器学习算法，用于进行分类和回归分析。它通过对数据集的分析和学习，以树形结构的方式表示出决策规则，并可以用于预测新样本的类别或数值。



决策树的构建过程包含以下步骤：

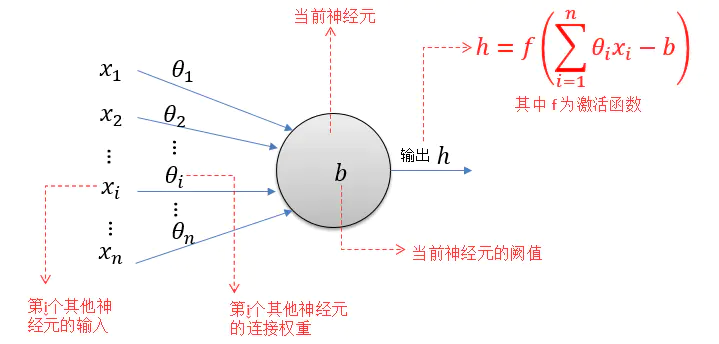
1. 特征选择：从所有特征中选出一个最优特征，用于对数据集进行划分。
2. 数据集划分：依据上一步中选择的最优特征，将数据集划分为多个子集。
3. 构建子树： 对于每个子集，重复以上两个步骤进行递归，直到子集中所有样本属于同一个类别或者满足预先设定的停止条件。
4. 剪枝：在构建好整棵决策树后，进行剪枝处理，以防止该模型过度拟合，并达到更好的泛化能力。

决策树的优点包括易于理解、可解释性强，能够同时处理离散型数据和连续型数据，并且能够对缺失属性值进行处理。此外，决策树也具有较高的准确性和相对较快的训练速度。

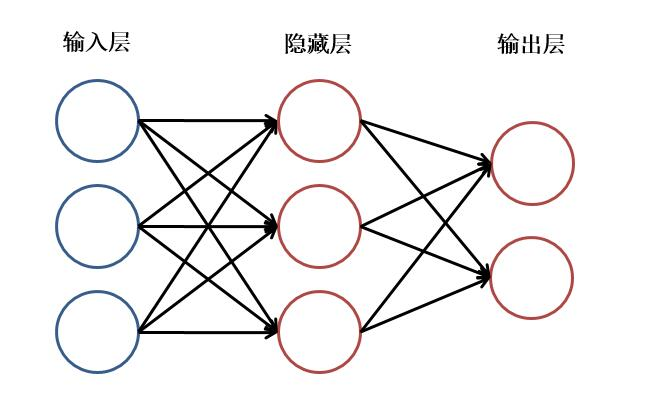
决策树的应用非常广泛，包括数据挖掘、机器学习、统计学和人工智能等领域。它可以用于分析数据中包含的规律和模式，用于预测股票价格、信用风险、疾病诊断、客户分类等不同的场景

#### 10：神经网络（Neural Network）：

是模拟人类神经元网络结构的一种计算模型，用于处理和解决各种类型的机器学习问题。它由许多个简单的神经元节点组成，每个神经元节点接收输入，经过加权求和和激活函数的处理后，将计算结果与其他节点连接并向前传递，直到达到输出层产生预测值或结果。



神经网络通常包括输入层、隐藏层和输出层。



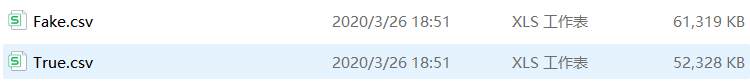
输入层接收原始数据作为输入，隐藏层和输出层由一些神经元节点组成，隐藏层用于提取和表示数据的特征，最终输出层给出了一个或多个预测的结果。神经网络的训练过程就是对网络的参数进行调整，使网络的预测结果与真实值之间的误差最小化，常用的优化算法包括反向传播算法和随机梯度下降。

神经网络可以用于各种机器学习任务，例如模式识别、分类、回归、聚类和自然语言处理等。它在图像识别、语音识别、自然语言处理和机器翻译等领域中具有非常出色的性能，并且在许多其他计算机科学应用领域中也有广泛的应用。尤其是在大规模数据和智能系统的开发中，神经网络的应用变得越来越流行。

### 六、实验过程：

#### 5.1 数据预处理：

一共有两个数据集，一共有两个数据集，一个True.csv，一个Fake.csv



数据集中分别有4个字段，新闻标题，新闻内容，新闻主题，新闻日期

新闻内容通常包含了相比于其他字段最多，最丰富和最直接的信息，可以提供关于新闻类别的重要特征。

接下来是进行文本的预处理操作



1：读取文本的text字段，使用正则表达式清洗掉文本中非字母数字字符和非空白字符，减少噪声。

2：将文本转换为小写形式。统一文本的大小写，避免因为大小写不一致而导致的词汇重复或匹配错误。

3：词性统一，将所有词的词性都转化为动词，例如：running->run，减少词向量的维度，减少噪声。

4：对于假新闻，文本类别标注为0，对于真新闻，文本类别标注为1

#### 5.2 机器学习模型：

我们做机器学习模型的目的是为了探索最佳的数据预处理方式。

机器学习模型我们采用的是将文本数据转换为TF-IDF特征表示形式，并设置最大特征数量为5000，TF-IDF可以凸显在当前文档中频繁出现但在整个语料库中较为罕见的词，从而更好地反映词的重要性。

我们使用了几个常见的机器学习模型来训练该分类任务（逻辑回归，朴素贝叶斯，K邻近，神经网络（前馈神经网络（包含100个神经元的单隐层），决策树模型）来分类结果，并用准确率判断好坏。



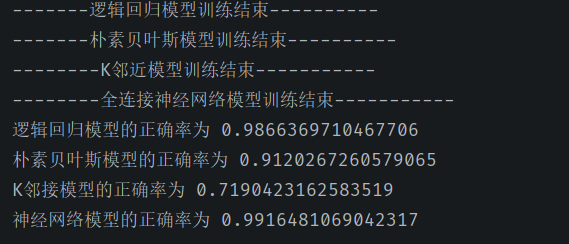


实验过程中，我们在使用朴素贝叶斯算法时，发现正确率低下，我们认为原因是若数据特征之间存在依赖性，或是特征之间并非真正的独立，则会导致朴素贝叶斯算法的正确率降低。这是因为朴素贝叶斯算法的基本假设是特征之间相互独立，当这个假设不成立时，就会导致算法的错误决策。

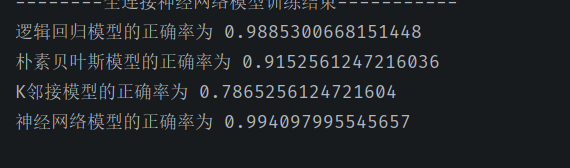
所以我们深度学习模型选择了词嵌入模型。

在考虑预处理过程时：

1：删除停用词后的正确率如下：



2：不删除停用词后正确率如下：



一开始我们数据预处理的时候是删除了停用词的，即删除语料中一些类似于is a the的词，但是后来发现不删除停用词，各个模型的正确率均得到了一定的提高，看来删除停用词的操作是有争议性的，对于有些语料，删除停用词会使正确率得到提高，而有些语料，似乎语境和一些逻辑是比较重要的信息，删除了停用词反而会导致模型学习不到一些知识。所以我们需要多实验多尝试对比实验结果。

如下是我们多次对比的结果：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 删除停用词的TFIDF | 不删除停用词的TFIDF | 删除停用词的词袋模型 | 不删除停用词的词袋模型 |
| 逻辑回归模型 | 0.98660 | 0.98853 | 0.96423 | 0.97211 |
| 朴素贝叶斯模型 | 0.91202 | 0.91525 | 0.90344 | 0.91001 |
| K邻接模型 | 0.71904 | 0.78655 | 0.70102 | 0.73655 |
| 神经网络模型 | 0.99164 | 0.99409 | 0.98012 | 0.99121 |

#### 5.4 深度学习模型：

选择词嵌入层作为我们所有深度学习模型的第一层，在深度学习中，词嵌入是更常用的表示方式。词嵌入通过将每个单词映射到一个低维连续向量空间中的向量来表示单词，使得具有相似含义的单词在向量空间中更加接近。词嵌入可以通过学习单词的分布式表示来捕捉单词之间的语义关系和上下文信息，从而提供更丰富和有意义的表示。

对于二分类问题，binary\_crossentropy是更常用和合适的选择。

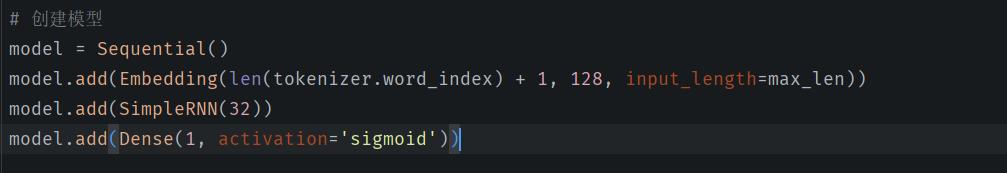
1：使用适合问题类型的损失函数可以更好地反映问题的特性，提高模型的性能和效果。

2：在二分类问题中，只需计算一个类别的损失，相比多分类问题需要计算多个类别的损失，节省了计算资源和时间。

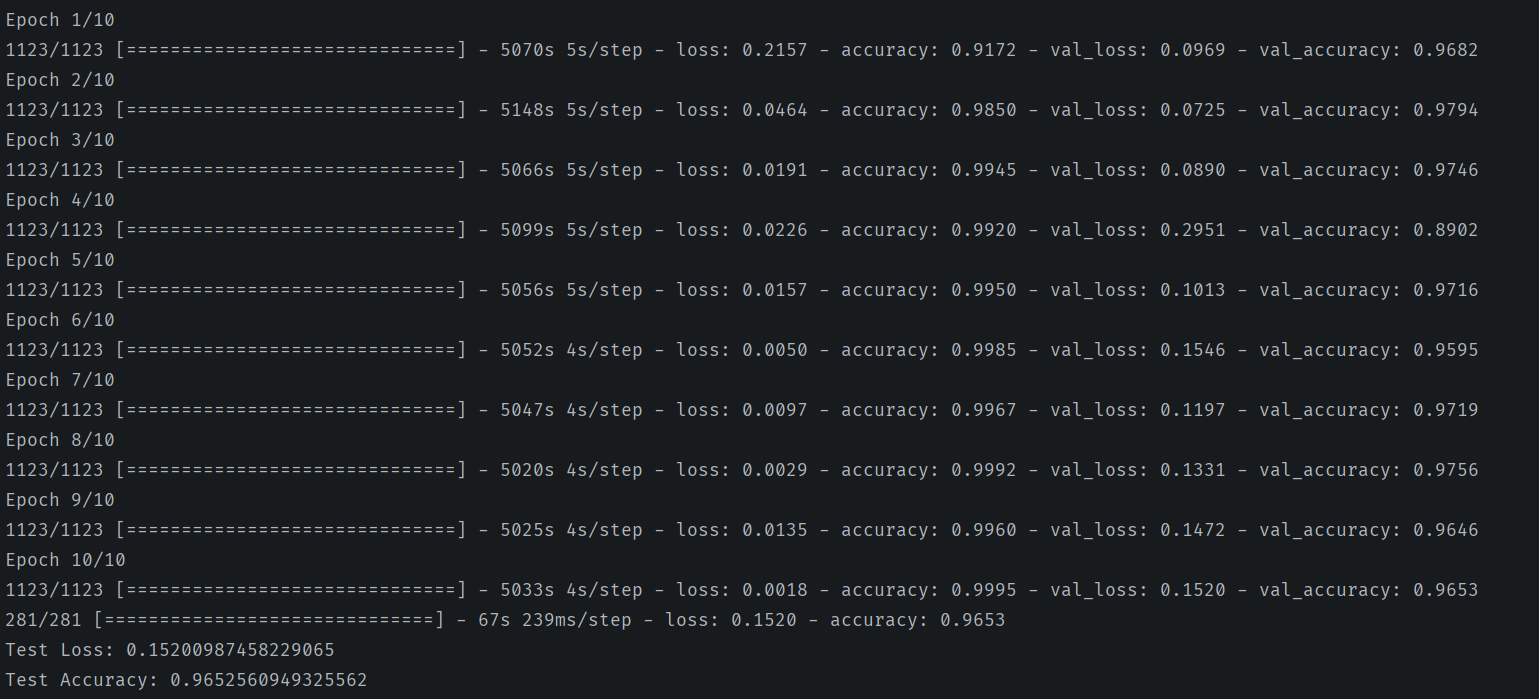
3：在数值计算上，binary\_crossentropy通常更稳定。它不会受到类别数量的影响，避免了出现数值上的不稳定性，如分母为0等问题。

4：binary\_crossentropy的损失函数对模型输出的解释更直接。它度量了模型对于二分类问题中两个类别之间的区分度和准确性，更加直观。

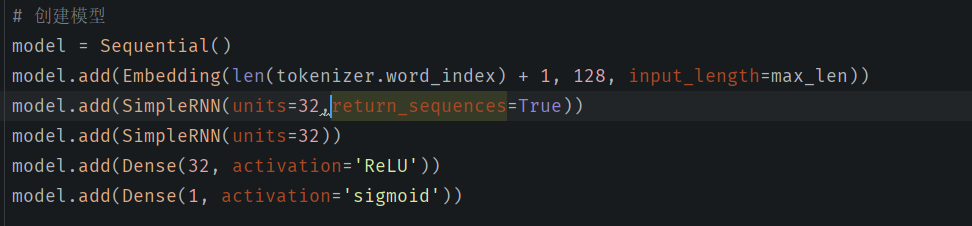
1：单层RNN模型：



我们建立了一个简单的RNN模型来做训练，在迭代10次后正确率达到96.525%。



2：双层RNN模型：



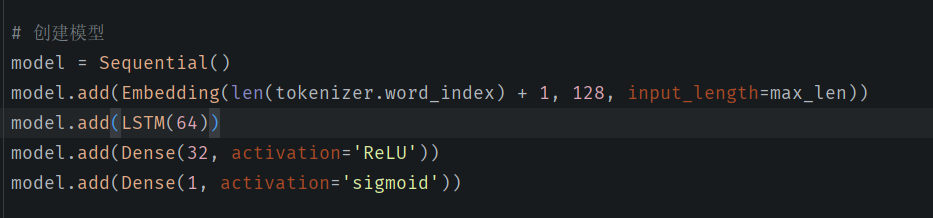
一开始我们采用了教材上的RNN结构，即两层RNN后面再添加一个全连接层。该层具有32个神经元，并使用ReLU激活函数。末尾添加一个全连接层。该层具有一个神经元，用于二分类问题，并使用sigmoid激活函数将输出限制在0到1之间。



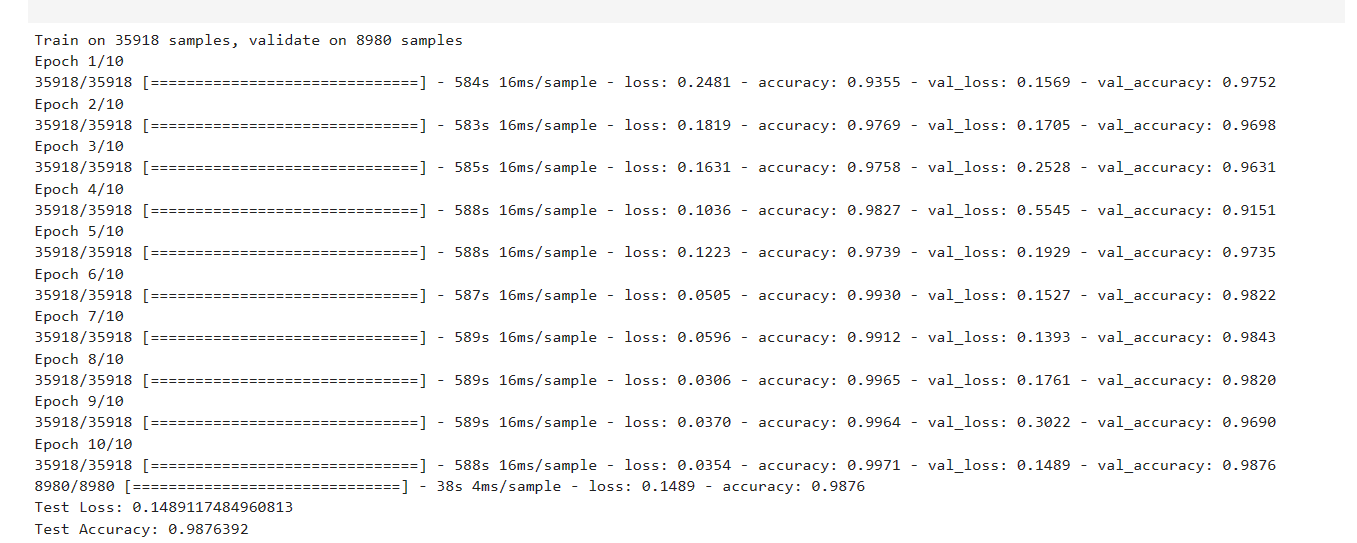
我们发现难以收敛，而且训练速度极慢

多个SimpleRNN层叠加时，梯度消失和梯度爆炸问题可能会出现。这意味着在反向传播过程中，梯度可能会变得非常小或非常大，导致参数更新不稳定。这可能是因为SimpleRNN层在时间步的展开中，会导致梯度的连乘或连加，从而使梯度变得非常小或非常大。

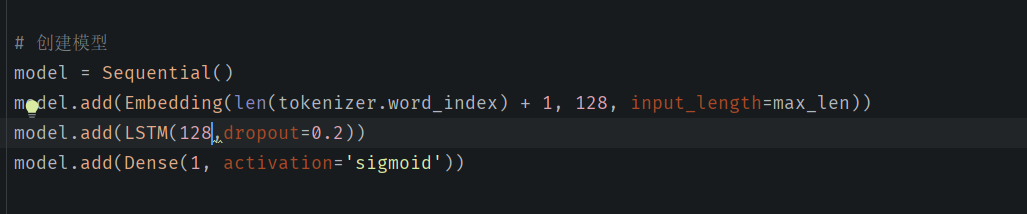
3：LSTM模型：（带一个全连接层）：

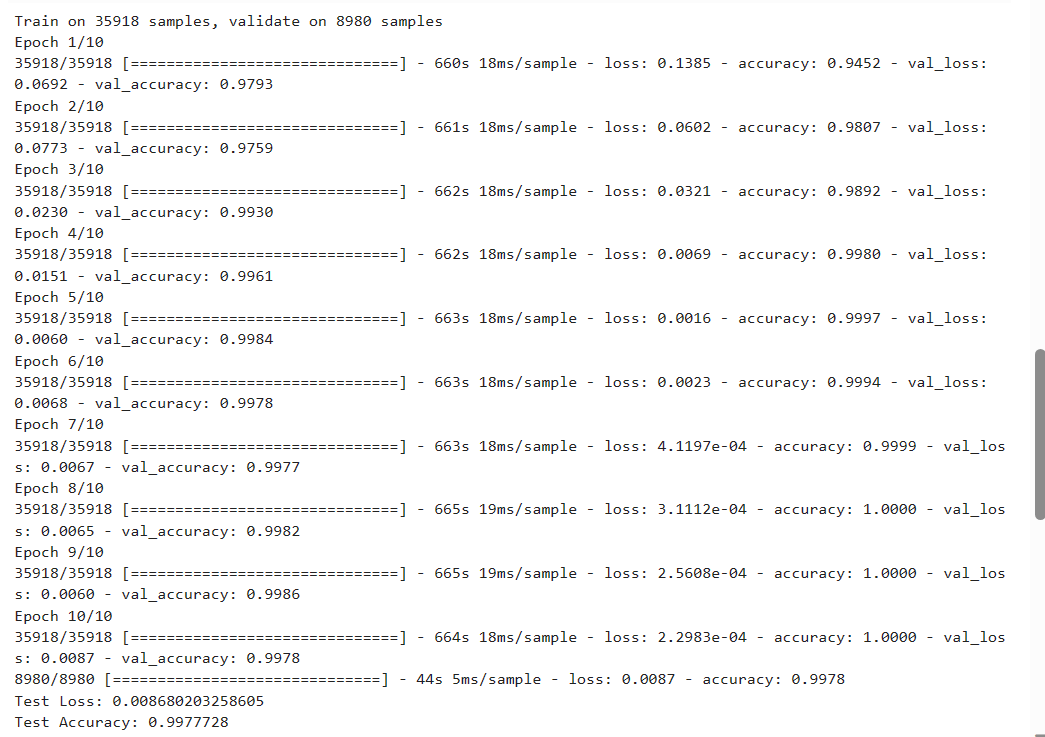


我们使用了LSTM，10次迭代后正确率为98.763%左右，至少收敛了而且正确率超过部分机器学习模型，但是也出现了训练过程不稳定的问题，我认为是层数过深，LSTM连接一个全连接层导致的。参数过多。而且全连接层导致了LSTM的信息丢失。



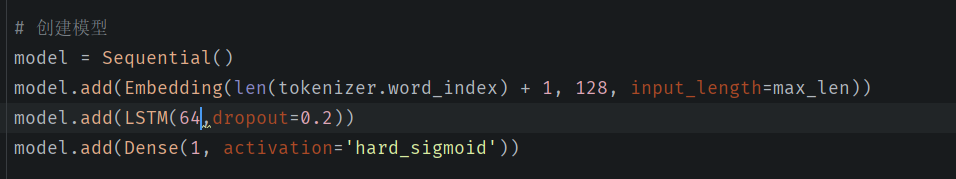
4：LSTM模型：（不带全连接层）：



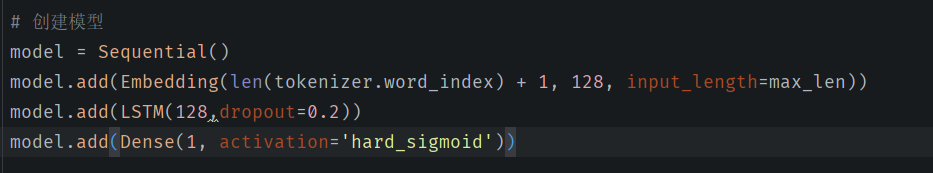


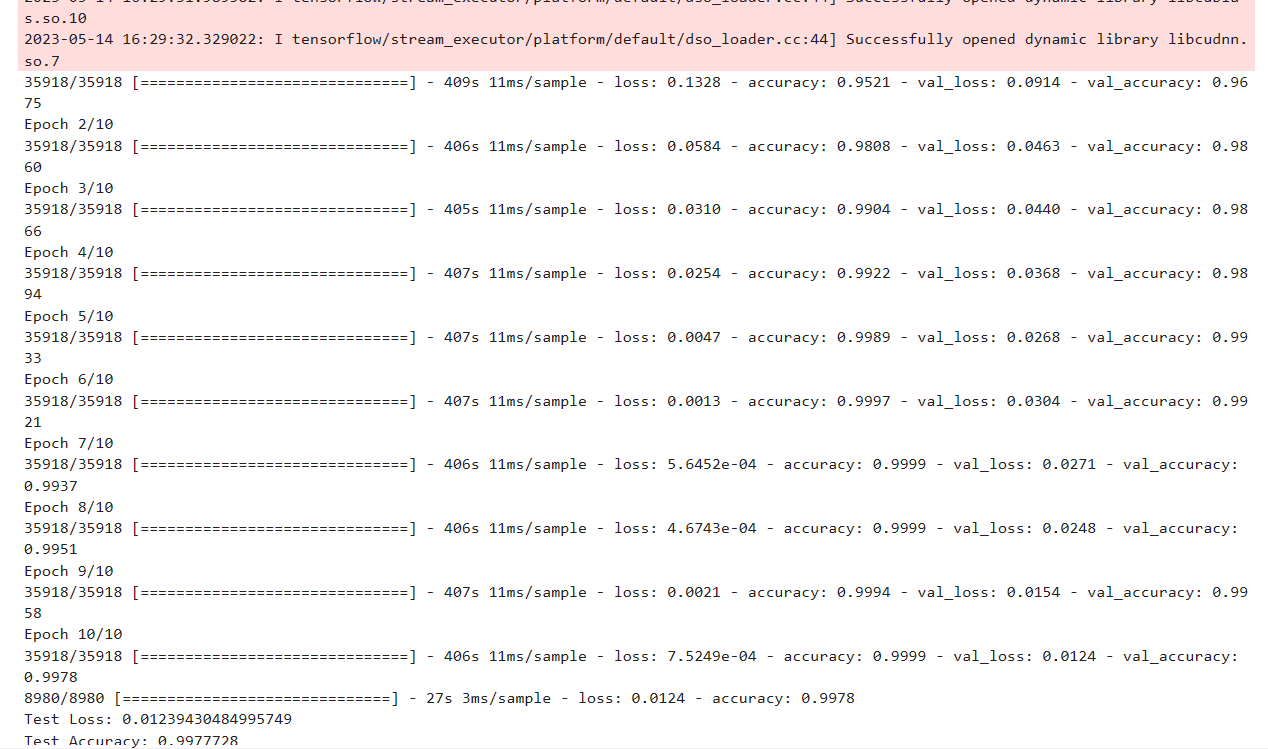
5：LSTM模型：（不带全连接层，使用hard-sigmoid）：

后面尝试使用老师在课上提到的hard-sigmoid进行输出，减小训练时间，但是依然收敛很慢，不过的确训练时间上降低了很多。我们经过多次尝试认为是网络容量的问题，hard-sigmoid输出的值在一些情况下不足以表示某些模式或特征



我们通过提高LSTM单元数量增加网络容量，最后做出来的确收敛，正确率达到99.77%，相比于之前的99.86%略有降低，但是训练时间缩短了，模型计算量更小，所以我们综合考虑多个因素，选择了这种结构的模型作为最终的模型





5.5 前端web及后端设计:

代码部分：



这里我们实现了一个针对新闻真假分类的JavaScript的前端界面，它通过获取用户输入的新闻内容并将其发送给后端进行真假判断

具体实现如下：

* 加载 axios 库，用于发送 POST 请求到后端
* 在 HTML 中创建一个包含输入框和按钮的表单，当用户点击按钮时会调用 handleClick 函数
* handleClick 函数会从输入框中获取用户输入的文本，然后使用 axios 库将其提交到后端路由 ‘/predict’
* 如果后端返回的预测结果 prediction 为 0，则在页面上显示 这是假新闻。如果 prediction 为 1，则在页面上显示 这是真新闻。
* 如果发送 POST 请求失败，则控制台会打印出错误

后端：

from flask import Flask, render\_template, request, jsonify

import re

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.linear\_model import LogisticRegression

app = Flask(\_\_name\_\_)

@app.route('/')

def home():

    return render\_template('index.html')

@app.route('/predict', methods=['POST'])

def predict():

    text = request.form['text']

    text = re.sub(r'[^\w\s]', '', text)

    text = text.lower()

    input\_vector = vectorizer.transform([text]).toarray()

    prediction = model.predict(input\_vector)[0]

    return jsonify({'prediction': prediction})

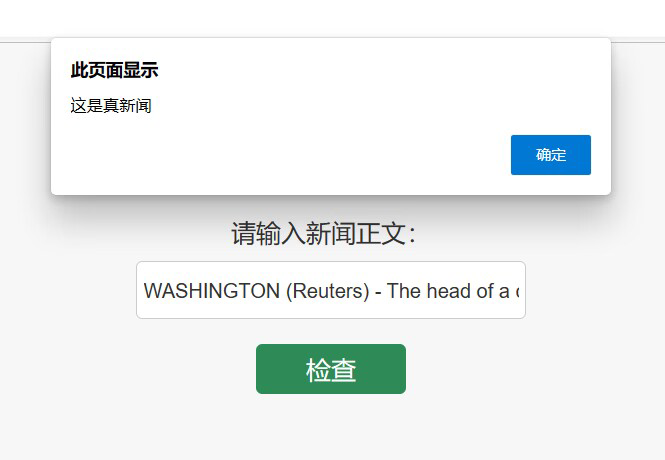
if \_\_name\_\_ == '\_\_main\_\_':

    app.run(debug=True)

这里实现了与前端的数据交互：

* 首先加载Flask模块：Flask、render\_template、request 和 jsonify，并实例化一个Flask应用。
* 定义’/’路由，调用 render\_template 函数将前端页面模板 index.html 渲染显示
* 定义’/predict’路由，使用 request.form 获取前端页面发送的 text 参数，对其进行相应处理后使用model进行真假新闻分类预测
* 使用 jsonify 函数将预测结果以 JSON 格式返回给前端页面

效果图：



### 七、结论：

对机器学习模型而言

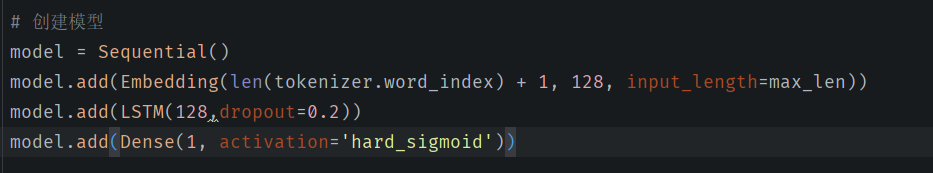
（TFIDF,max-feature=5000,正常清洗数据，词性统一，不删除停用词）

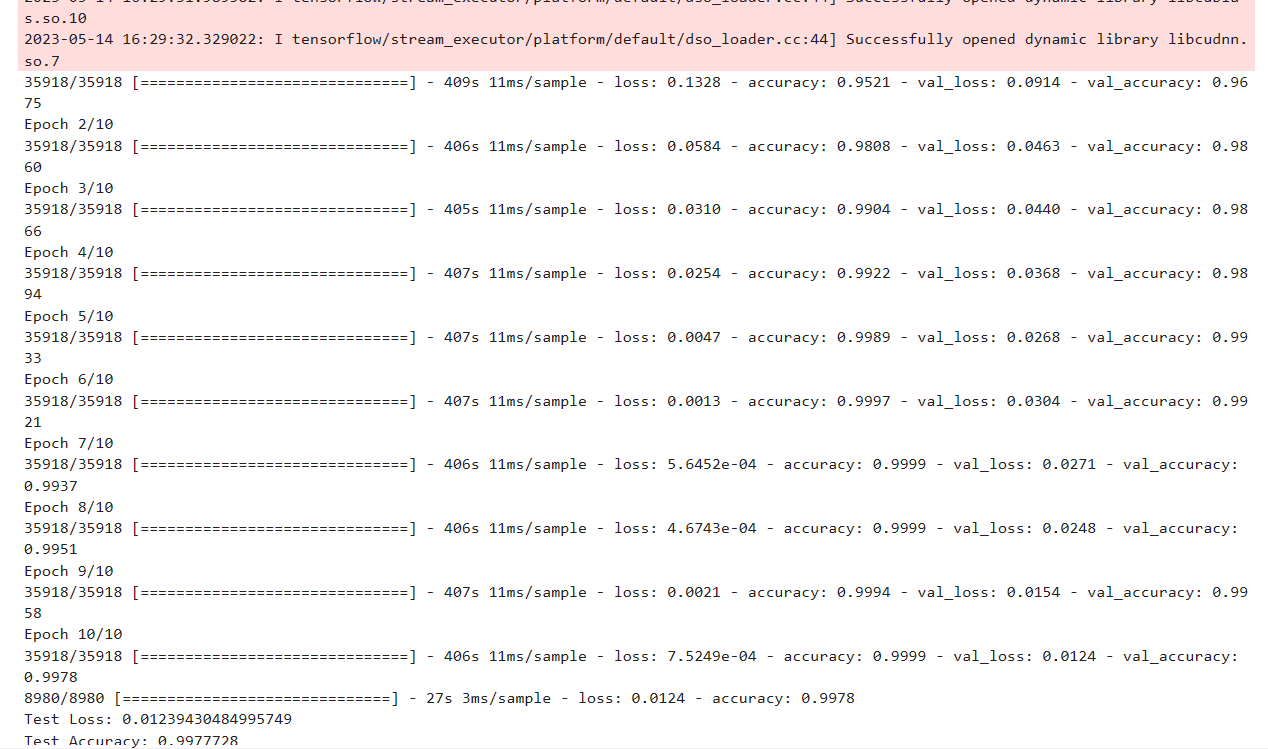


对深度学习模型而言：

我们最后选用的是老师上课提到过的hard\_sigmoid为激活函数的模型

折中了正确率和训练复杂度



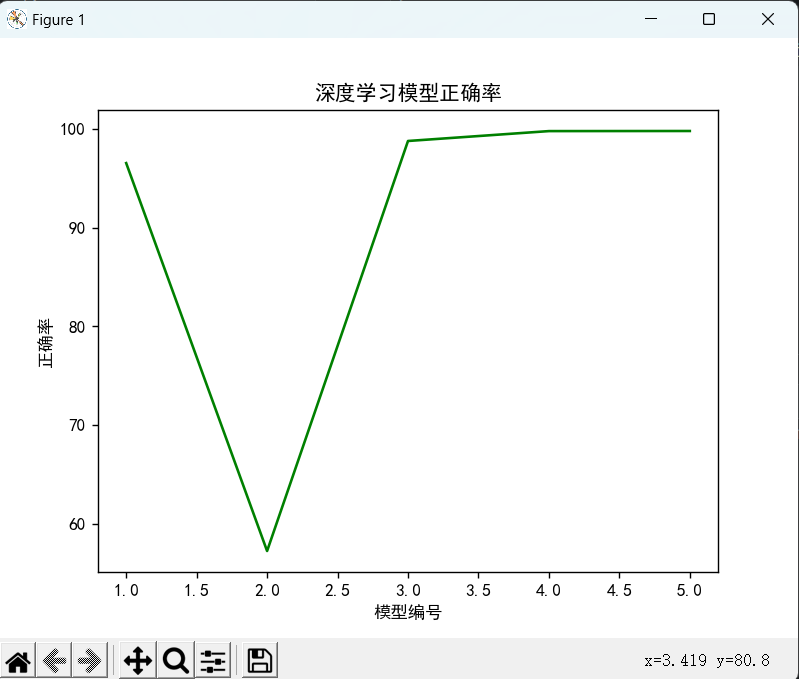


我们最后选用的模型如上所示，正确率为99.77%

对于其他深度学习模型的结论如下表所示

|  |  |
| --- | --- |
| 单层RNN模型 | 96.525% |
| 双层RNN模型 | 57.29%（未收敛，和乱猜的50%没区别基本） |
| LSTM模型：（带一个全连接层） | 98.763%（虽然收敛，但是训练过程不稳定的问题依然存在） |
| LSTM模型：（不带全连接层） | 99.865% |
| LSTM模型：（不带全连接层，使用hard-sigmoid） | 99.777% |

下图是可视化正确率分析



### 八、总结和不足:

在使用LSTM进行新闻分类的项目中，我们首先构建了一个词汇表，并将文本数据转换为序列表示。通过对序列进行填充，使其长度相同，然后构建了一个LSTM模型进行训练和预测。通过对训练数据的学习，LSTM模型能够学习到输入文本的语义和上下文信息，并对其进行分类。在训练过程中，我们使用了适当的损失函数和优化算法，并通过验证集的评估来监控模型的性能。最终，我们可以使用该模型对新的新闻文本进行分类。

LSTM能够有效地处理序列数据并捕捉长期依赖关系，因此在文本分类任务中表现出色。

但是同时也存在很多不足

1：数据量限制：LSTM模型通常需要大量的训练数据来学习有效的表示和模式。然而，在新闻分类项目中，获取大规模的标注数据可能是一项挑战。数据量的不足可能会导致模型的泛化能力不足，无法准确地分类新的未见过的新闻文本。

2：可解释性不足，该模型只是做了一个判别，但是其原因并没有解释，我认为更好的研究方式应该和论文查重软件类似，将假新闻”哪里假“标红，并给出原因。

3：LSTM模型可能无法准确地理解和处理文本中的复杂语义和含义。处理多义词、歧义和情感等方面仍然是一个挑战。

4：未处理文本噪声，在真实世界的文本数据中，可能存在一些噪声或不规范的文本，例如拼写错误。

5：缺乏处理特定领域文本的技术，专业术语的处理、实体识别等，以提高分类任务的性能。

### 九、参考文献：

**Shu, K., Mahudeswaran, D., Wang, S., Lee, D., &amp; Liu, H. (2019). Fakenewsnet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. arXiv preprint arXiv:1901.00124.**

**Thorne, J., Vlachos, A., Christodoulopoulos, C., &amp; Mittal, A. (2018). FEVER: a large-scale dataset for Fact Extraction and VERification. arXiv preprint arXiv:1803.05355.**

**Zhang, J., Su, Y., Lyu, M. R., &amp; King, I. (2019). Fake news detection on social media using geometric deep learning. IEEE Transactions on Computational Social Systems, 6(6), 1164-1175.**