**【P18】-【预测新闻头条的讽刺性】实验结果与问题报告**

**成员：陈德福**

1. **项目背景与目标**

（一）项目背景

随着互联网信息爆炸式增长，新闻平台（如社交媒体、新闻聚合网站）的头条内容成为舆论传播的核心载体。讽刺性新闻标题常通过反语、夸张、双关等手法隐含真实意图（如 “恭喜某公司产品‘再创销量新低’”），但其语义的隐蔽性导致传统文本分析技术难以精准识别。

新闻舆情误判风险：主流媒体或自媒体通过讽刺标题传递立场时，若检测模型误判为正向或中立，可能导致舆情分析偏离真实导向（如将揭露性讽刺标题误判为正面宣传，影响公众对事件的认知）。

信息过滤与推荐失效：个性化推荐系统若无法识别讽刺标题，可能向用户推送与其价值观冲突的内容，降低用户体验；同时，虚假新闻可能借助讽刺包装规避审核，加剧信息污染。

学术研究与传播需求：新闻传播学领域对讽刺修辞的量化研究依赖精准的语义检测技术，但现有模型在短文本场景下泛化能力不足，亟需针对性优化。

从舆情监测角度来看，错误解读带有讽刺意味的舆情信息，可能导致对公众情绪的误判，使相关部门或企业无法及时准确地掌握真实民意，进而制定出不恰当的决策。例如，在某产品的网络评论中，若将消费者带有讽刺的负面评价误判为正常反馈，企业可能会忽视产品的关键问题，影响产品改进和品牌形象维护。

现状痛点：新闻标题的讽刺性高度依赖语境、标点符号（如引号、感叹号）及领域术语（如 “专家” 在特定语境下的反讽用法），传统方法难以捕捉多维度语义特征，导致检测准确率普遍较低。

（二）项目目标

本项目旨在深度融合 GloVe 与 Word2Vec 技术，依托 PySpark MLlib 或 Deep Learning Pipelines 框架，构建适用于分布式环境的高效讽刺语义检测模型。通过精心设计的特征提取流程，充分挖掘文本中的语义和句法特征，运用较先进的机器学习与深度学习算法进行模型训练与优化，致力于将讽刺语义检测的准确率提升至全新高度。此模型建成后，能够精准识别各类文本中的讽刺语义，有效弥补现有技术在准确性和泛化能力上的短板，切实满足舆情监测、客户反馈分析、智能客服等多领域对于高精度文本分析的迫切需求，助力相关行业实现智能化升级，具有极为重要的现实意义与广泛的应用价值。

核心技术与框架定位：

技术定位：GloVe提供通用语义基础，Word2Vec进行领域适配，结合SVM和MLP模型实现高效分类。

框架优化：引入PySpark MLlib或Deep Learning Pipelines实现分布式训练，提高大规模数据处理效率，满足实时检测需求。构建适用于短文本场景的高精度讽刺语义检测模型，在公开新闻标题数据集（如 NTCIR-15 讽刺标题数据集）上实现85% 以上的 F1 值，满足以下需求：

领域适配性：有效识别新闻标题中依赖标点符号、领域术语、上下文语境的讽刺语义（如 “某明星‘低调’出席豪华派对” 中的引号反讽）。

短文本优化：针对标题长度短（平均 15-20 词）、语义密度高的特点，设计轻量化特征提取流程，平衡模型效率与精度。

工程落地性：基于 PySpark 实现分布式训练，支持日均百万级新闻标题的实时检测，延迟控制在 50ms 以内。

**二、实验环境与工具**

1. **硬件环境**：
   * 处理器：Intel Core i5
   * 内存：8GB 或以上
   * 存储：500GB 或以上
2. **软件环境**：
   * 操作系统：Windows 11
   * 编程语言：Python
   * 大数据处理平台：Hadoop、Spark
   * 数据分析工具：Pandas、NumPy、Matplotlib、Scikit-learn 、seaborn、nltk
   * 集成开发环境：Jupyter Notebook

**三、项目内容与步骤**

1.**数据收集与加载：**

数据来源：使用kaggle公开的 Sarcasm\_Headlines\_Dataset\_v2.json 数据集,该数据集包含以下字段：headline：新闻标题is\_sarcastic：标签（0 表示非讽刺，1 表示讽刺）article\_link：文章链接（无用字段）

加载方式：使用 pandas.read\_json() 加载 JSON 格式文件,只保留有用字段。

清洗字段：去除无用字段（如链接），仅保留 headline 和 is\_sarcastic（标签）。

2.**数据预处理**：

步骤说明：

小写转换 ：统一所有字符为小写。

去除 URL ：使用正则表达式匹配并删除网址。

去除标点符号 ：移除英文标点如逗号、句号等。

去除数字和特殊字符 ：删除带有数字的词或特殊符号。

分词 Tokenization ：将句子拆分成单词列表。

去除停用词 Stopwords ：过滤掉常见但无实际意义的词汇（如 "the", "and" 等）。

**3.特征提取与嵌入表示**：

Word2Vec 训练:使用 gensim 库训练自定义的 Word2Vec 模型，基于训练集中的 token 构建词向量。

GloVe 预训练词向量:加载外部的 GloVe 向量文件（如 glove.twitter.27B.100d.txt），构建词典映射,返回字典 {word: embedding\_vector}，每个词也是 100 维。

组合词向量（Word2Vec + GloVe):将两个模型的词向量拼接起来，形成一个更强的组合词向量，总维度为 200。

4.**模型训练:**

**模型结构概述：**

Embedding 层 ：使用固定权重的组合词向量。

BiLSTM 层 ：双向 LSTM 提取上下文信息。

Dropout 层 ：防止过拟合。

Dense 层 ：全连接层进行分类。

Sigmoid 输出层 ：输出概率值（0~1）判断是否为讽刺。

1. **模型评估及可视化：**

分类报告：使用 classification\_report 输出了模型的精确率、召回率和 F1 分数，详细分析了模型在两类样本上的表现。

混淆矩阵：通过 confusion\_matrix 计算了模型的预测结果，并使用 Seaborn 的热力图进行可视化。混淆矩阵清晰地展示了模型的错误分布，例如哪些样本被误判为讽刺或非讽刺。

**四、实验结果**

1. **数据预处理结果**：

通过一系列预处理步骤，将原始数据转化为一个高质量的数据集，具体表现为：

每条新闻标题被清洗为简洁、规范的文本。

标题被分词并去停用词，生成了干净的词列表。

数据集结构清晰，标签分布合理，为后续的特征提取和模型训练提供了坚实的基础。

这些预处理工作是整个项目的基石，确保了模型能够高效地学习和泛化。

1. **特征提取与嵌入表示结果：**

Word2Vec 和 GloVe 的联合嵌入捕捉了单词的上下文信息和全局语义信息。这种组合方式显著提升了模型对复杂语言模式的理解能力。

标准化输入 :文本序列化和填充确保了输入数据的一致性，便于神经网络处理。

高效表示 :每个单词被映射为一个 200 维的密集向量，既保留了语义信息，又减少了稀疏性问题。

这些特征提取与嵌入表示的结果为后续的深度学习模型（如双向 LSTM）奠定了坚实的基础。

**3.模型评估及可视化结果：**

Keras 模型表现优异 :

准确率较高，在非讽刺类别上的表现尤为突出。

混淆矩阵显示模型在讽刺类别上的召回率略低，可能是由于讽刺语言的复杂性导致。

Spark 模型具有扩展性 :

Spark 模型在分布式环境中表现良好，适合处理大规模数据集。

准确率和 F1 分数接近 Keras 模型，说明其具备较强的泛化能力。

模型改进空间 :

对于讽刺类别召回率较低的问题，可以尝试增加更多讽刺样本或调整模型结构。

引入更复杂的特征（如情感分析、句法特征）可能会进一步提升模型性能。

模型训练结果充分证明了所选方法的有效性，为讽刺新闻检测任务提供了可靠的解决方案。

**五、遇到的问题与解决方案**

**1.数据质量问题**

**问题描述 :**原始数据中存在噪声（如 URL、标点符号、停用词等），影响了模型的学习效果。数据集中两类样本的数量不均衡（非讽刺样本多于讽刺样本），可能导致模型对少数类别的预测能力不足。

**解决方案：**

设计了一个全面的 clean\_text 函数，去除了 URL、标点符号、数字和特殊字符，显著提高了文本质量。通过上采样（upsampling）解决了数据不平衡问题，使两类样本的数量更加均衡。

2.**特征提取中的词汇覆盖率问题**：

**问题描述**：Word2Vec 和 GloVe 嵌入可能无法覆盖所有词汇，尤其是训练数据中出现的罕见词或未登录词（OOV，Out-of-Vocabulary）。这可能导致某些单词无法获得有效的向量表示，进而影响模型性能。

**解决方案**：组合嵌入矩阵 :将 Word2Vec 和 GloVe 的嵌入结合起来，为每个单词生成了一个 200 维的联合嵌入向量。这种方法充分利用了两种嵌入的优点，增强了模型对语义的理解能力。固定嵌入权重 :将嵌入层设置为不可训练（trainable=False），防止模型在训练过程中更新嵌入权重，从而保留了预训练嵌入的高质量语义信息。

3.**模型过拟合问题**：

**问题描述**：深度学习模型（如双向 LSTM）容易在小规模数据集上发生过拟合，导致验证集和测试集的表现下降。

**解决方案**：在模型中添加了 Dropout 层，随机丢弃部分神经元（比例为 0.5），有效防止了过拟合。使用二元交叉熵损失函数和 Adam 优化器，进一步提升了模型的泛化能力。

**4. 数据分布可视化问题**

**问题描述 :**

在分析数据分布时，如何直观地展示两类样本的数量差异以及模型的错误分布是一个挑战。

**解决方案 :**

条形图展示样本分布 :使用 Matplotlib 和 Seaborn 绘制条形图，清晰展示了两类样本的数量差异。

**六、总结与计划**

1、项目总结

通过本项目的实施，成功构建了一个用于讽刺新闻检测的机器学习系统，并在数据预处理、特征提取、模型训练和评估等方面取得了显著成果。以下是项目的主要成就和总结：

主要成就：对原始数据进行了全面的清洗（如去除噪声、分词、去停用词等），为后续建模奠定了坚实基础。数据平衡技术（如上采样）有效解决了类别不平衡问题。Word2Vec 和 GloVe 嵌入生成了高质量的词向量，并通过组合嵌入矩阵增强了语义表达能力。文本序列化和填充确保了输入数据的一致性，便于神经网络处理。Keras 模型采用双向 LSTM 和 Dropout 层，捕获了文本的上下文信息并防止过拟合。Spark 模型利用分布式计算框架，展示了扩展性和高效性。分类报告和混淆矩阵详细分析了模型的表现。热力图和条形图直观地展示了模型的错误分布和数据分布。

经验总结：数据清洗和平衡是提升模型性能的关键步骤，尤其是在处理文本数据时。组合嵌入矩阵充分利用了不同嵌入的优点，能够显著增强模型对复杂语言模式的理解能力。在深度学习模型中，防止过拟合和优化超参数是提升性能的重要手段。在分布式环境中，合理配置资源和调整流水线结构是提高效率的关键。条形图、热力图等可视化工具不仅帮助理解模型表现，还为后续改进提供了方向。

2.后续计划：

数据层面 :引入更多数据源（如社交媒体数据）和多语言数据，增强数据多样性。优化预处理流程，尝试更高级的分词和词形还原方法。

特征提取 :使用上下文感知嵌入（如 BERT、RoBERTa）代替静态嵌入。结合句法特征或图像信息进行多模态学习。

模型改进 :尝试复杂架构（如 Transformer、Attention 机制）或迁移学习（微调预训练模型）。优化超参数，提升模型性能。

评估与部署 :增加评估指标（如 AUC-ROC、PR 曲线），全面衡量模型表现。优化推理速度，部署到实际应用场景（如移动端或在线系统）。

扩展研究 :探索多任务学习、无监督学习等技术。将讽刺检测应用于其他领域（如舆情分析、社交媒体监控）。