邹卓

男 | 28岁 | 籍贯：长沙 15717514161 zouzhuo0@gmail.com

7年工作经验 | 求职意向：自然语言处理算法工程师

个人优势

具有7年 python 工作经验，能够承担核心模块的设计与研发工作，完成高质量交付；
熟悉深度学习任务中 RNN/LSTM/GRU 相关模型以及 Transformer 等深度学习模型；
熟悉 BERT 及其的变体 RoBERTa、XLNET 等以及 GPT 系列模型等预训练模型；
熟练使用Bge 等文本向量化（Embedding）模型；
熟悉大语言模型提升工程，包括CoT，Few-shot等提示学习；
熟悉大模型ChatGPT4.0，ChatGPT3.5，minimax，Doubao接口
Prompt应用；以及ChatGLM系列、Qwen、LLaMa系列、Mistral系列等大模型 LoRA 、QLora、P-tuning v1、P-tuning v2 等微调微调方式；
熟悉RLHF等多种大模型对齐方式；
熟练使Pytorch,Transformers,Langchain，Milvus，FAISS等深度学习工具，熟悉知识蒸馏，并行训练；
了解多模态，并将 CLIP 模型多模态特征词嵌入应用到文本分类任务上。

工作经历

CSDN 算法工程师 2024.01-至今

1.使用机器学习模型对全站海量数据进行特征提取、识别、分类等
2.针对用户query建词权重模型，分词算法基础模型
3.应用Elasticsearch来对csdn进行全文检索，针对用户query，能高性能精准检索出用户意图需求
4.基于LLM api构架C知道大模型RAG服务
5.构建技术标签体系，应用于亿万级CSDN站内资源IT技术标签分类

美银宝网络信息服务（上海）有限公司 大模型算法工程师 2021.10-2023.12

内容:

1.辅助公司精准合规风控项目；
2.负责开发合规；
3.研究大语言模型提示工程，微调，进行大规模的文本数据的特征分析，提取建模，并对NLP深度学习模型调优；
4.负责对文本分类，文本相似度，大模型生成，NL2SQL等相关的NLP模型设计和优化。
5.将大语言模型集成到公司的合规项目上，通过提示工程，RAG，实现文本精准召回和文本生成。

业绩:

1.将提高合规风控控制目标召回率到由10%提升到92%。
2.通过ChatGPT，推荐更符合用户的模型输出，业务指标提升到70%。

平安银行股份有限公司 算法工程师 2018.03-2021.08

内容:

自然语言处理工程师
1. 基于商品数据(含结构化和非结构化字段)，用 python，pandas, excel，sql 等工具进行统计分析；

2. 定义数据质量指标，持续提升数据质量；

3.研究深度学习算法和自然语言处理（中文），进行大规模数据文本的特征分析、提取建模，并对 NLP 深度学习模型的调优；

4.负责参与文本分类、命名实体识别、语义理解，文本生成等相关 NLP 模型的设计和优化；

5.负责尝试各种算法，深度学习和注意力模型，提高识别的准确率以及召回率；

6.利用 NLP 相关技术、深度学习理论和方法解决实际问题，改进产品功能；

业绩:

1.客户流失预警准确率提升约 11%，间接规避资金损失约数百万元

2. 年度节省人工标注与审核成本约 500万元人民币

3.人力减少约 60%，由原来平均每天产出 10 份提升到 25+

项目经历

CSDN C知道 AI搜索与问答系统研发　算法工程师　2024.01-2024.12

内容:

CSDN平台累积了海量技术内容，传统站内搜索只能基于关键词检索，用户在提问时经常遇到：搜索结果过于分散，难以快速聚合答案；对冷门问题召回能力差，内容覆盖有限；LLM直接对话缺乏上下文和可信引用；为提升搜索体验，平台决定开发基于RAG（Retrieval-Augmented Generation）的AI搜索问答系统，将站内内容与大模型结合，实现更准确、可解释的回答，尤其对冷门长尾问题显著提升答案质量。

技术栈:RAG,Prompt Learning，PyTorch、Transformers、bge-small-zh-v1.5、LLM API

V1.0

基础检索生成流程打通；用户问题 → 站内关键词搜索（BM25）返回Top-N文章拼接到Prompt；大模型直接生成答案。回答引用准确:50%;用户点击采纳率:21%。

V2.0

引入分片+向量化相似度；微调BGE模型；问题与片段分别向量化；余弦相似度筛选Top-K片段；Prompt结构优化；片段显式编号；回答时强制引用片段；公域向量库初步整理。回答引用准确:78%;用户点击采纳率:34%。

V3.0

向量库进一步扩展至全域资料（图书、行业规范）；完成C知道产品和全站90%以上的Prompt，多轮Prompt组合（摘要 + 引用 + 回答分步生成）；分段召回+分层聚合策略；Answer Confidence Score模型（对生成回答置信度打分）；联合评估标准上线（自动+人工）回答引用准确:90%;用户点击采纳率:48%。

业绩:

AI搜索与问答替代传统站内搜索成为入口

用户对冷门问题的留存与复访率显著增长

搜索流量转化知识付费订单明显增长

全站内容结构化工程　算法工程师　2024.10-至今

内容:

CSDN拥有海量非结构化内容，包括博客、下载资源和文库文件，内容体量超过5.9亿条数据（博客约4600万，下载资源和文档约1300万）。

传统的内容搜索、推荐和运营均依赖于有限的手工标签和规则，存在：

数据规模大、处理难度高；内容理解不足，标签覆盖率低；新增内容无及时结构化能力。为此，团队研发自动化内容清洗与结构化

平台，利用大模型和分布式计算对全站内容进行清洗、意图识别、技术标签打标和批量入库，提升搜索、推荐及AI问答的底座数据能力。

技术栈:Prompt Learning,LLM api (Minimax,Doubao,Qwen,Chatgpt)

V1.0：基础内容清洗与意图识别

对全站博客和下载资源完成：分批拉取；文本提取；基础规则去重

调用大模型进行意图分类（仅一级分类）输出结构化结果入库，构建增量清洗流程

每日生成清洗进度报告

完成博客约5500万条、下载资源800万条清洗；覆盖率约60%

V2.0：技术标签与AI、标题、摘要、关键字生成

建立计算机技术标签体系（三级分类）；对全站资源自动多标签打标

在清洗流程中集成AI摘要生成；搭建定时机器人推送进度与指标；标签挂接至三级技术标签体系；

累计清洗博客5000万条、下载资源1200万条，技术标签准确率提升至88%

结构化覆盖率约85%

业绩:

提升搜索与推荐体验

AI问答质量和引用准确性显著提升

促进内容运营效率大幅提高

自建ES搜索系统　算法工程师　2024.11-至今

内容:

随着CSDN平台内容量和用户访问量的持续增长，全文检索服务面临多重挑战，本项目面向CSDN平台站内全文检索系统，针对搜索结果相关性差、性能扩展性不足和技术语义理解能力弱等问题，设计并实施了一套自建Elasticsearch集群、定制分词算法和词权重优化方案。

技术栈：词权重,BGE , HanLP

1.HanLP 分词算法优化

1.1基于业务场景，不断扩充领域内词典

1.2应用下游搜推业务场景，完成HanLP内部规则以优化

2.搜索Query词权重建模与优化:

1.1 基于BGE,构建token到词的映射，并基于HanLP的分词结果，借用大模型辅助完成数据标注，完成SFT。

1.2 分析badcase，扩充样本量，平衡样本分布，提高模型效果。

业绩:

用户点选Top3结果的点击率提升 15%

平台整体搜索转化率提升 10%

用户留存率、活跃度均有增长

博客付费意愿分模型　算法工程师　2024.01-2024.06

内容:

为了提升内容变现能力，平台希望通过自动化模型识别优质博客内容，对博客进行付费意愿打分，精准判断哪些内容更适合设置付费阅读，提高整体收入。此前缺乏科学的打分体系，依赖人工经验，导致高价值内容无法有效挖掘。

v1.0

基于分词和搜索结果构建初版稀缺分模型

利用基础关键词热度统计生成需求分

仅用简单规则计算质量分

付费内容识别准确率：约65%

v2.0

引入泰勒公式与聚类方法优化需求分热度预测

使用XGBoost代替规则质量分模型

加入用户行为多特征（阅读时长、评论数）

付费内容识别准确率提升至：78%

日均收入提升约：150%

v3.0

集成多数据源（站内外搜索数据）增强稀缺分鲁棒性

高频词库基于词向量相似度去噪

针对新博客增加内容特征预测质量分（冷启动优化）

完善分数加权策略

付费内容识别准确率提升至：88%

日均收入提升约：250%

业绩：

博客日均收入提升约250%

Miss control Identification　LLM Algorithm engineer　2023.06-2023.12

项目概述：

全球政府组织机构每年针对PayPal支付项目进行风险合规管控，需要针对合规内容文件，公司拆分针对不同的文件拆解称为单条Citation。然后进行不同的Citation，使用不同Control objective来管控每笔交易是符合当地的法律法规的。

V1.0

合规风控推荐算法baseline.基于ChatGPT4，从海量的PDF合规文件中获取提取不同的Citation，同时对不同文档，使用Langchain抽取出Citation，然后针对不同Citation 基于数据库中的Control objective推荐Top20的Control objective,及其推荐原因。

相关技术:Langchain,ChatGPT4,Col,RAG

Recall top20: 9.85%

V2.0

优化精排粗排方式，训练citation 和 control objective二分类模型。

相关技术:Minilm Loss build

Recall top20: 42.25%

V3.0

topicmodel以及keyword embedding召回方式(优化RAG向量化embedding)，提高模型的召回准确率。

相关技术:keybert,ChatGPT4.0 Few shot

Recall top20:82.25%

NL2SQL智能查询与分析系统　算法工程师　2022.12-2023.09

内容：

项目概述：

该项目在PayPal风控与销售业务中，通过集成大语言模型，实现自然语言和语音交互快速生成SQL查询，获取高风险客户、交易行为、销售转化等多维度数据，提升查询效率和数据利用率。系统显著降低了对技术能力的依赖，帮助一线员工以对话式方式完成复

杂分析，助力“人人都是数据分析师”。
技术框架：ChatGPT、Llama、Whisper、NL2SQL、Slot-Filling、Python、PostgreSQL、FastAPI、Docker

工作内容：
基于大语言模型设计NL2SQL转换引擎与多模态交互界面，构建风控/销售数据库的意图识别体系，实现自然语言到SQL的映射与安全执行，并输出可视化报告。

v1.0
集成Llama（13B）与Whisper模型：支持语音和文本输入；基于Slot-Filling检测意图和关键字段；自动生成SQL查询执行数据库；查询结果以文本和图表反馈；初版SQL生成准确率74%。

v2.0
优化复杂查询与语义理解：引入Prompt Engineering和Few-shot Learning提升复杂SQL稳定性；增强领域术语词典覆盖，提升识别准确率；SQL生成准确率提升至90%，查询响应时延缩短40%。

v3.0
多轮对话与上下文管理：支持连续查询和上下文记忆；优化API服务高并发能力；部署CRM、风控系统集成接口；销售分析工单减少约65%，风控报告生成时间缩短至30秒。

业绩：
风控团队能在30秒内生成高风险客户报告
销售人员通过语音直接查询季度转化率
年度节省数据分析支持人力成本约200万美元

美银宝智能风险文本分析与异常检测系统　自然语言处理工程师　2022.05-2022.11

内容：
自动识别和分析风控业务中大量非结构化文本数据，如客户投诉、交易备注、合规报告等，及时发现潜在风险和异常行为。传统人工审核成本高、响应慢，难以满足大规模实时监控需求。系统基于深度学习文本理解模型，实现风险信息的自动抽取、分类和告警，提升风险识别准确率和效率，辅助风控专家进行决策。
技术框架：BERT预训练模型、BiLSTM-CRF实体识别、文本分类、异常检测算法

● V1.0
基于BERT微调文本分类模型，实现三分类风险等级判定，准确率达到85%
使用BiLSTM-CRF进行风险实体抽取，F1分数约为82%

● V2.0
引入对抗训练和数据增强，提升模型鲁棒性
开发多标签分类能力，支持风险类型细粒度划分
异常检测算法融合统计规则和机器学习，召回率提升20%

● V3.0
集成风控知识图谱，增强实体识别与关联分析能力
部署在线实时风控文本分析系统，支持高并发请求
风险识别整体F1提升至90%，系统响应时延降低30%

业绩：

风控团队自动筛查文本工单比例提升60%

预警响应时间从小时级缩短至分钟级

有效降低风控人工成本，减少潜在经济损失

美银宝客户服务自动工单分类系统　算法工程师　2021.10-2022.06

内容：

项目概述：PayPal客户服务中心每天接收海量客户工单，涉及付款失败、账户安全、退款请求等多个类别。传统人工分发工单效率低，分类准确率不稳定，影响后续处理时效和客户满意度。为此，项目团队基于自然语言处理与机器学习技术，设计了一套自动工单分类系统，实现多类别文本快速准确分类，提升客户服务的自动化水平。

● 技术框架：TF-IDF XGBoost

● 工作内容：构建大模型生成式算法

● V1.

基于TF-IDF特征，训练XGBoost模型进行三分类（支付、账户、退款），分类准确率约78%

● V2.

优化XGBoost参数（max_depth、eta、subsample等）提升模型效果

扩展支持更多业务类别标签，分类准确率提升至85%

v3.0

集成Word2Vec词向量增强文本语义特征

增加特征重要性筛选和标签一致性校验

实现线上API自动分类工单，日均分发效率提升30%

业绩：

人工分单比例降低60%，客服人力成本节约显著

客户满意度提升，投诉率下降

工单处理效率显著改善

HiSiri,智能对话系统；自然语言处理研究南洋理工大学组　自然语言处理工程师　2022.03-2023.01

项目概述： 该项目针对用户基于数据集 LCCC （包含超过一千万个 session 的闲聊对话，是清华大学提供的一个大规模中文对话数据集），智能对话系统主要实现了自动与用户进行对话的功能，并且帮助用户完成特定的任务。该项目主要是应用于对话任务中，同时也可以几乎应用在所有文本生成任务，包括文本摘要，机器翻译等。

技术框架： Pytorch GPT

工作内容：实现了基于预训练 Transformer的 decoder 端的生成式 NLG 模型 GPT 中文预训练模型，实现了一个自己的基于字级别的 tokeneizer 并通过多种解码技术包括 beam search,greedy decoding ，使用数据并行的方式利用多张 GPU 训练模型，并得到的一个生成式的模型。

智能风控报告自动生成系统　自然语言处理工程师　2020.08-2021.07

内容：

在平安银行的信贷与风控业务中，员工需要定期撰写大量客户信用分析报告、贷后风险审查文书和合规说明，工作量大、重复性高、人工撰写效率低，且质量不稳定。为了降低人工成本并提升文档生成质量，项目团队设计并实现了一个基于NLP文本生成技术的智能报告生成系统，支持从结构化数据自动生成标准化报告草稿，并可进行人工校对后直接使用。

技术框架：BART、T5、Slot-Filling模板生成、Jinja2模板引擎、Copy Mechanism、Rule-enhanced Generation、Python、Beam Search、Top-k Decoding

工作内容：

基于BART深度学习生成模型，结合模板填充和规则引擎，对结构化数据进行报告文本生成；搭建多段落生成与校对流程，实现自动化风险报告的生产和人工审阅闭环；优化生成质量与一致性，持续迭代提升业务覆盖率。

v1.0
基于Slot-Filling模板生成：
设计标准化合规模板，将客户信用评分、收入、负债等字段直接填充生成高合规性段落；
使用Jinja2模板引擎实现模板自动渲染；
支持基础字段到文本的映射，保证术语准确。

v2.0
引入BART文本生成模型：
收集并脱敏约5万条历史风控报告，进行分段标注（如客户基本情况、信用分析、风险结论）；
微调BART模型，输入JSON结构化数据及客户交易摘要，输出自由表达的分析性文本；
采用Beam Search + Top-k Decoding策略控制生成多样性与稳定性。
(ROUGE-L：0.41；BLEU-4•：0.32 人工评分一致性（满分5分）：2.8 生成效率（平均每段）：2.1s)

v3.0
多段落协同生成与模型融合：
针对不同段落分别训练小模型，实现分段落生成后拼接；
引入规则模板与语言模型融合（Rule-enhanced Generation），部分句式由规则生成，部分句式由模型生成；
集成Copy Mechanism，提升对原始字段的忠实表达能力，减少关键字段误写。

v4.0
使用蒸馏后的 T5-small 模型做在线部署，主模型用于批量生成。
（ROUGE-L：0.58；BLEU-4：0.46 ；人工评分一致性（满分5分）：4.2；生成效率（平均每段）：0.7s（distilled 模型））

业绩：
实体识别系统上线后，信贷审批资料自动审核覆盖率提高约 30%。
风控系统企业黑名单匹配准确率提高，有效降低人工审核成本。
企业画像构建更为完整，支持后续精准营销与信用评估。
年度节省人工标注与审核成本约 500万元人民币。

企业与法人智能抽取项目　自然语言处理工程师　2019.12-2020.09

内容：
平安银行在信贷审批、风控审核、企业尽调等场景中，亟需精准、自动化地从非结构化文本（如企业公告、新闻报道、工商信息、合同条款等）中提取企业名称及法人实体信息。传统基于规则的抽取方法存在覆盖不足、泛化能力差、人工维护成本高的问题，无法满足高频复杂的业务场景。为此，平安决定基于自然语言处理（NLP）技术，搭建一套高准确率、可扩展的企业及法人实体识别系统。

技术框架：BERT-CRF（哈工大BERT-wwm-ext）、BIO标注体系、API服务化部署

工作内容：
使用命名实体识别（NER）技术，结合预训练BERT模型进行企业名称与法人实体的抽取，基于BIO标注法构建标签体系，整合多源数据完成标注和清洗，进行模型训练与部署，并与信贷、风控、企业画像等系统集成。

v1.0
基于BERT-CRF（Bidirectional Encoder Representations from Transformers + Conditional Random Field）架构进行企业名称与法人实体的抽取。

v2.0
引入BERT预训练模型（BERT-wwm-ext），进行迁移学习与Fine-tuning，采用分词、正则过滤、数据清洗等多步提升数据质量；在约5万条企业文本数据上进行训练，准确率和召回率显著提升。

v3.0
针对嵌套与重叠实体问题，引入层次化命名实体识别（Nested NER），并结合对抗训练与数据增强，显著提升小样本实体的识别效果，企业名称识别准确率提升至95%，法人实体召回率提升至88%。

v4.0
集成企业知识图谱，构建企业与法人的实体关系库，增强识别准确性和实体消歧能力，系统实现日处理文本量100万条以上，支持高并发业务需求，实体整体F1值提升至92%。

业绩：
实体识别系统上线后，信贷审批资料自动审核覆盖率提高约 30%。
风控系统企业黑名单匹配准确率提高，有效降低人工审核成本。
企业画像构建更为完整，支持后续精准营销与信用评估。
年度节省人工标注与审核成本约 300万元人民币。

企业舆情情感分析系统　算法工程师　2018.12-2019.07

内容：
在金融风险管控与客户管理过程中，平安银行需实时掌握企业相关的舆情态势，尤其是负面情感信号。项目旨在通过文本分类技术构建一个高效、可解释的舆情情感分析系统，为授信审批、风控预警、客户画像等业务提供智能支撑。

技术框架：Python、PyTorch、Sklearn、LSTM、BERT、FastText、Flask

工作内容：
基于LSTM与BERT双模型融合方案进行文本情感三分类（正向/负向/中性），针对金融领域文本进行微调；结合规则引擎提升模型解释性；搭建API服务，实现与内部信贷风控平台、客户画像系统的无缝集成；优化模型推理性能，提升系统处理能力。

v1.0
构建初版LSTM模型，输入为TF-IDF特征与Word2Vec/ FastText词向量，对金融领域舆情文本进行情感分类；结合金融黑名单词库与规则引擎进行后处理，情感识别准确率达到84%。

v2.0
引入BERT预训练模型，采用“预训练+微调”策略，增加分类头进行三分类；通过Flask API部署在线服务，支持多业务线接入；情感识别准确率提升至91%以上，但模型推理时延较高。

v3.0
开发模型融合层，采用stacking方法对BERT与LSTM输出进行加权融合，提升整体鲁棒性和解释性；引入领域专属词典进行分词与实体识别，评估指标达到：Precision：92.3%；Recall：89.7%；F1-score：90.9%

v4.0

针对性能瓶颈，对BERT模型进行量化与蒸馏，将大型模型压缩为小模型以降低推理延迟，单条文本推理时延优化至≤120ms；构建分层处理机制，使用FastText+LSTM轻量模型处理低优先级文本，BERT处理高优先级文本。

业绩：

在广州分行部署后，负面舆情识别提前率提升 2~3 天，协助信贷部门提前介入风险企业；

客户画像系统接入舆情情感维度后，客户流失预警准确率提升约 11%；

作为企业合规风控的一部分，辅助风控系统完成多起高风险客户预警，间接规避资金损失约数百万元

教育经历

湖南工业大学　本科　材料学　2014-2018

1. 在校期间，参加过数据分析与挖掘竞赛，锻炼了自己的数据分析能力。
2. 学科成绩优异，数学、统计学、计算机语言等相关科目均获得了较高的分数，具备扎实的理论基础。
3. 参与了校内数据分析项目，负责数据清洗和建模工作，熟悉了数据处理流程和常用工具，具备较强的数据分析实战能力。

资格证书

大学英语四级