多功能在线文本分析平台

CONTENTS

产品背景

在当今社交网络上，虚假新闻的流行和传播速度超过了真相，这已成为信息生态系统的一个重要挑战。根据《科学》杂志上发表的研究，对Twitter上12.6万个被质疑新闻故事的分析显示，虚假信息在社交网络中占据主导地位。研究报告指出，无论从用户吸引度还是传播速度来看，虚假新闻都表现出更强大的影响力。在过去10年间，这些虚假新闻和谣言比真实的信息更具有吸引力，更深入地渗透到用户之间的互动中。然而，大量虚假信息和误导性新闻的大范围传播和长时间存在，对社会、政治和经济稳定带来了潜在威胁。在这个背景下，开发一种多功能文本分析系统具有重要意义。

产品概述

这个多功能文本分析系统提供了一站式的解决方案，该系统整合了自然语言处理和机器学习技术，旨在帮助用户从新闻文本中获取关键信息、情感分析、主题识别，并判断新闻真伪，为各领域用户提供可信赖的数据支持。通过自动化的分析流程，用户能够高效地获取文章核心信息，节省时间并提高决策效率。

产品概述

提供情感分析功能，以帮助用户了解新闻文章的情感色彩。

提取新闻文本中的关键词，以便用户快速获取文章的主要内容。

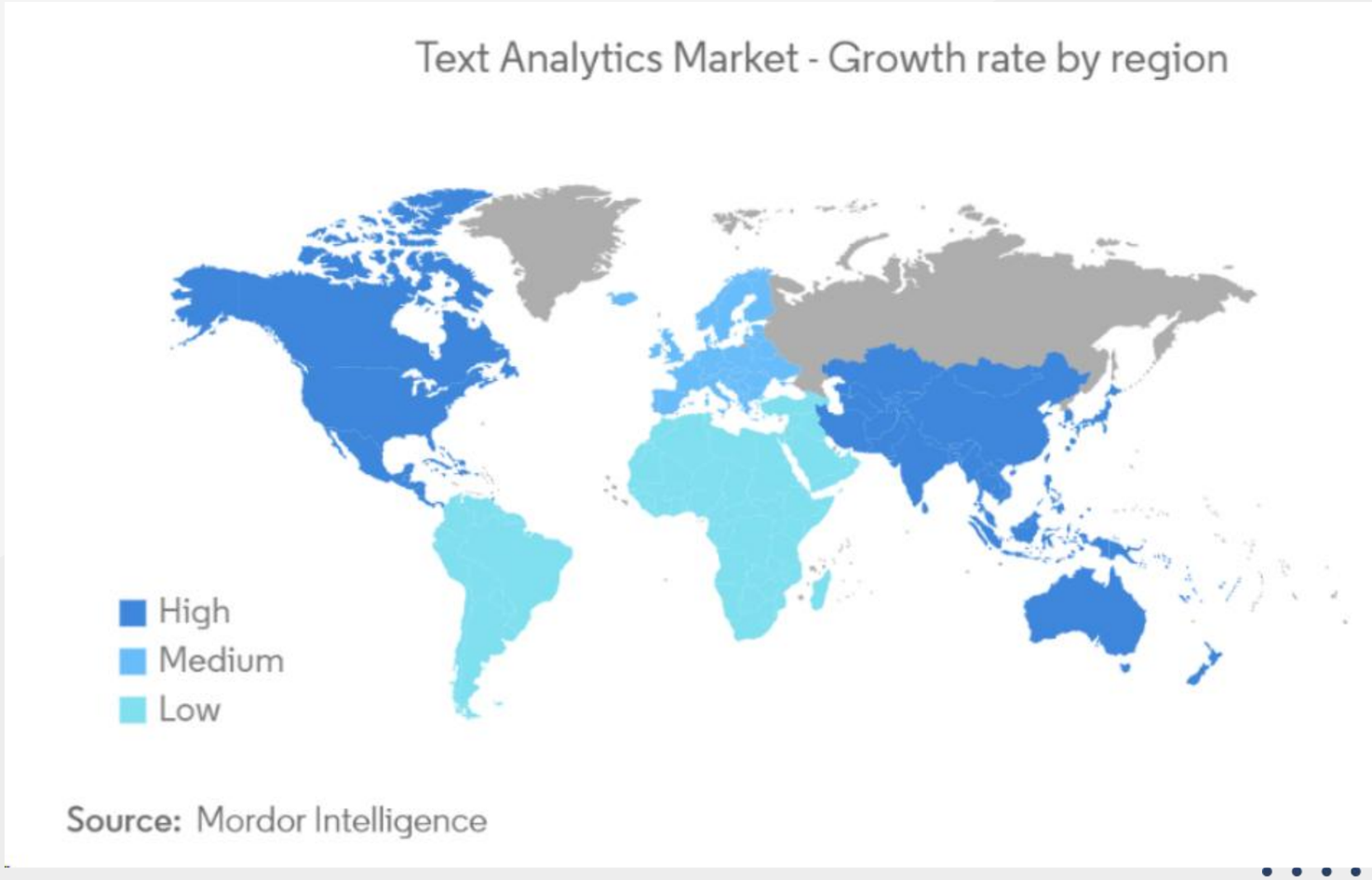
提供文本摘要，以节省用户的阅读时间并提供新闻的要点。

生成高频词词云图，可视化文章的关键词，以更好地理解新闻的重点。

识别新闻主题，帮助用户了解新闻所涵盖的主要话题。

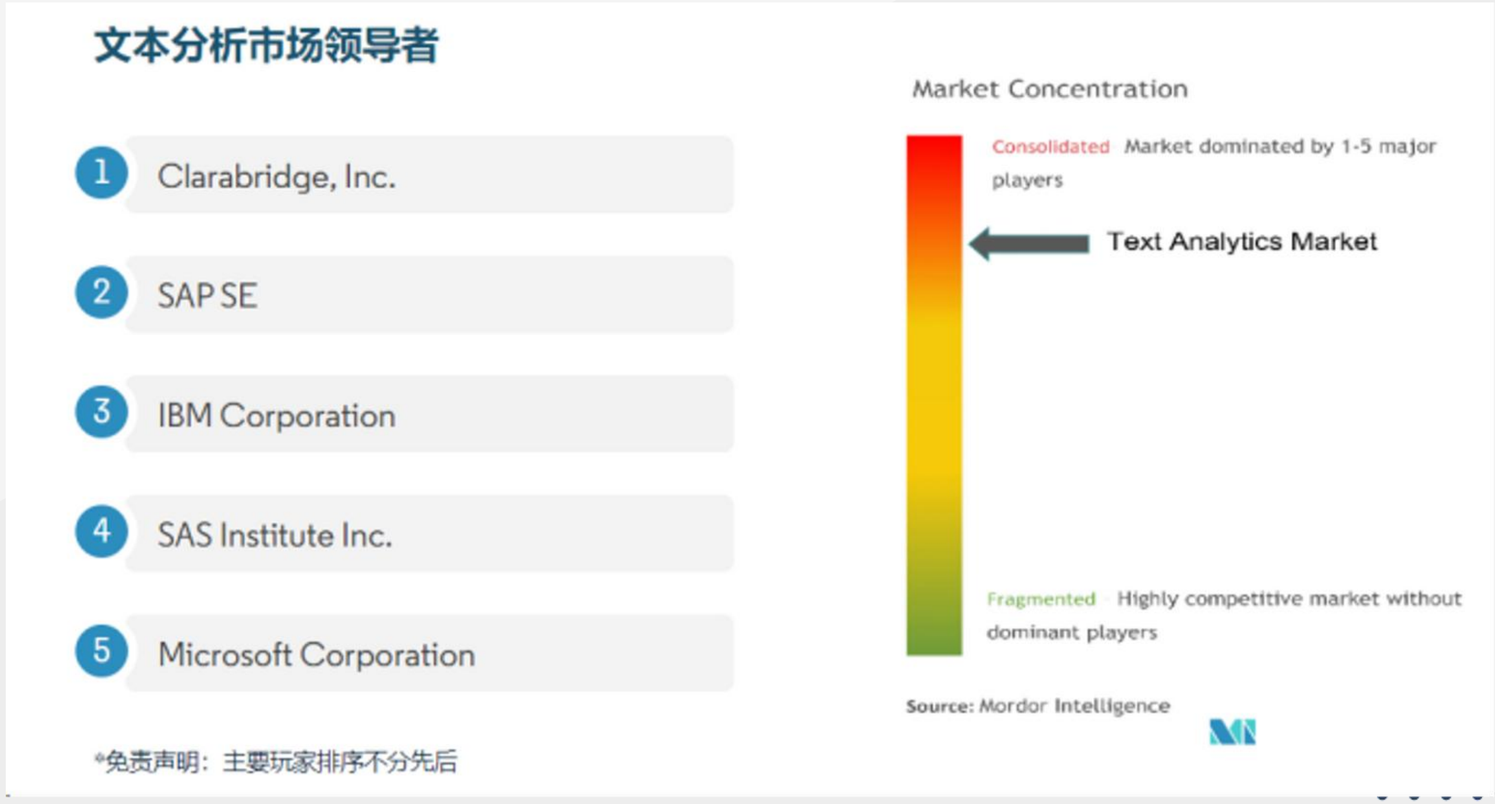
市场分析

在文本分析市场规模和份额分析-增长趋势和预测（2023-2028）里指出，文本分析市场在上一年的价值为64亿美元，预计在预测期内将以39.90%的复合年增长率增长，到未来五年将达到402亿美元。许多基于云的应用程序使用文本分析，包括预测分析、欺诈管理、风险管理、商业智能和网络犯罪预防。



市场需求

市场对于真实、可信的信息需求日益增长。新闻机构、社交媒体平台、政府机构、教育和研究机构以及普通用户都渴望有一种可靠的工具来验证信息真实性。这种系统能够满足他们对文本数据快速处理和准确分析的需求。



目标用户

新闻机构和媒体

新闻行业面临着信息真实性的挑战，机构需要保证其报道的可信度，以维护读者信任和品牌声誉。

普通用户

拥有一个工具可以自主验证信息真实性是至关重要的，避免分享或相信虚假信息。

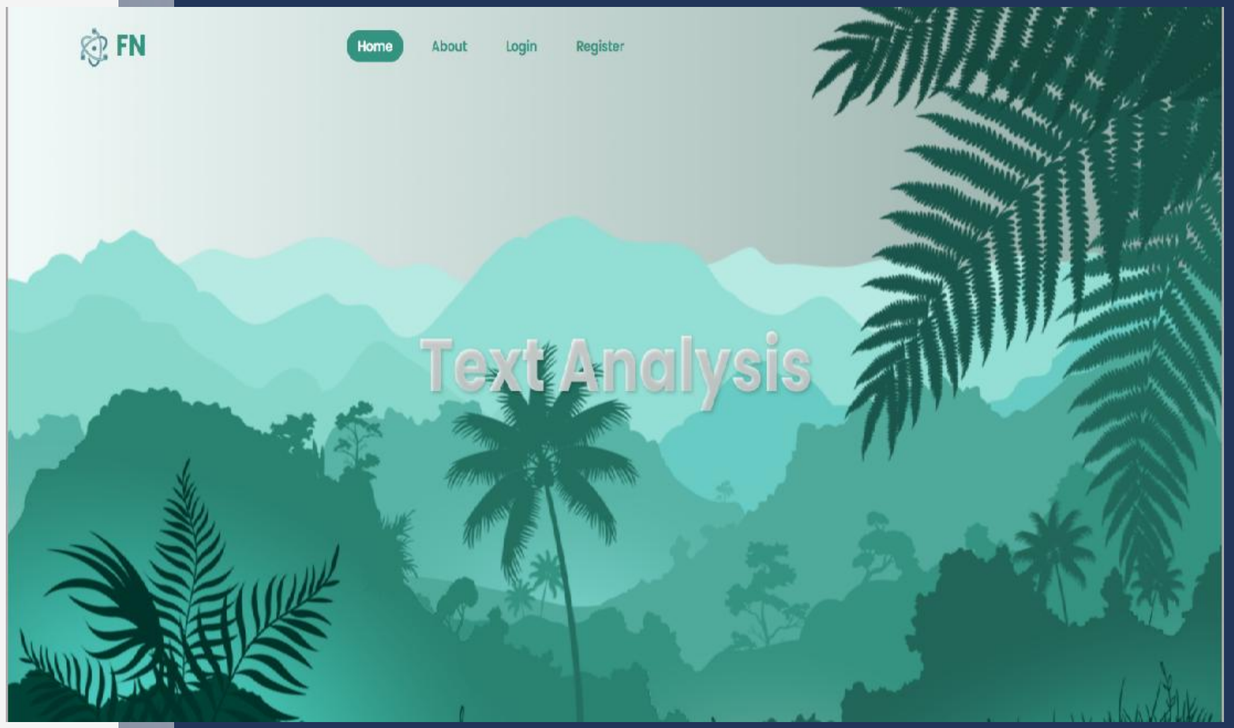
社交媒体平台

社交媒体作为信息传播的重要平台，需要防止虚假信息的传播，以维护用户信任度和平台声誉。

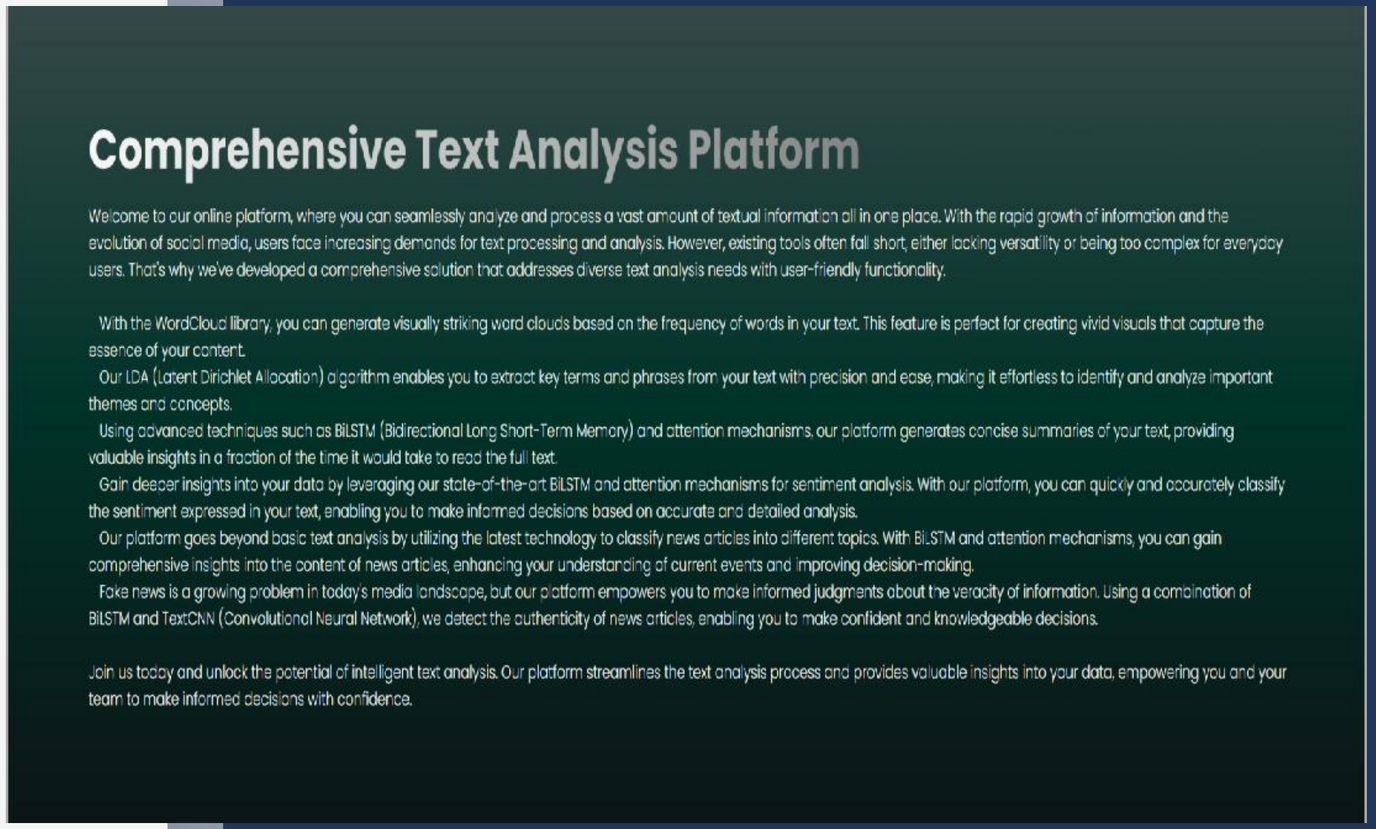
政府机构

防止虚假信息对公共秩序和国家安全带来的潜在威胁至关重要。

主页界面

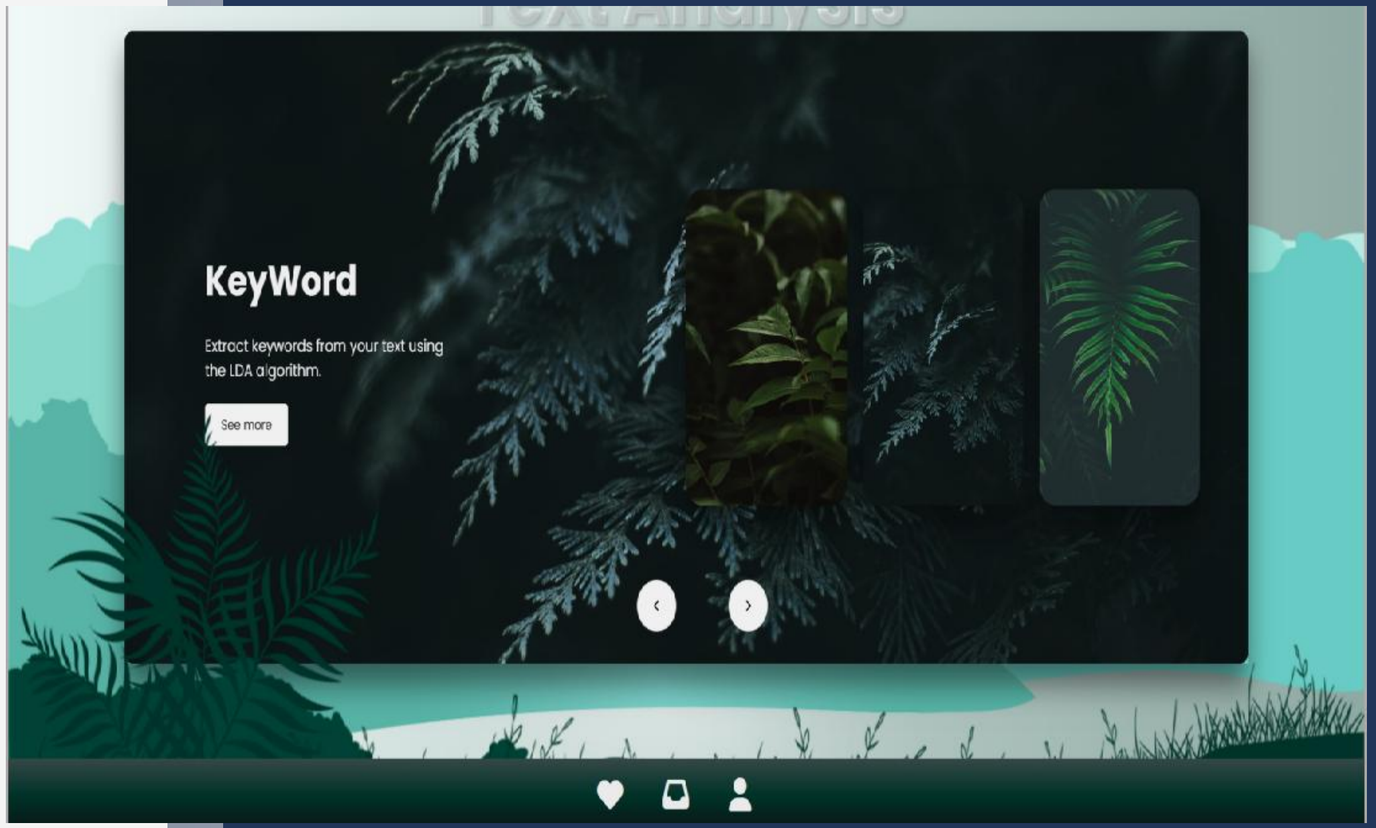


用户进入网站主页，里面有网站功能的介绍和说明，并且提供登录和注册按钮、功能选择按钮。用户可以查看功能说明了解本网站，并且可以点击按钮进行相应的页面跳转。



主页界面

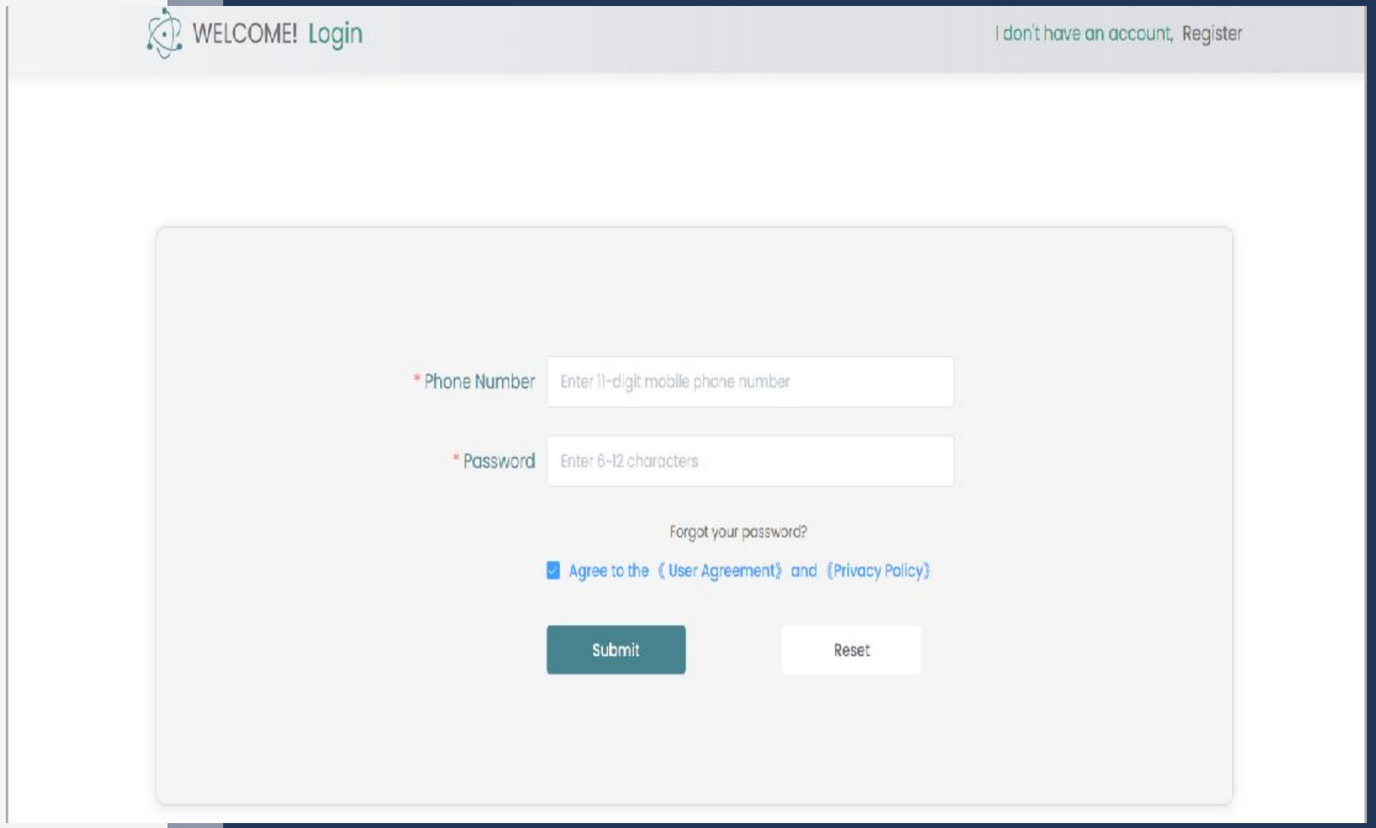
用户进入网站主页，里面有网站功能的介绍和说明，并且提供登录和注册按钮、功能选择按钮。用户可以查看功能说明了解本网站，并且可以点击按钮进行相应的页面跳转。
（图示为主页的网站介绍说明板块）
。



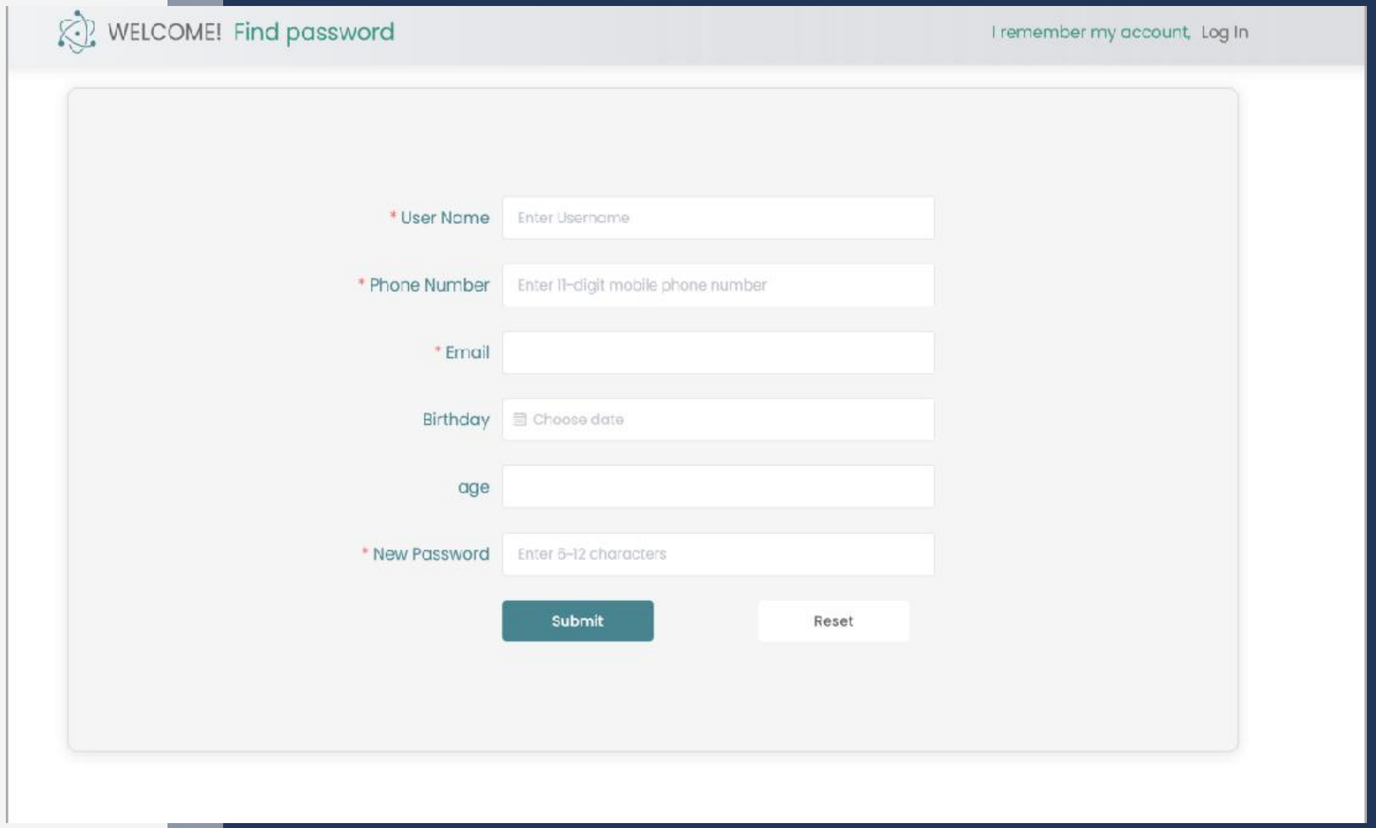
主页界面

用户进入网站主页，里面有网站功能的介绍和说明，并且提供登录和注册按钮、功能选择按钮。用户可以查看功能说明了解本网站，并且可以点击按钮进行相应的页面跳转。
（图示为网站主页的功能选择模块）

登录界面

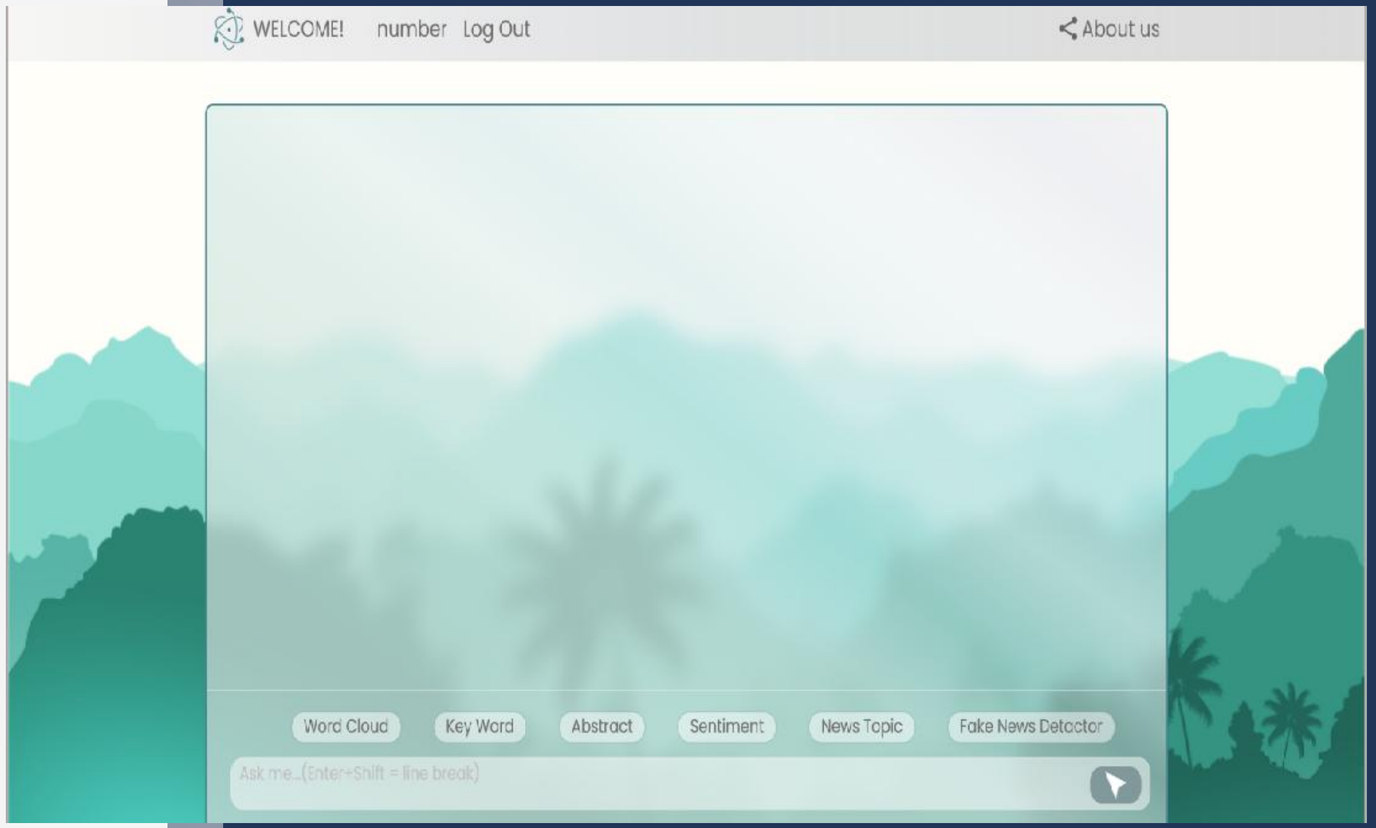


登录界面提供了用户输入已注册账号和密码的功能，用户可以通过提交按钮完成登录操作，同时还可以重置表单、跳转到注册页面、以及找回密码。界面还会对输入的手机号和密码进行简单的格式验证，确保输入符合要求。



找回密码界面

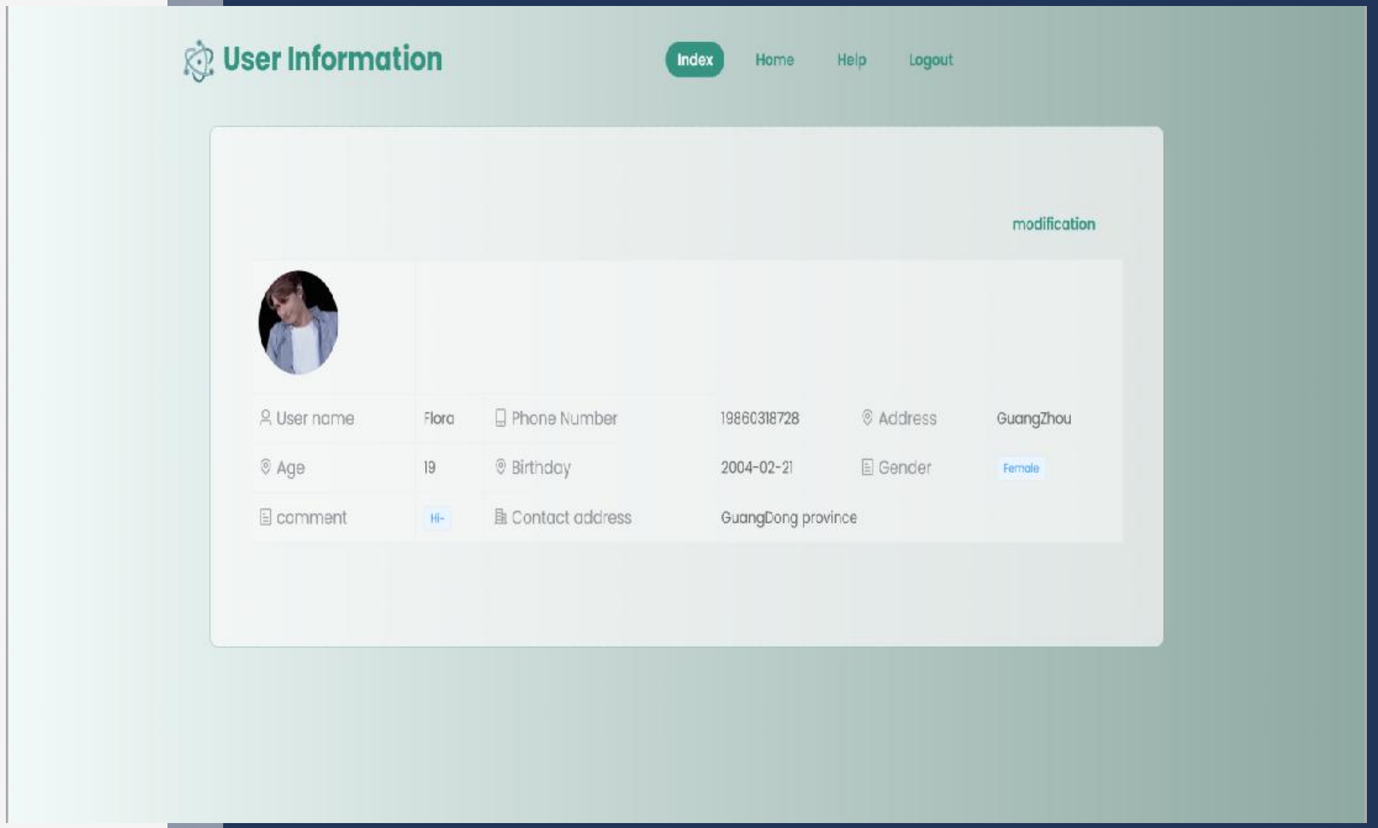
提供了已有账户但忘记密码的用户重新设置密码的功能，用户可以通过输入账号信息进行账号验证，验证通过后进行下一步信息验证，并通过新密码输入框替换旧密码。界面还提供了提交和重置表单的功能，以及跳转到登录页面的链接按钮。



功能界面

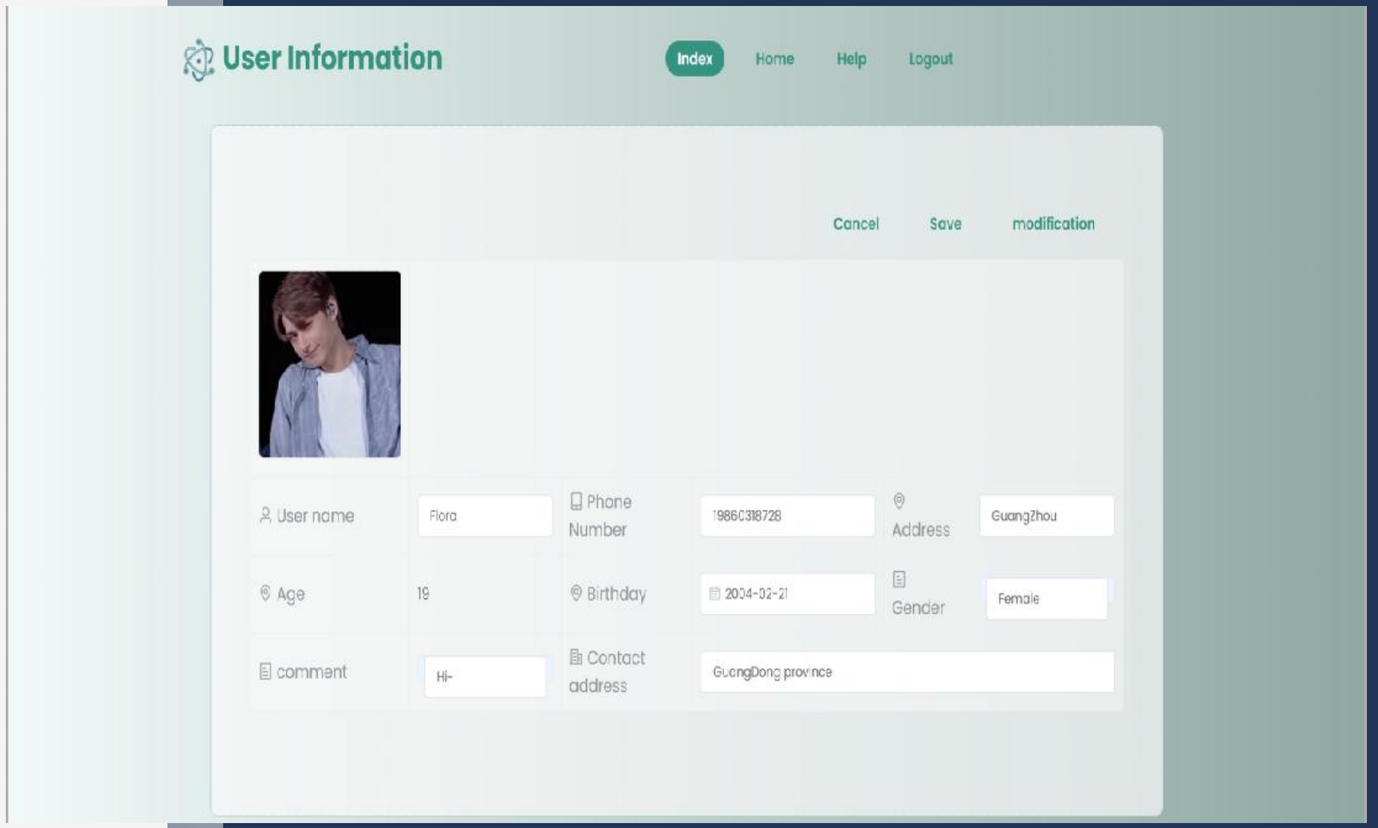
提供了用户进行文本输入并得到分析结果的功能，用户可以通过选择不同的功能按钮（如词云、关键词、摘要、情感分析、新闻主题、假新闻检测）进行文本分析，并通过文本输入框和发送按钮进行交互。界面还提供了注销、用户链接、历史记录链接以及首页链接的跳转按钮。用户选择功能后，经算法分析后返回结果给用户，并存储于数据库；若用户未进行功能选择，直接发送文本，则返回“Please choose a function by clicking the buttons!”（图为名为number用户的登录状况）

用户信息界面



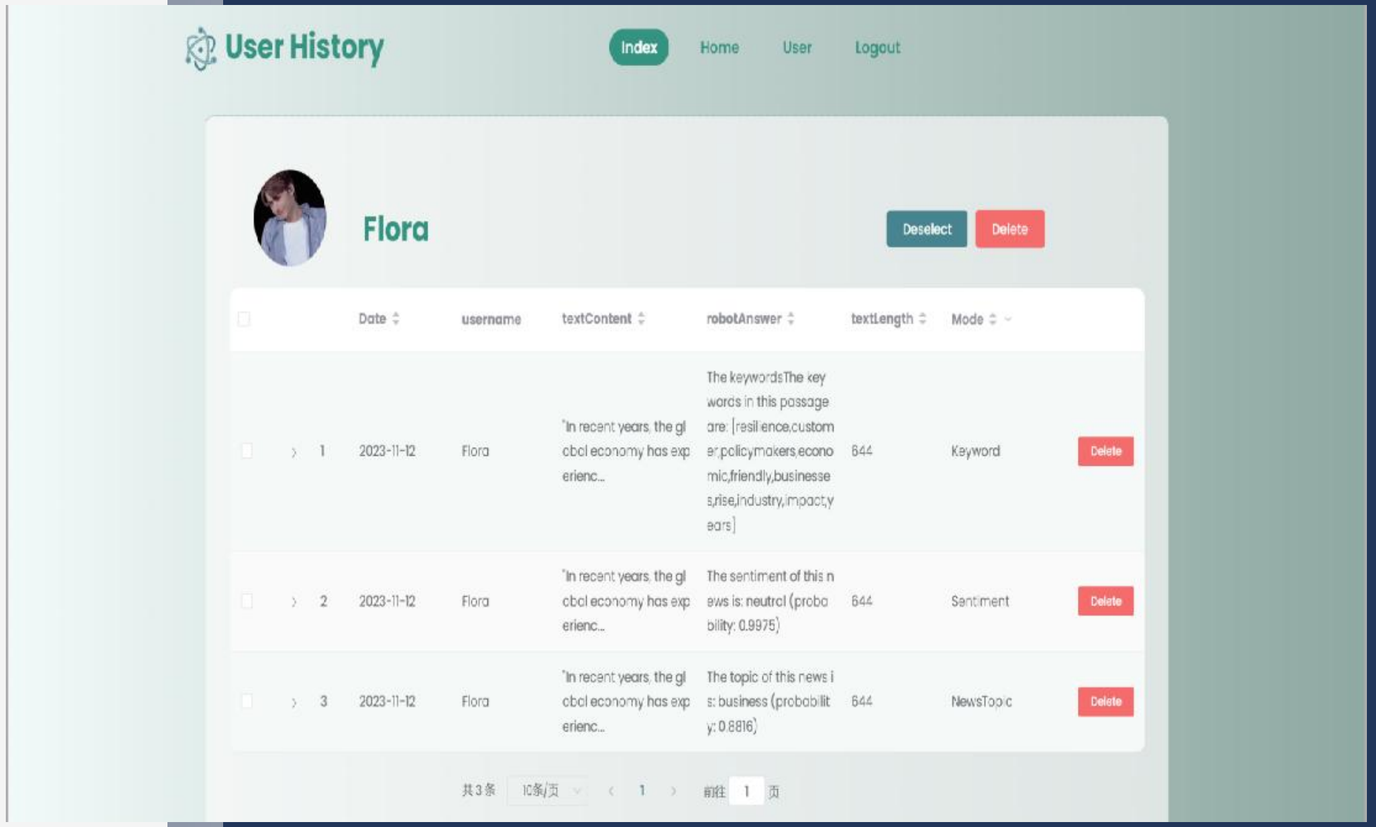
用户信息展示页面“User Information”用于展示用户的基本信息，包括用户名、手机号、地址、年龄、生日、性别、评价、联系地址以及用户头像。页面提供了修改按钮，点击后可以编辑用户信息并保存至数据库，同时也提供了首页、个人主页、历史记录和账户注销按钮，以实现页面间的跳转和用户信息的管理交互。

用户信息界面



如果用户点击modification按钮，可以进行用户信息的修改。修改完成后，点击save按钮可以保存修改，点击cancel按钮修改取消。

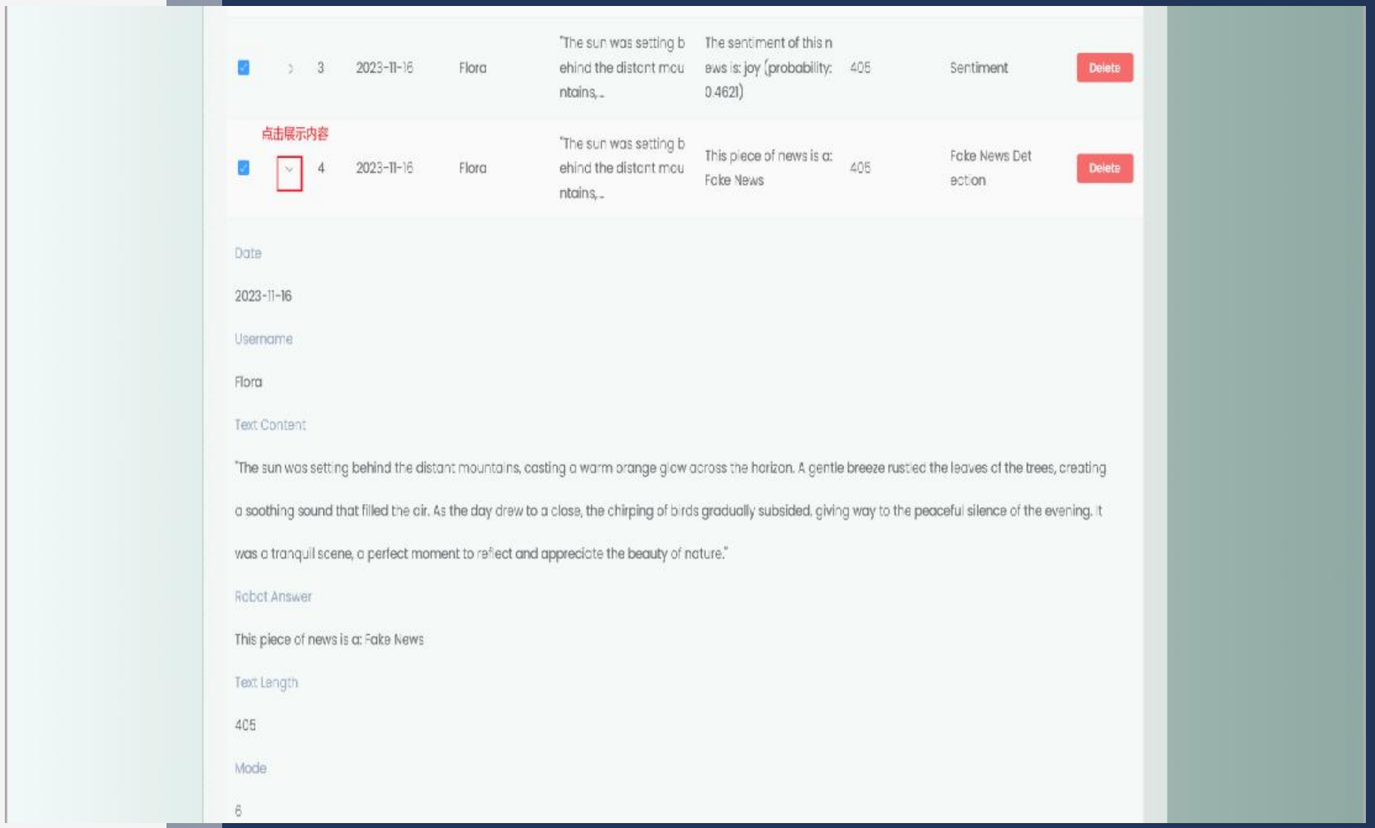
历史记录查询界面

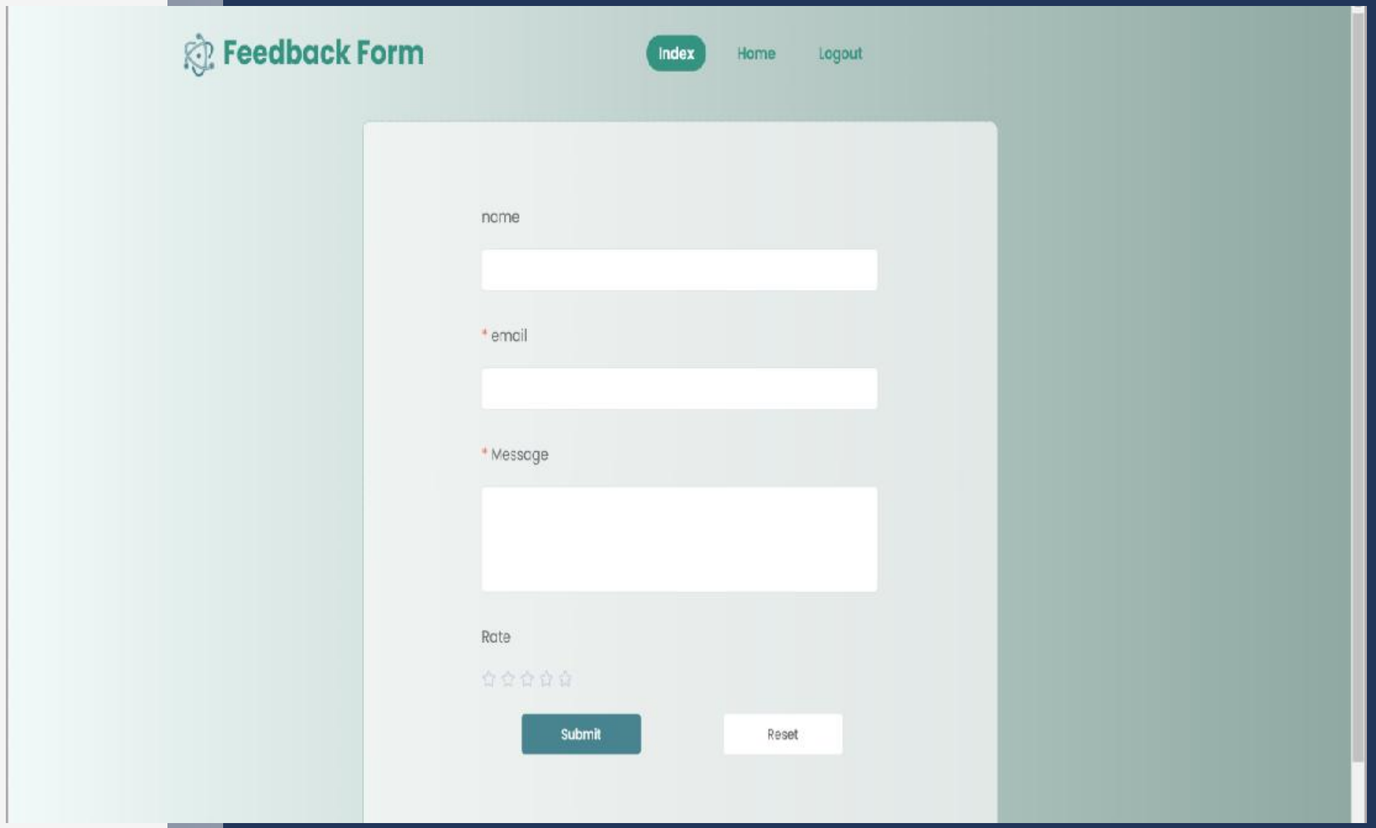


用于展示该用户上传的文本历史记录，包括输入日期、原文、分析结果、文本长度和模式等信息。页面提供了delete按钮和deselect按钮，可以实现单个或多选记录删除功能；同时也提供了首页、个人主页、用户信息和注销按钮，以实现页面间的跳转和用户管理交互。用户可以通过选择页面和页面展示记录的条数进行分页查询，也可以通过选择查询的模式作为条件进行过滤查询。通过点击按钮实现单个记录删除、多选记录删除和登出操作。

历史记录查询界面

用户可以点击箭头展开信息栏
查看详细的历史记录信息





用户反馈界面

用于给用户直接向开发人员提供意见、建议或报告问题。用户可以在这个界面对网站进行评分，通过调查用户的喜好与满意度，可以不断改进网站，加强用户体验。

功能运行结果

词云生成

关键词提取

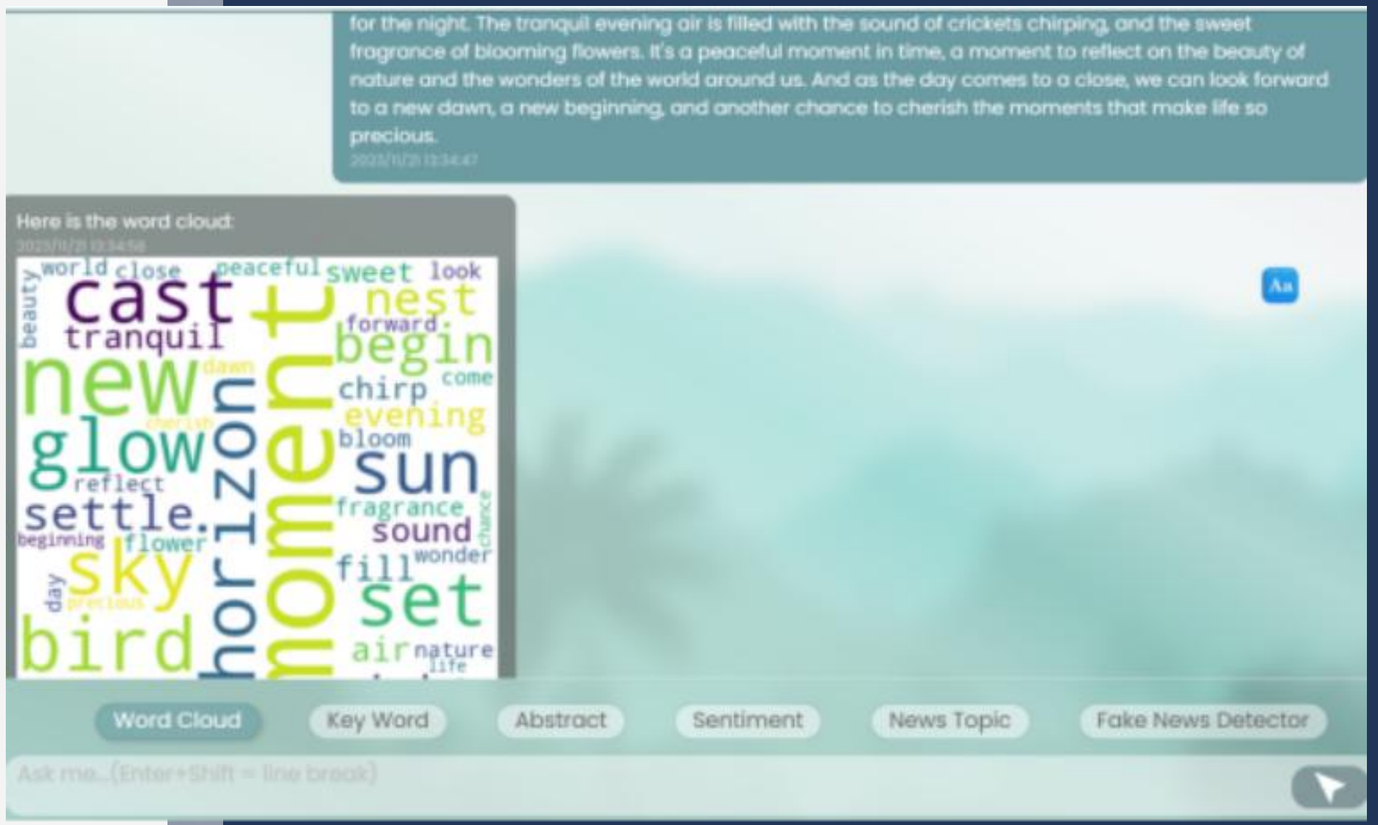
文本摘要

情感分析

主题预测

假新闻检测

词云生成

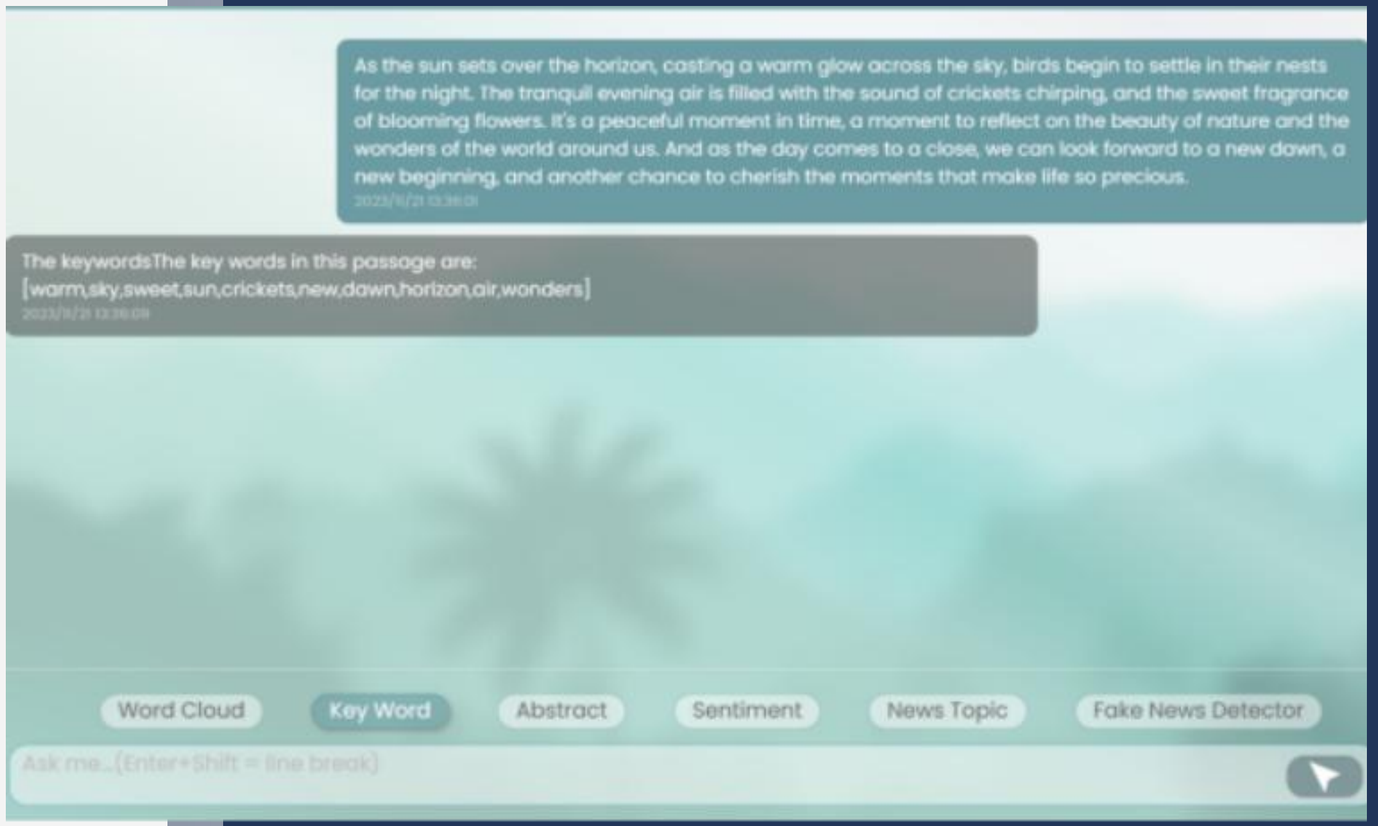


词云生成是一种数据可视化技术，通过视觉展示文本中单词频率分布，生成形象的词云图，使常出现单词在图像中更为突出，展示关键词的重要性。

实现细节：使用WordCloud类，配置参数如图像大小、背景颜色和字体样式。该类生成词云图，词语大小和颜色根据频率调整。生成的图像保存到BytesIO对象，并通过Base64编码发送，实现词云生成和网络传输。

适用于数据可视化、文本分析和关键词提取等场景。

关键词提取

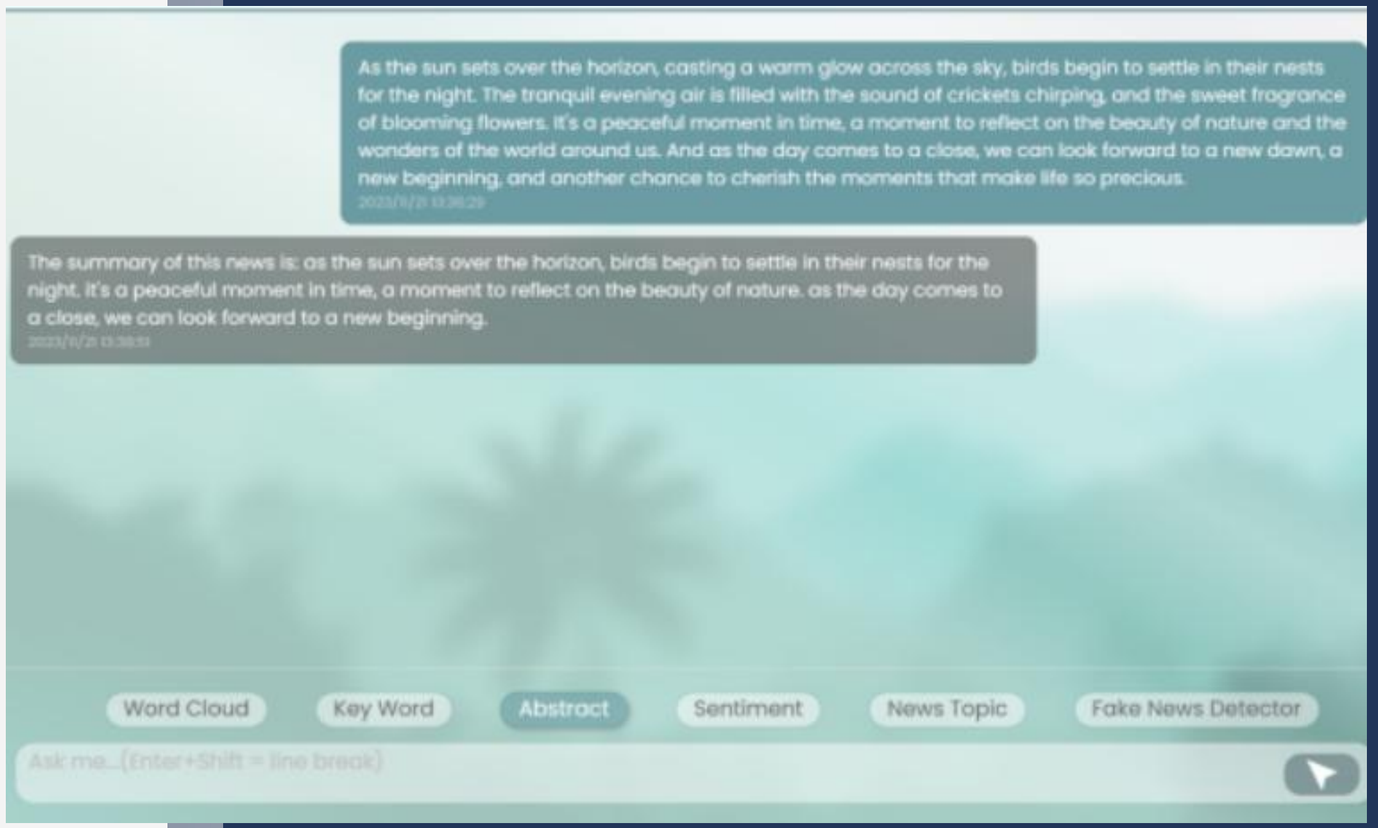


用关键词提取技术，从文本中抽取最具代表性和重要性的关键词，揭示文本的核心主题。

实现细节：利用Spacy库进行文本处理，实现关键词和短语提取。排除停用词和标点，专注于提取特定词性的词汇。通过Flask框架搭建Web服务，用户输入文本即可获取处理后的关键词。

适用于SEO优化、市场研究和内容策略等场景。

文本摘要

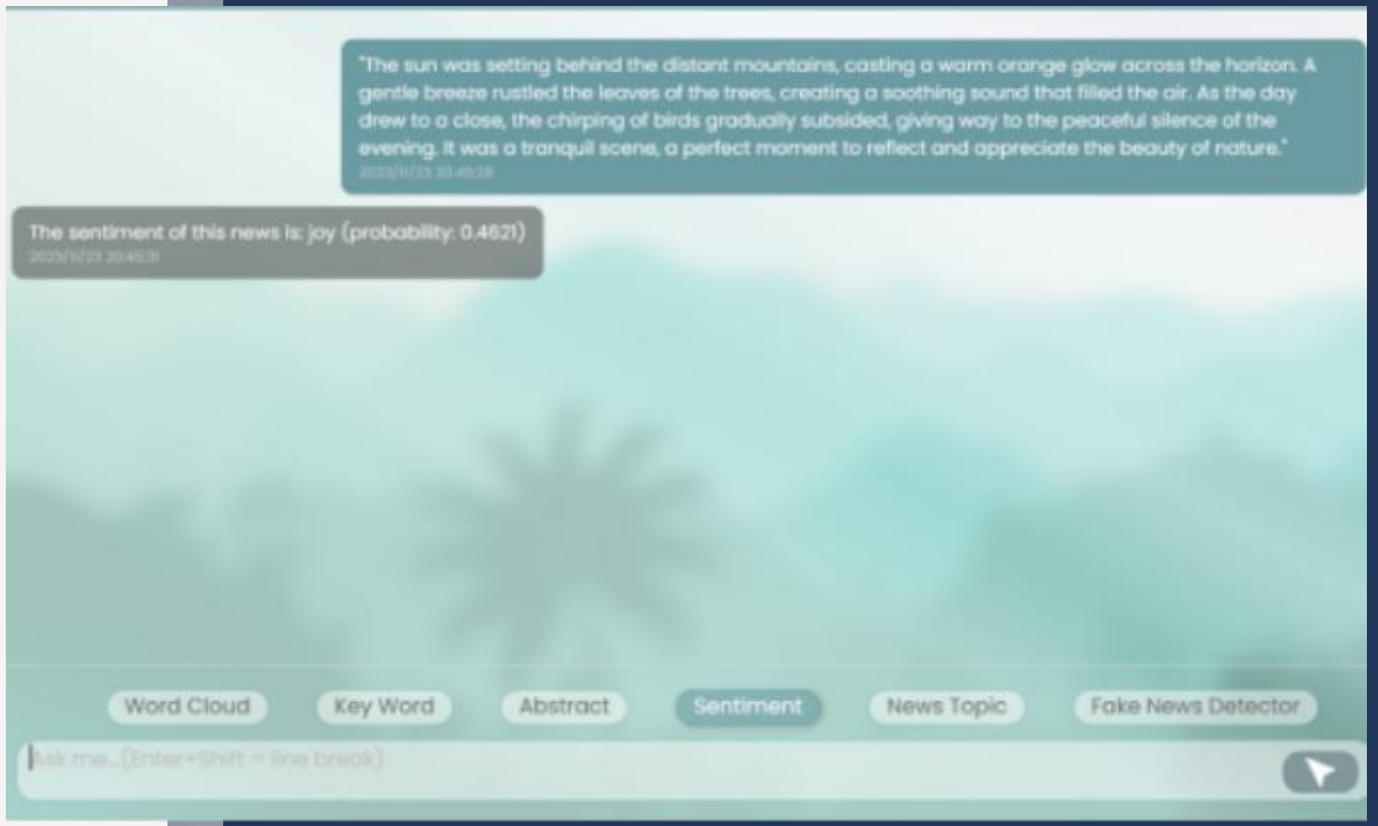


实施文本摘要功能，通过提取文本中的关键信息，生成简明扼要的摘要，帮助用户快速了解文本内容。

实现细节：通过T5模型提取关键信息生成简明摘要，使用Transformers和PyTorch库进行处理，Flask框架实现Web服务返回用户。

适用于新闻、学术文献概要和长文本精炼等场景。

情感分析

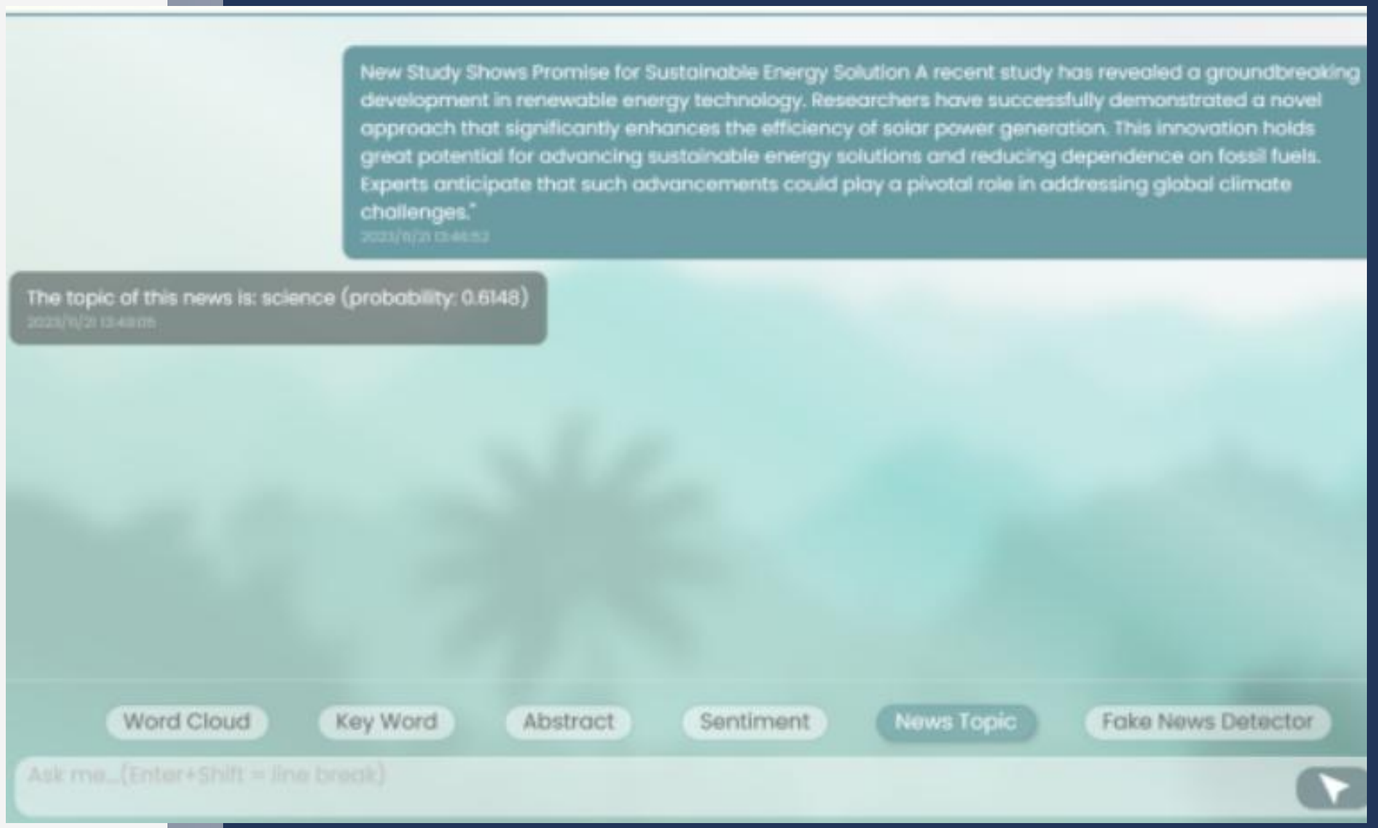


情感分析利用自然语言处理，通过RoBERTa模型判断文本情感，如正面、负面或中性。

实现细节：使用Transformers库提供的RoBERTa分词器编码输入文本，PyTorch库进行情感推断，Flask框架搭建Web服务返回预测结果。

适用于用户反馈、社交媒体情感监测和市场研究等场景。

主题预测

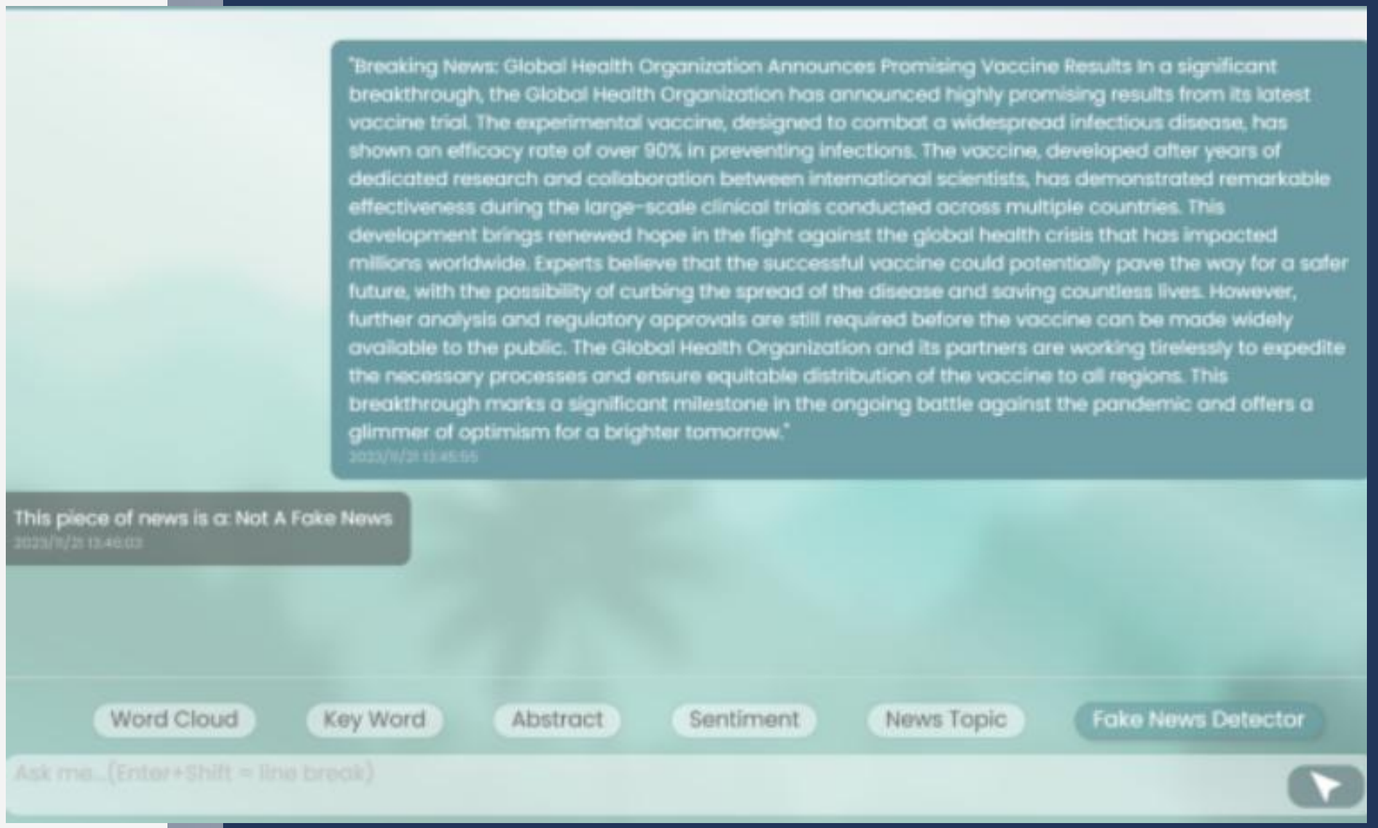


利用主题建模算法，对文本进行分析，预测并识别其中的主题，帮助用户更深入地理解文本背后的内容。

实现细节：使用Transformers库中的BART模型进行自然语言推理，实现文本主题的预测。关键步骤包括关键词提取、主题推断和结果排序。通过结合自然语言推理技术，预测文本与主题的关联，并返回按概率排序的主题列表。

适用于新闻分类、内容管理系统和个性化推荐等场景。

假新闻检测



运用自然语言处理和机器学习技术，对文本进行分析，识别其中是否存在虚假信息，有助于提高文本内容的可信度

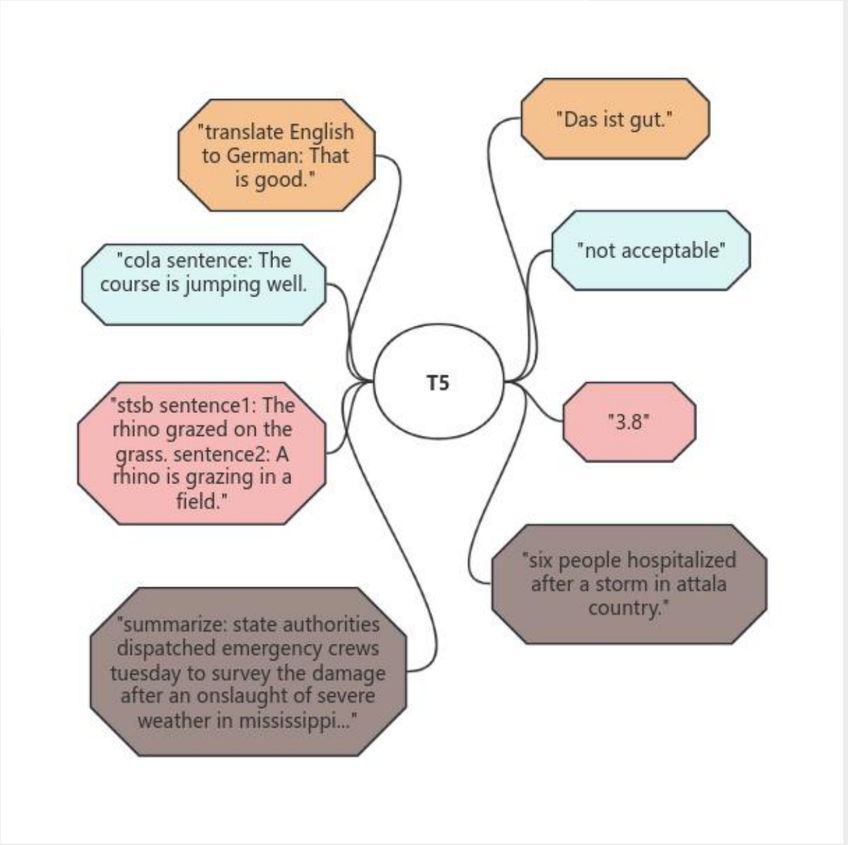
实现细节：采用Sklearn、Pandas、Pickle和Flask构建技术栈，使用随机森林模型进行文本分类，判断虚假新闻。实现包括关键词提取、TF-IDF文本向量化以及基于向量的文本预测与分类。

适用于新闻验证、社交媒体监控和信息验证等场景。

技术介绍

T5模型

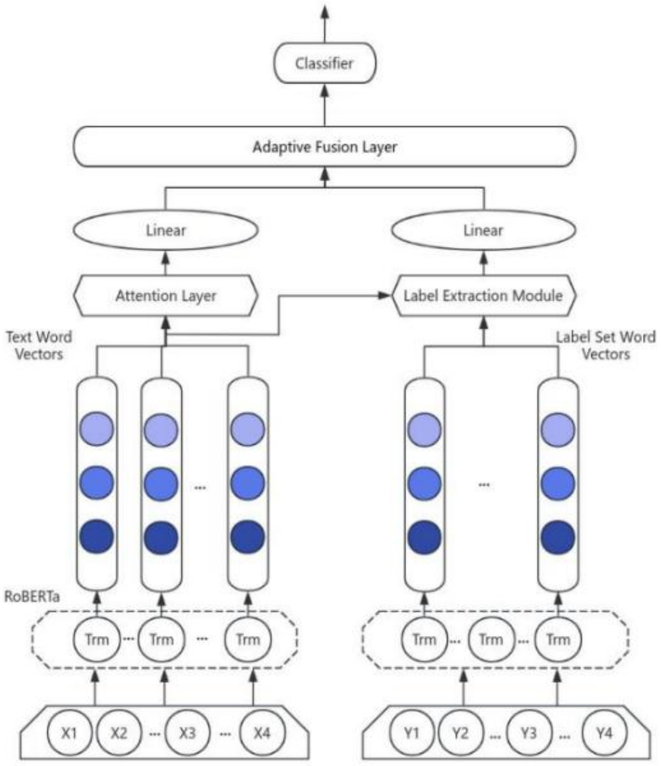
该模型建立在Transformer架构上，专为处理各种文本到文本任务而设计。它采用一种一致的方法来处理不同的自然语言处理任务，依赖于Transformer架构中的注意机制和编码器-解码器结构。关键组件包括Transformer中的自注意机制和位置编码，有助于模型理解文本序列之间的关系和顺序。



技术介绍

ROBERTA

基于BERT（来自Transformer的双向编码器表示），但融入了一系列训练技术和超参数设置以提高性能。
这个系统由五个关键组件组成：
1.词向量生成器：利用技术如Word2Vec、GloVe等将文本转换为向量形式。
2.注意力层：动态分配关注度，提高模型性能，关注不同部分的重要性。
3.标签信息提取模块：从标记数据中提取信息，包括分析、预处理和特征提取，帮助模型理解和预测数据。
4.全连接层：神经网络中连接前一层所有神经元到当前层所有神经元的基本结构，用于特征提取和转换。
5.自适应融合层：通过动态学习权重自动调整融合多个输入或特征的权重，提高模型性能和泛化能力。



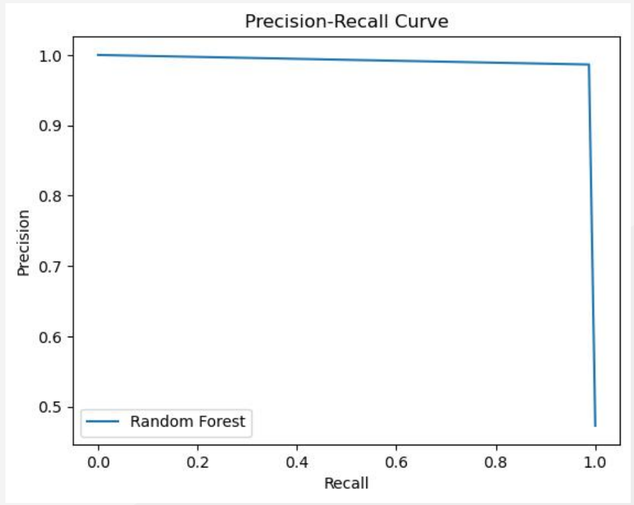
技术介绍

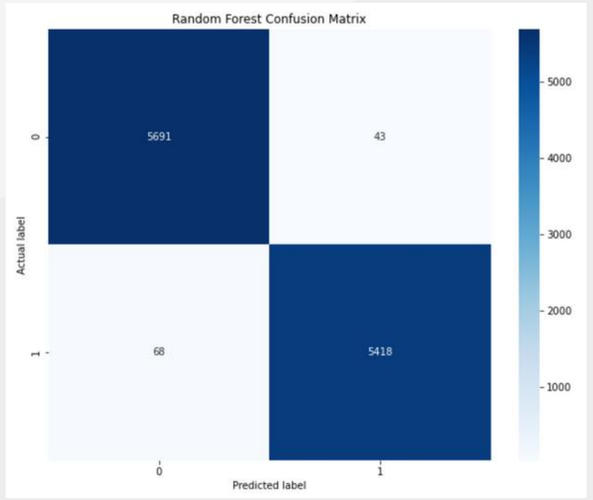
随机森林

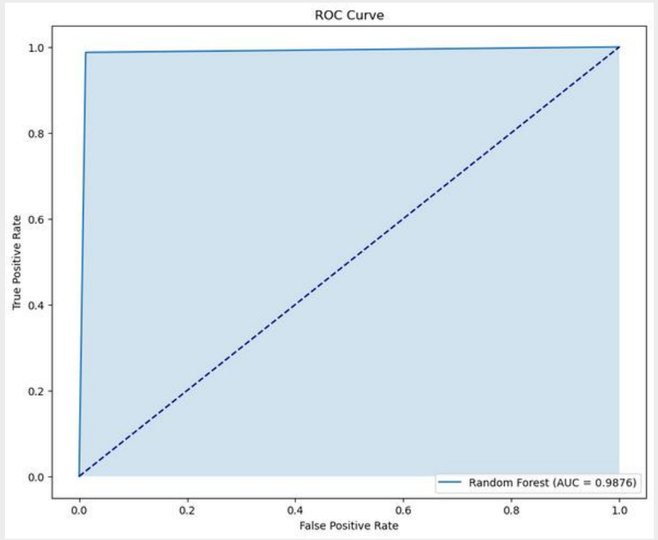
一种基于决策树的集成学习算法，用于提高整体模型性能。每个组件是一个决策树，独立训练以预测结果。

决策树通过递归分割数据集，选择最佳特征和阈值进行纯化。使用信息增益或基尼不纯度进行特征选择。随机森林采用自助聚合（Bagging）方法，通过多次随机抽样创建多个训练数据集。最终模型通过多数投票原则组合各决策树的预测。

适用于新闻验证、社交媒体监控等场景。







facebook/bart-large-mnli

基于BART架构的，并且在MultiNLI数据集上进行了训练。我们用此完成了话题分类功能

引用 API组件

SamLowe/roberta-base-go_emotions

基于 "go_emotions" 数据集进行多标签分类的模型概览。这个模型基于 "roberta-base" 构建，旨在将文本分类为多种情感标签。我们用此完成了情感分析功能

核心团队

方怡萱

后端

杨沣渝

前端

徐子珊

后端

未来计划

第一阶段：
功能完善和性能提升

确保平台在核心功能上稳定可靠，为未来的扩展和优化打下坚实的基础。

第二阶段：
用户体验优化和市场反馈整合

提升用户体验，让平台更加符合用户需求和市场趋势。

第三阶段：
功能细化和目标市场深化

根据特定行业和市场的需求，进一步细化和深化平台功能。

第四阶段：
平台扩展和用户社区建设

扩大平台的功能范围，建立一个活跃的用户社区。

第五阶段：
创新和技术升级

通过引入最新技术，提高平台的智能化水平和技术竞争力。

THANK YOU

We look forward to working with you

CONTACT US


华南师范大学 开发冲冲队


1250885980@qq.com