MDS5119 FINAL PROJECT PROPOSAL — FINANCIAL SENTIMENT ANALYSIS MINI PROGRAM FOR STOCK INVESTORS

Yixin ZHOU
224040090@link.cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong Province, China

1 Introduction

1.1 Core Project Objectives

This project aims to develop a cloud computing-based, stock investor-friendly financial sentiment analysis platform that deeply integrates web crawling, Natural Language Processing (NLP), and sentiment analysis technologies. It will create an all-in-one financial sentiment service system for the general public of investors, focusing on three core functionalities: one-click sentiment information summarization, which can quickly aggregate and extract key insights from multi-platform sentiment content related to specific stocks; professional sentiment analysis, which employs advanced algorithms to judge the emotional tendency of sentiment content and generate intuitive sentiment trends; and bundled download of research report PDFs, enabling users to batch acquire authoritative financial research reports for local storage. This platform will effectively address the inefficiencies and operational difficulties faced by ordinary investors in accessing and analyzing financial sentiment.

1.2 Market Gap Positioning

The current market for financial sentiment tools exhibits a significant mismatch between supply and demand, failing to meet the actual needs of ordinary stock investors in three key aspects. Firstly, the technical threshold is excessively high: most existing open-source sentiment tools require users to master specialized skills such as Python programming, Selenium automation testing, and MongoDB database management [7], while ordinary investors generally lack relevant technical backgrounds, making it impossible for them to use these tools effectively. Secondly, functionalities are fragmented: most tools only offer single functions, either limited to sentiment data crawling or simple sentiment analysis, lacking end-to-end coverage of the "data collection - in-depth analysis - result output" process. Investors must use multiple tools simultaneously to complete sentiment analysis, resulting in cumbersome operations. Finally, commercial services are prohibitively expensive: professional financial data platforms targeting institutional users, such as East Money Choice and Wind Financial Terminal, typically charge annual fees exceeding 10,000 RMB, far beyond the affordability of ordinary investors. This prevents ordinary investors from accessing high-quality sentiment analysis services. By leveraging technological optimization and model innovation, this project will fill the market gap for stock investor-friendly financial sentiment analysis tools.

2 Background and Motivation

2.1 Analysis of Current Market Pain Points

The current market for financial sentiment tools faces numerous pain points that severely hinder ordinary investors' effective utilization of sentiment information. Regarding technical barriers, open-source projects impose high technical requirements on users. For instance, the EastMoney_Crawler project on GitHub [7] only provides script files for crawling stock bar data; users must master Python programming and database operations to run the scripts. Furthermore, the project lacks a visualization interface, requiring data inspection and analysis to be performed through code, which is

inaccessible to ordinary investors. In terms of functional completeness, most tools only support single functionalities and lack end-to-end service capabilities. For example, the stock comment analysis project launched by Lianxianghui can only score the sentiment of stock comments but does not support research report downloads [5]; in contrast, some information crawling scripts shared on Zhihu [6] can only crawl and store news and research reports without sentiment analysis capabilities. This forces users to switch between multiple tools, leading to low operational efficiency. Concerning service costs, commercial platforms primarily target institutional users with pricing far beyond the means of ordinary investors. East Money Choice charges an annual fee exceeding 10,000 RMB, while Wind Financial Terminal's annual fee reaches over 30,000 RMB [1, 3]. Unable to afford such high costs, ordinary investors can only rely on free but fragmented sentiment channels with questionable accuracy.

2.2 Project Vision

The core vision of this project is to provide ordinary stock investors with convenient, efficient, and low-cost financial sentiment analysis services, promoting the inclusiveness of financial information services. In terms of reducing operational barriers, the project will adopt a cloud-based WeChat Mini Program to achieve "zero-programming operation." Users do not need to master any professional technologies; simple operations on the mini program interface will suffice to meet their sentiment analysis needs, effectively covering investor groups without technical backgrounds. In functional integration, the platform will integrate end-to-end services including "sentiment collection + sentiment analysis + research report download + visualization reporting." Users will no longer need to switch between multiple tools, as all operations from sentiment acquisition to analytical application can be completed within a single platform, significantly improving operational efficiency. In cost optimization, the project will leverage open-source technologies and cloud computing architectures to reduce system development and operational costs. Additionally, it will adopt a flexible pricing strategy to launch cost-effective service plans, enabling ordinary investors to access high-quality sentiment analysis services at a low cost and breaking the monopoly of commercial platforms on professional financial information services.

3 Technical Approach

3.1 Data Collection Module

As the core of the platform for acquiring sentiment information, the data collection module will focus on covering mainstream financial information platforms to ensure data comprehensiveness and timeliness. In terms of data source selection, it will primarily include East Money Stock Bar and News Platform. As an important communication venue for investors, East Money Stock Bar contains a wealth of authentic investor opinions and emotional feedback, from which post content and comment information will be crawled. The East Money News Platform will serve as the source for research reports and professional news, enabling the crawling of various authoritative research report PDFs and the latest financial news. In technical implementation, Selenium will be selected for crawling dynamic pages on East Money to ensure the acquisition of complete page data, while the Requests tool will be used for data requests on static pages such as research report PDFs to improve crawling efficiency. Regarding data storage, a phased plan will be implemented: the first phase will adopt local storage, saving crawled data in CSV or JSON formats locally to facilitate development and testing; the second phase will migrate data storage to a MongoDB database, leveraging MongoDB's advantages in unstructured data storage to better manage massive volumes of sentiment data.

3.2 Natural Language Processing (NLP) Module

As the key to realizing sentiment analysis and sentiment summarization, the NLP module will adopt a layered processing approach to ensure the accuracy and practicality of analytical results. In the text preprocessing phase, word segmentation, stopword removal, and Named Entity Recognition (NER) will be sequentially performed. Word segmentation will use the Jieba tool combined with a professional financial vocabulary library to improve segmentation accuracy; stopword removal will be based on a custom financial stopword list to filter meaningless vocabulary and reduce data interference; NER will focus on identifying stock codes and company names in the text to lay the foundation for subsequent stock-specific sentiment analysis.

In the sentiment analysis phase, a dual-model collaborative approach will be adopted to respectively generate sentiment scores for individual stock comments and comprehensive sentiment analysis reports. The SKEP model [2], an open-source sentiment pre-trained model developed by Baidu, will be used to score the emotional tendency of each stock comment, obtaining positive, negative, or neutral orientations for individual comments due to its high accuracy in Chinese text sentiment analysis. The FinGPT model [4], a specialized large language model for the

financial field, will integrate sentiment scores from all stock comments with research report and news information to generate comprehensive sentiment analysis reports. These reports will include overall market sentiment trends, key influencing factors, potential risks, and opportunities, providing users with decision-making references.

In the visualization phase, analytical results will be intuitively presented through various charts: the WordCloud library will generate word clouds of high-frequency keywords to help users quickly grasp sentiment hotspots; the Matplotlib tool will create sentiment distribution histograms to display the proportion of stock comments with different emotional tendencies; simultaneously, the K-Means algorithm combined with TF-IDF feature extraction technology will perform topic clustering on sentiment content, grouping similar thematic sentiment information to facilitate user classification and in-depth analysis.

3.3 Cloud Computing Architecture

The cloud computing architecture will be designed around the WeChat Mini Program to ensure system stability, scalability, and user experience. On the front-end, the WeChat Mini Program will be adopted as the core form, leveraging its advantages of a large user base, no requirement for download and installation, and convenient operation to effectively lower user access barriers. The front-end interface will consist of three core pages: the homepage (displaying platform function entrances and popular sentiment information), the analysis page (providing sentiment analysis parameter settings and result display), and the download page (supporting research report PDF viewing and downloading). Additionally, visualization charts such as sentiment trend graphs and word clouds will be integrated into the front-end to ensure intuitive presentation of analytical results.

On the back-end, the Flask framework will be used for back-end service development, offering lightweight flexibility to facilitate rapid development and deployment. REST API interfaces will be encapsulated to realize data interaction between the front-end and back-end, including interfaces for triggering data collection, returning sentiment analysis results, and downloading research reports. Furthermore, user permission management functionality will be developed to distinguish between free and paid users, providing differentiated services.

On the database layer, Tencent Cloud Base's cloud database and cloud storage will be adopted as the core database to meet the storage requirements of sentiment data. Three types of data will be primarily stored: crawled raw sentiment data (including stock bar posts, comments, and news content), sentiment analysis results (including sentiment scores for individual comments, comprehensive sentiment analysis reports, and clustering labels), and user data (including basic user information, usage history, and payment status). Comprehensive data indexes will be established to improve data query efficiency.

For cloud service deployment, Tencent Cloud Base (CloudBase) will be selected as a WeChat Mini Program-friendly free cloud platform. Specifically designed for WeChat Mini Program development, it provides free basic resource quotas, including cloud functions, database storage, and static website hosting, which can meet the initial deployment needs of the project. Additionally, AWS cloud services (e.g., elastic computing and CDN acceleration) will be considered to ensure stable system operation. A comprehensive data backup strategy will be implemented, with daily automatic backups to ensure data security.

4 Competitive Analysis and Benchmarking

4.1 Comparison with Open-Source Projects

Current open-source financial sentiment-related projects in the market have obvious deficiencies in functional coverage and user-friendliness, failing to meet the needs of ordinary stock investors. The zcyeee project only supports crawling stock comment data from East Money Stock Bar, lacking sentiment analysis and research report download functionalities. Although it supports MongoDB database storage, it has no visualization interface, requiring users to operate through programming with an extremely high access threshold. The information crawling script shared by Wang Yutao can crawl news and research reports and support PDF downloads but similarly lacks sentiment analysis capabilities; data can only be stored locally without database management, and users need a certain level of programming foundation to use it. The Lianxianghui stock comment analysis project focuses on sentiment scoring using the SKEP model and supports the acquisition of minute-level stock comment data, but it does not include research report download functionality, database support, or visualization display, with user operations also relying on programming.

4.2 Comparison with Paid Software Services

Commercial financial service providers in the market charge high prices, making them unsuitable for investors with small capital scales and a limited number of focused stocks. iFinD's iAskCai provides sentiment analysis and research report downloads with paid subscriptions (monthly fees ranging from 30 to 100 RMB) but features closed functionalities that do not support custom analysis. East Money Choice, a professional financial data platform, offers sentiment analysis, research reports, and sentiment analytics with an annual fee of over 10,000 RMB, targeting institutional users. Wind Financial Terminal provides high-end financial data services, including sentiment monitoring, research reports, and sentiment analysis, with an annual fee exceeding 30,000 RMB, primarily serving securities firms, fund companies, and other institutions.

In contrast, this project demonstrates significant advantages across phases. In the MVP phase, it already covers multi-source data including stock bars, news, and research reports, adopts dual-model collaboration (SKEP and FinGPT) for sentiment analysis, supports research report downloads, stores data in local CSV files, and provides visualization functionalities such as word clouds and histograms. Although basic operations are still required, the technical threshold has been significantly reduced. In the cloud deployment phase, the project will achieve comprehensive coverage of multi-platform data sources, further optimize sentiment analysis models, support bundled research report downloads, enrich visualization dimensions, and most importantly, realize one-click operation through the WeChat Mini Program, completely eliminating technical barriers and truly achieving stock investor-friendliness. Comparisons with open-source projects confirm that this project holds distinct advantages in functional integration and user experience, effectively filling the market gap.

5 Implementation Plan

5.1 Phase 1: Local Minimum Viable Product (MVP)

The first phase focuses on constructing a Local Minimum Viable Product (MVP) with a planned duration of 28 days, prioritizing the development and testing of core functionalities to ensure the usability and stability of basic features. In the data collection development task, scripts for crawling stock bar comments will be written, leveraging Selenium to crawl dynamic pages on East Money Stock Bar and ensuring the acquisition of post content, comment information, publication times, and other key data. Subsequently, functionalities for crawling news and research report PDFs will be expanded to realize the crawling of news content from financial news platforms and the saving of research report PDFs. Simultaneously, local data storage functionalities will be developed to support CSV and JSON formats, ultimately outputting a `crawler.py` script and a data folder for storing raw crawled data.

The NLP processing development task will focus on the dual-model collaborative sentiment analysis functionality. First, a text preprocessing module will be developed, integrating the `Jieba` word segmentation tool with a custom professional financial vocabulary library to optimize segmentation accuracy. A financial stopword list will be constructed to automatically filter meaningless vocabulary through code logic, and Named Entity Recognition (NER) functionality will be debugged to ensure accurate extraction of stock codes and company names from text. On this basis, the SKEP model will be integrated and parameter-optimized to automatically calculate sentiment scores for individual stock comments. Simultaneously, local deployment of the FinGPT model will be completed, and model calling interfaces will be written to realize integrated analysis logic for multi-source data (stock comments, news, and research report abstracts). Finally, an `nlp_process.py` script and a `score.csv` sentiment score table will be output, with the latter including core fields such as comment content, emotional tendency (positive/negative/neutral), and sentiment score (0-10 points).

In the visualization development and testing task, visualization scripts will be designed using the `WordCloud` library to generate high-frequency keyword word clouds, adjusting color schemes and layouts to align with financial sentiment characteristics. The `Matplotlib` tool will be used to construct sentiment distribution histograms, supporting the display of proportional changes in stock comments with different emotional tendencies over time (e.g., daily/weekly). Additionally, visual presentation of K-Means clustering results will be implemented, displaying the distribution of different thematic sentiment information in scatter plots. A `visualization.py` script and an output folder will be output, containing word cloud images, histograms, and clustering analysis charts. Finally, comprehensive testing and debugging will be conducted: first, verifying the accuracy of the data collection module by comparing crawled data with original webpage content to ensure data accuracy; second, conducting consistency testing of the sentiment analysis module by comparing manually annotated emotional tendencies of 50 stock comments with model output results and

correcting model misjudgment logic; finally, integrating all module functionalities to write a `main.py` script, realizing one-click operation of "data collection - sentiment analysis - visualization output," and outputting a `test_report.md` report documenting the testing process, encountered issues, and solutions.

5.2 Phase 2: Cloud Deployment and Mini Program Development

The second phase is scheduled for 28 days, with core objectives including migrating local functionalities to the cloud, developing the front-end of the WeChat Mini Program, writing a project report, and realizing cloud accessibility and user access to the platform.

In the WeChat Mini Program development task, the front-end interface will first be designed and developed, constructing three core pages based on the native WeChat Mini Program framework (WXML/WXSS/JS). Finally, functional testing of the mini program will be conducted, inviting 20-30 ordinary stock investors to participate in user experience testing, collecting feedback on interface operations, functional usage, response speeds, and other aspects. Interface interaction logic and loading speeds will be optimized to ensure compatibility across different mobile phone models, ultimately outputting a deployable WeChat Mini Program code package.

Finally, a project report will be completed, and the platform functionalities will be finalized and adjusted.

6 Conclusion

The core competitiveness of this project lies in four aspects: user-friendliness, functional comprehensiveness, cost-effectiveness, and scalability, enabling it to provide differentiated financial sentiment analysis services to ordinary stock investors. In terms of user-friendliness, the project adopts the WeChat Mini Program as the front-end carrier, allowing users to access services directly through WeChat without download or installation. The operational process is designed to be concise and intuitive, supporting "one-click trigger analysis + result push." Users do not need to master any programming technologies or professional knowledge, enabling them to easily meet their sentiment analysis needs. This completely addresses the pain point of ordinary investors being "unable to use professional sentiment tools" and effectively expands the user coverage.

In terms of functional comprehensiveness, the project integrates end-to-end services including "sentiment collection - sentiment analysis - visualization display - research report download." Compared with fragmented competitors in the market, users no longer need to switch between multiple tools, as all operations from sentiment acquisition to analytical application can be completed within a single platform. Simultaneously, the dual-model collaborative sentiment analysis approach not only achieves accurate scoring of individual stock comments but also generates comprehensive overall sentiment analysis reports. The diverse visualization presentation methods meet users' analytical needs at different levels, with functional completeness far exceeding similar tools.

References

[1] Ltd. East Money Information Co. *Introduction to Choice Data Services*. `https://choice.eastmoney.com/`. Accessed: [2025-10-19]. 2025.

T. Hao et al. "SKEP: Sentiment Knowledge Enhanced Pre-Training for Sentiment Analysis". In: *arXiv Preprint* (2020). arXiv: `2005.05635 [cs.CL]`.

iFinD. *Introduction to iAskCai Functions*. `https://www.10jqka.com.cn/`. Accessed: [2025-10-19]. 2025.

Y. Liang et al. "FinGPT: Enhancing Sentiment-Based Stock Movement Prediction with Dissemination-Aware and Context-Enriched LLMs". In: *arXiv Preprint* (2024). arXiv: `2412.10823 [q-fin.ST]`.

Lianxianghui. *Stock Comment Sentiment Analysis*. `https://www.lianxh.cn/details/440.html`. Accessed: [2025-10-19]. n.d.

Y. Wang. *How to Batch Download Research Reports from East Money Using Python?* Zhihu Column. Accessed: [2025-10-19]. Oct. 2022. URL: `https://zhuanlan.zhihu.com/p/558497407`.

zcyeee. *EastMoney_Crawler*. GitHub Repository. Accessed: [2025-10-19]. n.d. URL: `https://github.com/zcyeee/EastMoney_Crawler`.