National Research University Higher School of Economics
Faculty of Computer Science
Bachelor's Programme Applied Mathematics and Information Science
01.03.02 Applied Mathematics and Information Science

Internship Report

Fulfilled by:

Group: 2110

Todorov Denis Andreevich

(signature)

Supervised by:

Senior Software Developer, Google UK Limited

Zakharov D. S.

(signature)

Moscow, 2025

AI-Driven Financial and Real Estate Investment Recommendation Platform

I. Goals and Objectives of the Internship

The primary objective of the internship was to lay the theoretical and practical groundwork for the successful implementation of the bachelor's final project. The internship period was dedicated to both deep research and initial system prototyping, with the ambition to achieve a well-structured, modular, and scalable solution capable of providing investment recommendations based on real-time financial data and sentiment analysis.

This project addresses a critical problem faced by modern investors: the complexity of processing massive volumes of both structured and unstructured financial data. While traditional quantitative models focus on historical price movements and technical indicators, they often fail to incorporate the crucial dimension of market sentiment, which significantly influences financial decision-making in today's rapidly evolving landscape. The recent emergence of Natural Language Processing (NLP) and Reinforcement Learning (RL) technologies offers an opportunity to enrich financial forecasting with sentiment-based signals and adaptive decision-making agents [2].

As a conclusion, the internship's mission was framed around two key areas:

1) **Exploratory Research** — focused on identifying, understanding, and selecting appropriate machine learning techniques, data sources, and software tools
2) **Practical Prototyping** — aiming to build, test, and evaluate the initial components of the advisory platform, which includes a microservices-based backend, machine learning models for sentiment analysis and prediction, and a frontend interface for user interaction

A. Specific Objectives

The objectives of the internship were systematically broken down into the following tasks:

1) **Comprehensive Literature Review and Problem Domain Analysis**:
   - Study state-of-the-art approaches to financial market prediction and investment decision support systems
   - Investigate the role of NLP models, particularly transformer-based architectures like BERT and FinBERT, in financial sentiment analysis
   - Explore the application of Reinforcement Learning methods (such as Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO)) in automated trading strategies
   - Analyze the strengths and limitations of integrating both supervised learning and reinforcement learning models in the context of financial forecasting
2) **Design of System Architecture and Component Selection**:
   - Define the architectural vision for the advisory platform, with an emphasis on modularity, scalability, and maintainability
   - Choose appropriate technologies for backend, frontend, and machine learning components
3) **Prototype Development of Core Modules**:
   - Implement backend infrastructure supporting user management and integration with the ML module
   - Develop the initial version of the sentiment analysis component using FinBERT
   - Conduct early-stage experiments with predictive models for time-series-based financial forecasting
4) **Performance Evaluation and Validation of Selected Approaches**:
   - Define performance metrics (accuracy, F1-score, Sharpe ratio, backtesting results)
   - Conduct comparative experiments on selected models using real-world financial datasets
   - Perform backtesting of forecasting models on historical market data
5) **Definition of a Roadmap for Further Development and Integration**:
   - Formulate a clear plan for subsequent project phases, including ML module integration, conversational AI bot development, and reinforcement learning agent implementation

B. Skills Development and Learning Outcomes

In addition to the technical objectives, the internship also targeted the development of the following professional competencies:

- Proficiency in modern AI and ML technologies with a focus on their application in financial contexts
- Experience with software architecture design, API development, and inter-service communication using microservices
- Ability to conduct quantitative analysis, including model training, evaluation, and backtesting

II. Problem Statement and Relevance of the Topic

The process of making informed investment decisions has become increasingly challenging due to the exponential growth of financial data originating from diverse sources. Market participants are confronted with not only structured numerical data, such as historical price movements, trading volumes, and economic indicators, but also a growing body of unstructured textual data, including financial news, analyst reports, and social media discussions. These heterogeneous data streams, while rich in potential insights, pose significant difficulties in terms of efficient aggregation, processing, and analysis.

Traditionally, quantitative investment strategies have relied on statistical models and machine learning approaches trained on structured data. While such models have achieved notable success in forecasting price trends and evaluating market risks, they often fail to capture the emotional and psychological factors driving investor behavior. This gap is particularly evident during periods of high market volatility or speculative bubbles, where sentiment-driven dynamics can significantly impact asset prices.

In recent years, platforms like Twitter and TradingView have emerged as influential spaces where market participants express opinions, share analyses, and react to unfolding events [9]. The aggregation of such sentiment signals has been shown to correlate with short-term stock price movements [5], making social sentiment analysis a promising complement to traditional forecasting techniques. However, most existing investment advisory tools do not fully exploit the potential of these unstructured data sources, focusing instead on conventional quantitative indicators.

The primary challenge addressed in this project is the development of an intelligent, AI-driven system capable of integrating both structured financial data and unstructured textual sentiment data into a unified decision-support framework [6]. The core idea is to leverage recent advancements in **Natural Language Processing (NLP)** — particularly transformer-based models such as BERT and its financial adaptation FinBERT — to extract actionable insights from textual data, and to combine these insights with time-series forecasting models.

Furthermore, the project envisions the application of **Reinforcement Learning (RL)** agents as a future extension of the system. RL agents are capable of learning dynamic trading strategies by interacting with the market environment in a trial-and-error fashion, gradually improving their performance through feedback mechanisms. The potential integration of RL into the system is aimed at automating portfolio management, allowing the agent to adapt to changing market conditions without relying on fixed rule-based strategies.

A. Key Problems Addressed

The internship project focuses on the following key problem areas:

- **Heterogeneity of Data Sources**: Financial decision-making involves the simultaneous analysis of structured numerical data and unstructured textual data. Combining these heterogeneous sources in a meaningful way presents methodological challenges that require careful model design and data engineering solutions
- **Sentiment Extraction and Analysis**: Financial text often contains jargon, abbreviations, and sarcasm, making accurate sentiment classification a non-trivial task. General-purpose NLP models tend to underperform in financial contexts without domain-specific fine-tuning
- **Scalability and Real-Time Processing**: To be effective, the system must process data streams in real-time or near-real-time, enabling timely insights and rapid reaction to market developments. This requirement imposes constraints on model latency and infrastructure design
- **Trustworthiness and Explainability**: Investment recommendations must be interpretable and transparent to gain user trust. Black-box machine learning models, while potentially powerful, **can lack the explainability needed for financial applications** where regulatory compliance and decision justification are crucial [13]
- **Automation and Adaptability through Reinforcement Learning**: Traditional rule-based or supervised models may struggle to adapt to rapidly changing market environments. RL agents, by learning directly from market interactions, offer the potential for dynamic strategy optimization, but their effective deployment requires overcoming challenges related to training stability, exploration-exploitation balance, and real-world validation

B. Research Hypothesis

The central hypothesis of this work is that the integration of sentiment analysis from social media and market discussions with classical quantitative forecasting techniques can significantly enhance the quality of investment recommendations. Moreover, the introduction of reinforcement learning agents in future system iterations is expected to further optimize portfolio management through autonomous, adaptive decision-making.

C. Project Motivation

The motivation behind this project is driven by several converging trends in the financial industry:

- The increasing influence of public sentiment and online discourse on asset prices
- The rising demand for automated advisory solutions capable of handling complex data landscapes
- The need for AI systems that not only predict but also adapt to changing market conditions, improving both strategy resilience and investment outcomes

By addressing these challenges, the project aims to contribute to the development of more sophisticated, AI-driven investment tools, making advanced analytics accessible to a wider audience of market participants and democratizing financial decision support.

III. Overview of Existing Methods and Challenges

The problem of integrating both structured financial data and unstructured sentiment data for investment decision-making has been the focus of numerous academic studies and industrial implementations. Existing methods can be grouped into three primary categories: Natural Language Processing (NLP)-based sentiment analysis, time-series forecasting models, and reinforcement learning strategies for trading optimization.

A. NLP-based Sentiment Analysis Methods

Sentiment analysis has become one of the most actively explored areas in financial AI systems. Traditional sentiment analysis approaches were built upon lexicon-based methods or classical machine learning classifiers such as Naive Bayes and Support Vector Machines (SVM). However, these techniques often fail to capture contextual nuances in financial language, especially when dealing with sarcasm, abbreviations, or domain-specific expressions.

The introduction of deep learning, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), significantly improved sentiment analysis accuracy by learning complex language patterns. More recently, transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) revolutionized NLP by enabling deep contextual understanding of text through attention mechanisms.

In the financial domain, **FinBERT**, a fine-tuned variant of BERT on financial texts, has demonstrated superior performance over general-purpose models. According to Araci (2019), FinBERT achieved accuracy rates exceeding 98% in sentiment classification tasks on financial datasets, outperforming traditional NLP models.

Challenges in Sentiment Analysis:

- Financial text often contains ambiguous terminology, sarcasm, or rapidly evolving jargon
- General-purpose sentiment models tend to underperform without financial domain adaptation
- Aggregating sentiment from multiple social media platforms introduces heterogeneity in data formats and quality

B. Time-Series Prediction Methods

Financial forecasting has traditionally relied on statistical models such as ARIMA (AutoRegressive Integrated Moving Average) and GARCH (Generalized Autoregressive Conditional Heteroskedasticity) for predicting stock prices and volatility. While effective for stationary time-series data, these models struggle with capturing complex nonlinear dependencies [7] and regime changes in financial markets.

Machine learning-based approaches, including **LSTM (Long Short-Term Memory networks)** and **TCN (Temporal Convolutional Networks)**, have been widely adopted for time-series forecasting due to their ability to handle sequential dependencies and noisy data. LSTM models are specifically designed to learn long-term dependencies in sequential data through gated memory cells, while TCNs leverage causal convolutions with dilations to achieve stable and scalable forecasting performance.

Challenges in Time-Series Prediction:

- Financial data are inherently noisy and non-stationary
- Overfitting to historical data remains a significant risk without proper regularization and validation
- Traditional models may fail to capture sentiment-induced market movements without integrating external signals

C. Reinforcement Learning Approaches

Reinforcement Learning (RL) methods have shown promise in the development of adaptive trading strategies where the agent learns optimal actions through interaction with the environment. Notable algorithms include **Deep Q-Networks (DQN)**, **Proximal Policy Optimization (PPO)** [8], and **Deep Deterministic Policy Gradient (DDPG)**.

These methods enable agents to optimize decision-making processes by maximizing cumulative rewards, which may correspond to profit maximization or risk-adjusted returns in the context of trading. RL-based systems are capable of dynamically adjusting strategies according to changing market conditions, which traditional supervised models cannot easily achieve [4].

However, RL faces several practical challenges in the financial domain: **Challenges in Reinforcement Learning:**

- Training RL agents requires large amounts of historical or simulated data, which may not fully represent future market conditions (offline training process)
- Instability in policy convergence and sensitivity to hyperparameters
- Difficulty in defining appropriate reward functions that balance profitability and risk

D. General System Architecture Challenges

The successful integration of the above methods into a **unified system raises additional architectural and engineering challenges: [11], [12]**

- Ensuring scalability and low latency in real-time data processing
- Maintaining data privacy and security, especially when dealing with sensitive financial information
- Providing explainable and transparent model outputs for regulatory compliance and user trust

IV. Method Selection and Justification

Considering the challenges identified in the analysis of existing approaches, the selection of methodologies for this project was driven by the need for high accuracy, scalability, real-time processing capabilities, and adaptability to dynamic market conditions. The chosen methods and technologies reflect best practices from recent academic research and industrial implementations, while also addressing the specific requirements of the financial domain.

A. Selection of Sentiment Analysis Method

For the sentiment analysis component of the platform, **FinBERT** was selected as the primary model. FinBERT is a domain-specific adaptation of BERT (Bidirectional Encoder Representations from Transformers) that has been fine-tuned on large-scale financial text corpora, including earnings reports, analyst commentary, and financial news. This fine-tuning enables FinBERT to handle the complex linguistic structures and specialized vocabulary typical of financial texts.

Justification for Using FinBERT:

- Demonstrated superior performance in financial sentiment classification tasks compared to general-purpose NLP models
- Ability to capture contextual nuances, including negations, sarcasm, and domain-specific jargon, which are prevalent in financial discussions
- Availability of pre-trained models and datasets, which accelerates the development process while ensuring reliable baseline performance

B. Selection of Time-Series Forecasting Models

The project adopted two complementary time-series forecasting models: **Long Short-Term Memory networks (LSTM)** and **Temporal Convolutional Networks (TCN)**.

Justification for LSTM and TCN:

- LSTM networks are well-suited for learning long-term dependencies in sequential data, overcoming the vanishing gradient problem typical of traditional RNNs
- TCNs offer better parallelization capabilities and stable gradients through dilated convolutions, making them efficient and scalable for long time horizons
- Both models have demonstrated strong performance on financial time-series data, with the flexibility to incorporate multiple input features, including technical indicators and sentiment signals

C. Selection of Reinforcement Learning Strategies

Although reinforcement learning is positioned as a future enhancement of the platform, the following RL strategies were identified for experimental deployment: **Proximal Policy Optimization (PPO)** and **Deep Q-Networks (DQN)**.

Justification for PPO and DQN:

- PPO is known for its stable convergence properties and efficient policy optimization in continuous action spaces, which are critical in dynamic financial environments
- DQN provides a straightforward approach for learning discrete action policies and has shown success in multiple algorithmic trading research scenarios
- Both methods have extensive support in popular RL libraries such as TensorFlow Agents and Stable Baselines, facilitating experimentation and reproducibility

D. Backend Architecture and Communication Protocols

The platform's backend is designed as a **microservice-based architecture** implemented in Golang, with **gRPC** chosen as the primary inter-service communication protocol.

Justification for Golang and gRPC:

- Golang provides strong support for concurrency and parallelism through lightweight goroutines, which is critical for scalable real-time data processing
- gRPC ensures low-latency, type-safe communication between services using protocol buffers (protobuf), reducing the overhead of message serialization and deserialization
- The choice of gRPC facilitates seamless integration between backend services and the machine learning module, supporting both synchronous and asynchronous communication patterns

E. Data Storage Solutions

Two types of storage technologies were selected to address different data requirements:

- **PostgreSQL** was chosen for structured financial data storage, including user information, historical market data, and transactional logs
- **PineconeDB**, a vector database, was selected for storing high-dimensional embedding vectors generated by the sentiment analysis module. This enables efficient semantic search and retrieval operations on processed textual data

V. Plan for Solving the Task

The development of the AI-driven financial advisory platform has been structured into several key stages, focusing on modular implementation and progressive integration of its core components. The project architecture is divided into three main functional blocks: the backend microservices, the machine learning module (ML module), and the frontend user interface. Each of these components is designed to operate independently at the initial stage, with well-defined interfaces facilitating their eventual integration into a unified system.

Given the scope and complexity of the project, the work plan emphasizes the gradual development of each component, starting from foundational functionalities and moving toward advanced features such as reinforcement learning-based portfolio optimization in future iterations. As of the current stage of the project, the backend module is the most developed among the three, while the ML module and frontend remain in the prototype phase.

A. Backend Development Plan

The backend serves as the core of the system, responsible for managing user authentication, handling data aggregation, and coordinating communication between services. The following key tasks have been defined for this component:

- **Design and implementation of core microservices**: Development of user management, authentication (JWT-based), and session handling services
- **Data ingestion and aggregation**: Creation of pipelines for collecting structured market data from external sources, including price feeds and financial indicators
- **Integration with the ML module**: Defining gRPC-based contracts and data schemas to facilitate communication

between backend services and the machine learning models
- **Security and compliance considerations**: Ensuring secure service interactions, data encryption where necessary, and preparing the system architecture for future scalability
- **API gateway and external interface**: Exposing RESTful APIs for interaction with the frontend module, while keeping internal communication optimized via gRPC

The backend development is progressing steadily, with the foundational microservices and basic data flow already implemented. Further work will focus on enriching the feature set, improving stability, and finalizing the interfaces with the ML module.

B. Machine Learning Module Development Plan

The machine learning module is intended to process financial data and extract actionable insights through sentiment analysis and time-series forecasting. The current stage involves experimental prototyping and evaluation of different modeling approaches. The development plan includes the following tasks:

- **Implementation of the sentiment analysis pipeline**: Fine-tuning and deploying the FinBERT model for financial text classification, focusing on data from Twitter and TradingView discussions [9]
- **Forecasting module development**: Experimentation with time-series prediction models such as LSTM and TCN, using historical market data combined with sentiment signals
- **Evaluation and benchmarking**: Selection of suitable evaluation metrics (accuracy, F1-score, Sharpe ratio, backtesting performance) for validating sentiment models and forecasting algorithms
- **Preparation of integration interfaces**: Defining clear gRPC service contracts for interaction with the backend microservices
- **Research and groundwork for future reinforcement learning agent**: Designing the architecture for an RL-based trading agent as a future enhancement. The RL component will be developed after the successful integration of the core functionalities, with the goal of automating and optimizing portfolio management strategies

At the current stage, preliminary experiments on sentiment analysis have been conducted, and model selection processes are ongoing. Work on forecasting models and integration schemes is planned to proceed in parallel with the backend development.

C. Frontend Development Plan

The frontend module serves as the user-facing component of the platform, providing access to system insights, forecasts, and investment recommendations. The initial focus is on creating a simple, functional web interface, with plans to expand features based on user needs and system integration progress.

Key tasks for the frontend development include:

- **Design and implementation of basic user interfaces**: Development of interactive dashboards for data visualization, including sentiment analysis results and market trend predictions
- **Integration with the backend APIs**: Establishing seamless data flow between the frontend and backend using RESTful API endpoints exposed by the backend services
- **Prototyping of conversational AI elements**: Laying the foundation for a chatbot-style user interface, which will later serve as the main interaction channel between the user and the system
- **User experience (UX) optimization**: Refining the design based on feedback and ensuring accessibility, usability, and responsiveness across devices

The frontend is currently in the early stages of prototyping, with priority given to basic functionality and data visualization. Further enhancements will align with the progress of the backend and ML module integration.

D. Future Integration Plan

The long-term plan for the project envisions the seamless integration of the three primary modules into a unified, scalable system. The following principles will guide the integration process:

- Modular development with clear service contracts to facilitate independent testing and gradual integration
- Continuous evaluation of system performance and user feedback to inform iterative improvements
- Gradual introduction of advanced features, including the reinforcement learning agent, only after the stable operation of core functionalities

This stepwise approach ensures that each component reaches a sufficient level of maturity before full system integration, minimizing risks associated with complex system dependencies and allowing for flexible adaptation to changing requirements.

VI. Obtained Results

Although the project remains under active development, the internship period has resulted in significant preliminary achievements that form a strong basis for the continuation of work within the bachelor final project.

The following outcomes were achieved during the internship:

- **Initial backend infrastructure implemented**: A microservice-based architecture was designed and partially developed using Golang and gRPC for secure and efficient service communication. Key backend components, including user management and data aggregation services, have been prototyped and tested
- **Machine learning module prototyping**: The FinBERT model was fine-tuned on selected financial datasets (Twitter and TradingView posts), achieving well validation accuracy levels in initial sentiment classification tasks.

Experiments with time-series forecasting models, including LSTM and TCN architectures, were conducted to evaluate their applicability for price prediction task [9]
- **Defined integration contracts**: gRPC-based communication schemes between backend services and the machine learning module were outlined, establishing a foundation for seamless integration of these components in future stages
- **Frontend prototype designed**: A preliminary version of the web interface was created using React and TypeScript, focusing on basic minimal visualization of sentiment analysis results and forecast outputs
- **Literature review and methodological groundwork completed**: A comprehensive analysis of related work in the areas of financial NLP, time-series prediction, and reinforcement learning was carried out. This review serves as the theoretical foundation for the planned continuation of the project, including the future development of an RL-based trading agent

The current implementation represents a draft-level version of the system, with core architecture and essential components being prototyped and tested individually. Full system integration and production-ready functionality remain future objectives for the subsequent thesis development phase.

VII. Conclusions

The internship work has successfully established the conceptual and technical framework for the AI-driven financial advisory platform, with initial development efforts focused on prototyping the main system modules. During this period, valuable experience was gained in backend service development, machine learning model experimentation, and system architecture design.

The methods selected for sentiment analysis (FinBERT), time-series forecasting (LSTM, TCN), and backend microservice communication (Golang, gRPC) were validated through early-stage testing and half confirmed to be appropriate for the project's objectives.

While the current implementation remains at the prototype level, the work completed during the internship provides a robust foundation for the subsequent thesis project. The key architectural decisions, preliminary model evaluations, and research insights achieved at this stage will enable efficient and structured continuation of development toward a fully functional investment recommendation platform.

Future efforts will focus on expanding the ML module, completing the backend infrastructure, fully integrating the system components, and preparing for the experimental deployment of a reinforcement learning agent for dynamic portfolio optimization.

References

- **[1]** Wang, L., Cheng, Y., Xiang, A., Zhang, J., & Yang, H. (2024). Application of Natural Language Processing in Financial Risk Detection. *arXiv*. Available: https://arxiv.org/abs/2406.09765.
- **[2]** Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv*. Available: https://arxiv.org/abs/1908.10063.
- **[3]** Taylor, K., & Ng, J. (2023). Natural Language Processing and Multimodal Stock Price Prediction. *arXiv*. Available: https://arxiv.org/abs/2401.01487.
- **[4]** Liu, X.-Y., Xiong, Z., Zhong, S., Yang, H., & Walid, A. (2022). Practical Deep Reinforcement Learning Approach for Stock Trading. *arXiv*. Available: https://arxiv.org/abs/1811.07522.
- **[5]** Mokhtari, M., Seraj, A., Saeedi, N., & Karshenas, A. (2023). The Impact of Twitter Sentiments on Stock Market Trends. *arXiv*. Available: https://arxiv.org/abs/2302.07244.
- **[6]** Deloitte (2023). AI and Machine Learning in Financial Markets. Available: https://www2.deloitte.com.
- **[7]** Olorunnimbe, K., & Viktor, H. (2023). Deep learning in the stock market—a systematic survey of practice, backtesting, and applications. *Artificial Intelligence Review*, 56(3), 2057–2109.
- **[8]** Pricope, T.-V. (2021). Deep Reinforcement Learning in Quantitative Algorithmic Trading: A Review. *arXiv*. Available: https://arxiv.org/abs/2106.00123.
- **[9]** Twitter (2023). Twitter API pricing and limitations. Available: https://developer.twitter.com.
- **[10]** Chen, J. (2022). Basics of Algorithmic Trading: Concepts and Examples. *Investopedia*. Available: https://www.investopedia.com/terms/a/algorithmictrading.asp.
- **[11]** Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2023). Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology*. Available: https://dl.acm.org/doi/10.1145/3342911.
- **[12]** Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., and Nitin, G. (2021). Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*. Available: https://arxiv.org/abs/1912.04977.
- **[13]** Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*. Available: https://arxiv.org/abs/1908.09635.

VIII. Appendices

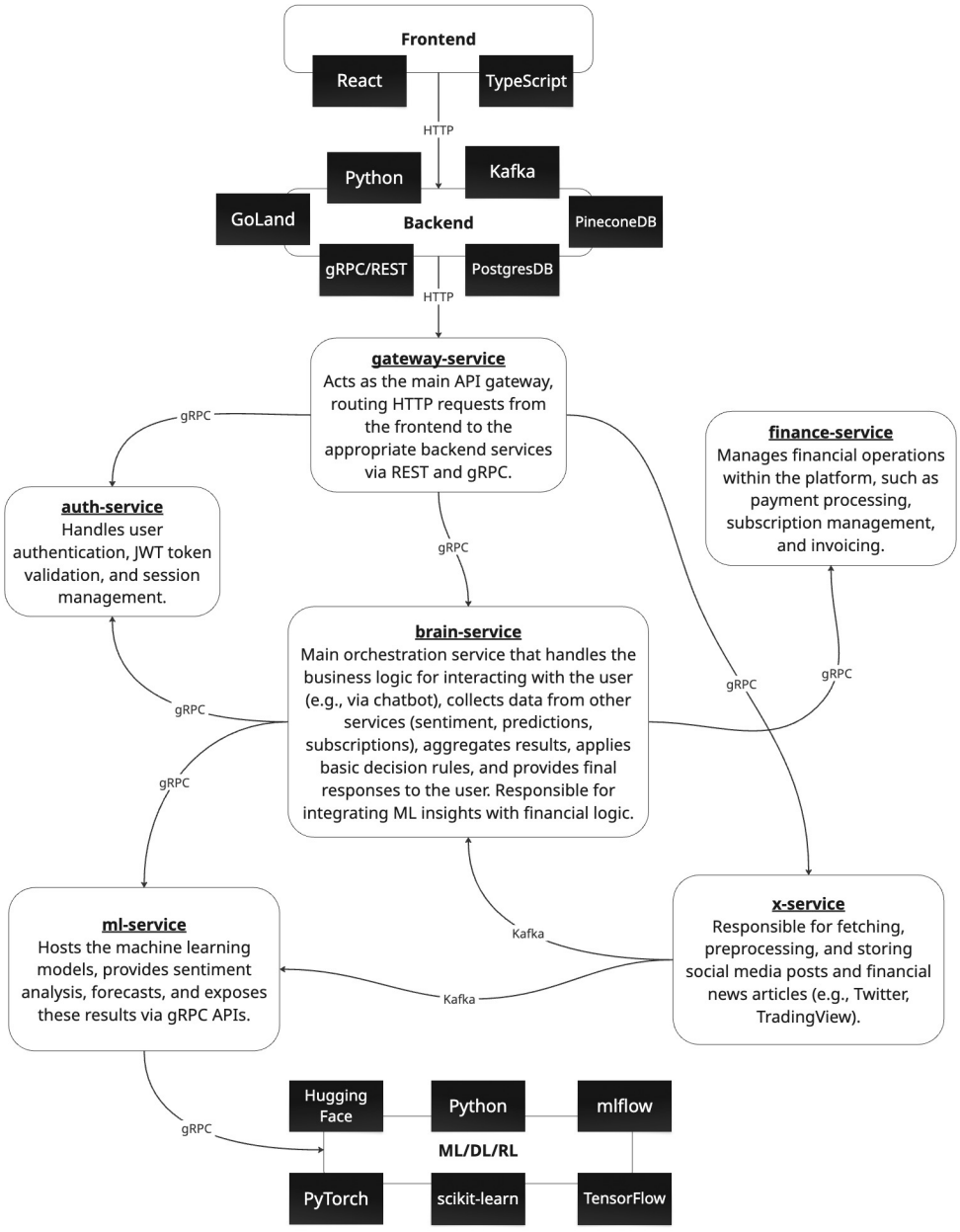

Fig. 1. Architecture of the AI-driven Financial and Real Estate Investment Recommendation Platform. This diagram illustrates the microservice-based backend structure and inter-service communication using REST and gRPC protocols.