Portfolio Optimization Using Online Learning and Alternative Data Sources

DISSERTATION - MID SEM REPORT

Submitted in partial fulfillment of the requirements of the

Degree : MTech in Data Science

By

Oindreela Bhowmick
2023DA04155

Under the supervision of

Kartik Pandithar
Data Engineer

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

July, 2025

Abstract

Key Words: Online learning, Portfolio optimization, Asset allocation, Machine learning, Streaming data, Incremental learning, Reddit sentiment, Social media signals, Alternative data sources, Natural language processing (NLP)

This project is designed to tackle the challenge of portfolio optimization in modern financial markets by using **online machine learning techniques** and **alternative data sources**. We have gone through multiple research papers and observed that traditional portfolio management approaches often depend on batch learning and historical price data, which is definitely not suitable for real-time decision-making in fast changing market conditions. Our approach uses **online (incremental) learning algorithms** to enable **adaptive portfolio allocation**, allowing the model to update continuously as new data becomes available.

The core of the system is built around the idea of streaming data inputs and evolving market behavior. To improve decision-making, this project uses alternative data like Reddit and finance news sentiment to understand what everyday investors are talking about and how public opinion might affect the stock market. Reddit data will be pruned (collected from subreddits like r/stocks and r/wallstreetbets) using Reddit's python API, then processed using **natural language processing (NLP)** techniques to extract general stock-level sentiment. Steps will be taken to ensure the social media data is free of noise as much as possible. This sentiment is converted to numerical values and incorporated as an additional feature in the learning algorithm, which will in turn enable the model to consider social media opinion in its portfolio decisions. This feature will be combined with traditional financial indicators such as price, volume, and volatility obtained from sources like yfinance.

The online learning model was chosen for this and implemented using either Python libraries such as **online gradient descent etc.**, which supports real-time learning. Risk metrics such as Sharpe Ratio, maximum drawdown, and cumulative returns will be used to evaluate performance and choose the best model.

Also, an **automated alert system** will be implemented to notify the user of significant market or sentiment events via email. A weekly or monthly email will be sent out to users to alert them of the current status of the stocks.

Lastly, this project aims to demonstrate a real-time, and new approach to portfolio optimization by combining various aspects like **machine learning**, **social sentiment analysis**, **alternative data**, and **automation**. We plan to highlight the significance of incorporating unconventional data sources and adaptive models to help make better financial decisions in this ever-changing market.

List of symbols and abbreviations used

Abbreviation/Symbol	Expanded form
yfinance	Yahoo Finance
API	Application Programming Interface
FinVader	Financial Valence Aware Dictionary and sEntiment Reasoner
FinBERT	Financial Bidirectional Encoder Representations from Transformers
RSI	Relative Strength Index
MACD	Moving Average Convergence Divergence
SMA	Simple Moving Average
EMA	Exponential Moving Average
BB	Bollinger Bands
XAI	Explainable AI
LIME	Local Interpretable Model-agnostic Explanations
SHAP	SHapley Additive exPlanations

List of Tables

List of Figures

Table of Contents

Chapter 1: Introduction and Objectives

1.1 Project Overview

The goal of this project is to create an automated daily portfolio rebalancing system that uses online learning models to change investment allocations on the fly. The system combines data from the financial markets with data from other sources, like news and social media sentiment, to make features for the model. A shell script runs the necessary Python files in order, managing and automating the entire workflow. This makes it simpler to deploy and monitor the system while ensuring that it runs smoothly and can be repeated.

1.2 Objectives as Stated in Abstract Submission

The key objectives of the project are as follows:

- **To build a portfolio optimization system** that can learn and adapt over time using online learning techniques.
- **To use alternative data** like Reddit posts and Google search trends to better understand investor sentiment and market behavior.
- **To combine this data with traditional stock market information** (like prices and volume) to make smarter investment decisions.
- **To build an alert system** that sends useful updates to the user.

If we break down the above objectives further, then below are the steps -

1. Finalize project scope, objective, and data sources
2. Collect historical stock data using yfinance/other APIs
3. Collect and preprocess Reddit or other social media data
4. Feature engineering
5. Select and implement online learning models
6. Train and validate models using combined data
7. Implement traditional models and comparison with our model
8. Build and test alert system
9. Final model tuning
10. Documentation (code + user manual + explanations)

1.3 Objectives Met Till Midterm

At the midterm stage, the following objectives have been achieved:

- Finalized project scope, objectives, and data sources: The scope and objectives were clearly defined, and data sources (Yahoo Finance for stock data, Reddit for social media sentiment) were identified and documented.
- Collected historical stock data using yfinance API: Automated scripts have been developed to fetch and store historical stock data.

- Collected and preprocessed Reddit data: Modular Python scripts fetch and preprocess Reddit posts and comments, with sentiment analysis features integrated.
- Feature engineering: Feature engineering modules have been implemented, including technical indicators, sentiment analysis and feature merging.
- Currently working on finalizing the batch and online learning model to use, have tried on SGDRegressor, and Online Gradient Descent with Momentum, along with LSTM on historical data.

Chapter 2: Literature Review

2.1 Automated Portfolio Rebalancing

Automated portfolio rebalancing shifted from basic rule-based tactics like periodic or threshold-based rebalancing to more data-driven adaptive methods as artificial intelligence and machine learning have greatly improved. Even though they can be simple customary methods that are widely used, they often fail in responding to the rapid changes that do occur in market conditions, resulting in suboptimal risk-return profiles. AI-powered modern rebalancing systems use real-time data analysis as well as predictive modeling plus dynamic optimization to adjust portfolios constantly given investor behavior and market movements. This shift improves efficiency and also reduces costs plus manages risk better through enabling timely adjustments that better align with investor goals and market realities.

2.2 Online Learning Algorithms in Finance

Because they adapt rapidly to new data online learning algorithms suit dynamic settings like financial markets therefore they are important for financial modeling. Reinforcement learning coupled with neural networks are now being applied within portfolio management. Therefore these models can learn a lot from the evolving market conditions and also can optimize asset allocations without end. When reacting to macroeconomic events, market sentiment shifts, along with other factors that may strongly affect asset prices gives these approaches a real edge beyond typical static models. Algorithms that can be complex could become just “black boxes” without each of the mechanisms to ensure appropriate transparency. The literature stresses model interpretability coupled with explainability because trust or validation is hard.

2.3 Alternative Data and Social Media Sentiment in Financial Modeling

Using alternative data, like social media sentiment, is now a key part of modern quantitative finance. Research shows that sentiment taken from news and social media can be very useful for investors, sometimes even matching or beating the results of traditional multi-factor strategies. Reddit, Twitter, and financial news aggregators are examples of platforms that let people quickly share information and opinions. This can change how investors act and how prices move in the market. One example is the "meme stock" phenomenon, in which people on sites like Reddit shared their opinions and caused prices to change dramatically, which went against traditional market models.

2.4 Explainability and Model Transparency in Financial AI

The need for explainable and transparent models has become a major concern as financial modeling increasingly uses cutting-edge machine learning and artificial intelligence techniques. Deep neural networks and ensemble methods are examples of complex algorithms that can achieve high predictive accuracy, but they frequently operate as "black boxes," making it challenging for end users, regulators, and practitioners to comprehend the reasoning behind

model decisions. In the financial industry, where risk management, regulatory compliance, and trust are crucial, this opacity presents serious difficulties.

The increasing use of explainable AI (XAI) methods in quantitative finance is highlighted in recent research. To interpret model outputs, pinpoint the most important features, and find possible biases in the data or modeling process, techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are employed. Also, explainability is not just a technical need; it's also a strategic benefit. Transparent models make it easier to talk to stakeholders, validate models more effectively, and keep up with changing rules in the financial industry. Because of this, explainability and model transparency are now seen as important parts of using AI responsibly in finance. This has changed both how researchers study the subject and how businesses use it.

Chapter 3: Methodology and Progress

3.1 Data Collection and Preprocessing

Table 1: Data Sources Overview			
Data Source	**Type**	**Description**	**Data Collected**
yfinance	Financial	Provides historical and real-time stock price, volume, and related financial metrics	Stock prices, volumes, returns
Reddit	Social Media	Provides user-generated posts and comments on stocks and financial markets	Posts, comments, upvotes
News API	News	News articles and headlines related to financial markets	News articles, headlines

The thorough gathering and preprocessing of financial and non-financial data sources forms the basis of this project. I periodically gathered historical stock prices, volumes, and associated market metrics for financial data using the yfinance API. To keep the dataset up to date and complete, the data pipeline is made to retrieve and update this data on a regular basis.

I concentrated on Reddit for social media data because of its increasing impact on market dynamics and sentiment among retail investors. I gathered posts and comments from pertinent subreddits using custom Python scripts, filtering for keywords and interesting tickers. Understanding that Reddit data is extremely noisy and unstructured, I applied a number of preprocessing techniques, including eliminating duplicates, dealing with missing values, standardizing text, and weeding out low-quality content.

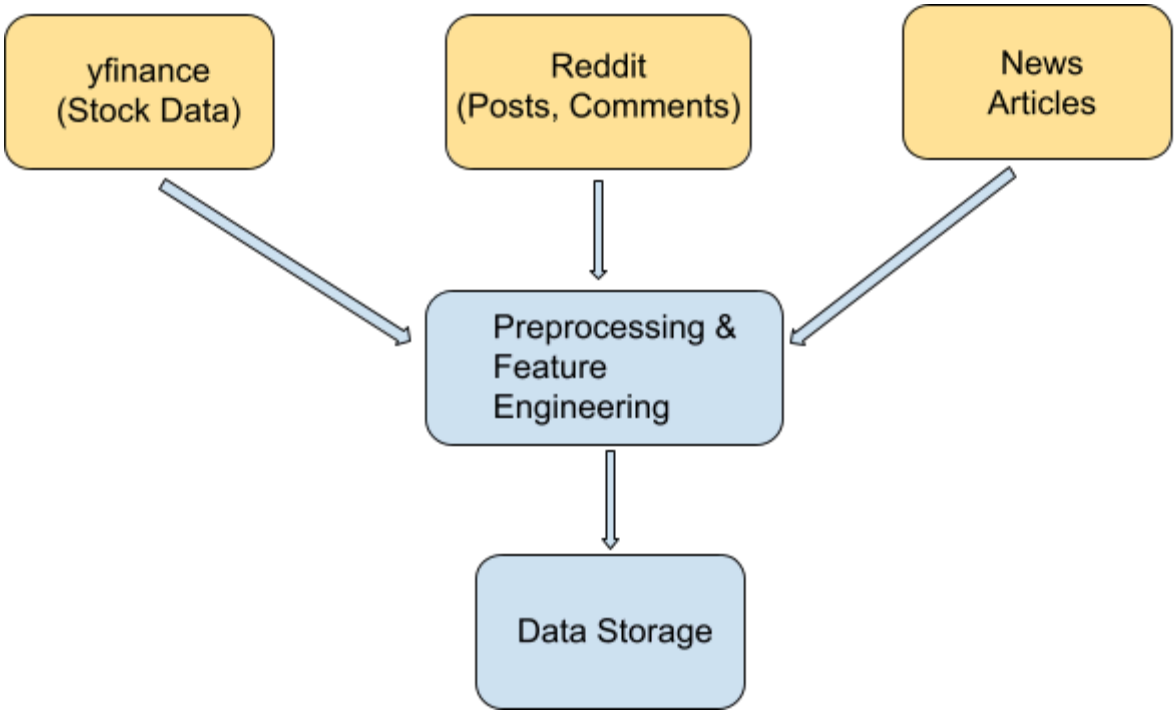

Figure 1: Data Collection Pipeline

Sentiment analysis was a critical step in transforming raw Reddit text into actionable features. I experimented with several open-source models, including FinBERT and Llama2-Financial. While Llama2-Financial initially showed promise, it was ultimately discarded due to its slow inference speed on my local setup. Instead, I adopted a hybrid approach: sentiment scores were generated using both FinBERT and FinVADER, and the final sentiment value for each post was computed as the mean of these two models' outputs. This ensemble method aimed to balance the strengths of both models and mitigate individual biases or weaknesses.

Feature engineering played a central role in bridging raw data and model-ready inputs. In addition to sentiment scores, I engineered a range of technical indicators from the financial data, such as rolling averages, RSI, MACD, and Bollinger Bands. Event-based features, such as abnormal volume spikes or sentiment surges, were also incorporated to capture market anomalies that might influence rebalancing decisions. The feature engineering workflow was carefully documented and modularized for reproducibility and future extension.

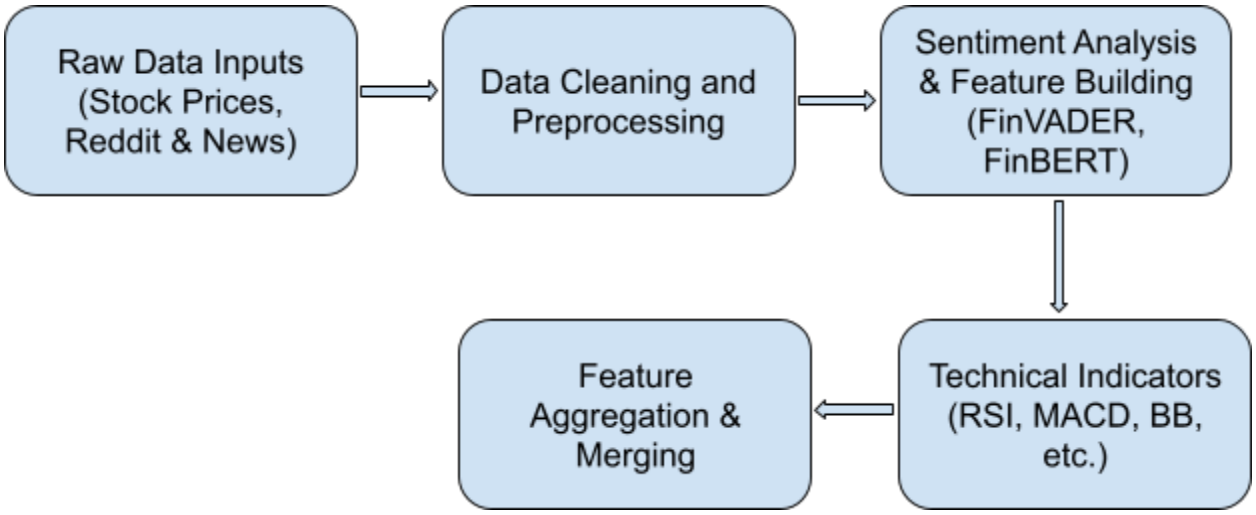

Figure 2: Feature Engineering Workflow

Table 2: Feature Engineering Summary			
Feature Name	**Source**	**Description**	**Purpose**
Sentiment Score	Reddit/News	Aggregated sentiment from posts and comments, using FinVADER and FinBERT	Capture market sentiment
RSI	Momentum	Relative Strength Index	Identifies overbought/oversold conditions
MACD	Trend/Momentum	Moving Average Convergence Divergence (difference between 12/26-period EMAs)	Signals trend direction and momentum
MACD Signal	Trend/Momentum	Signal line (9-period EMA of MACD)	Generates buy/sell signals
SMA 20	Trend	Simple Moving Average (20-period)	Smooths price trends, identifies support/resistance
EMA 20	Trend	Exponential Moving Average (20-period, more weight to recent	Responds faster to price changes than SMA

Table 2: Feature Engineering Summary			
		prices)	
BB Middle Band	Volatility	Bollinger Bands Middle (20-period SMA)	Midpoint of price range
BB Upper Band	Volatility	Bollinger Bands Upper (Middle + 2× standard deviation)	Upper boundary, identifies overbought
BB Lower Band	Volatility	Bollinger Bands Lower (Middle – 2× standard deviation)	Lower boundary, identifies oversold

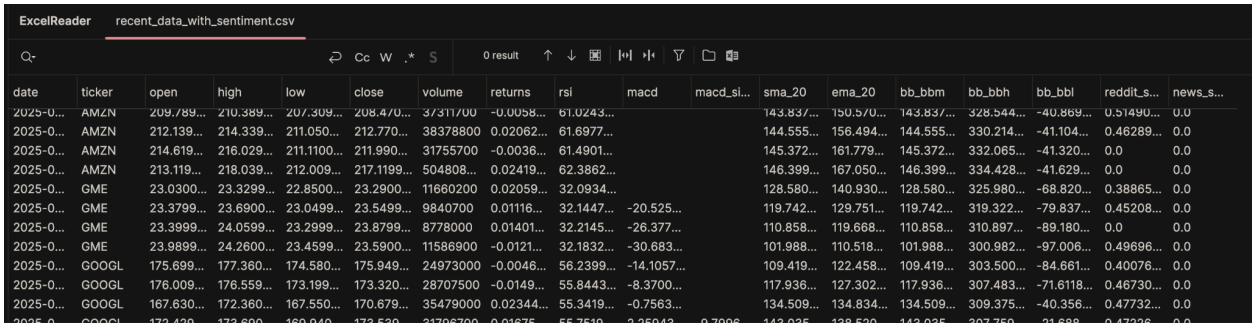

Figure 3: Data Snapshot

3.2 Model Development

The core modeling objective was to implement an online learning framework capable of adapting to new data in real time—a necessity in fast-moving financial markets. I began with the SGDRegressor, a well-established online learning algorithm, and then extended the approach by developing a custom Online Gradient Descent with Momentum (OGDM) model. The OGDM was designed to offer greater adaptability and responsiveness to shifting market conditions, which is essential for effective portfolio rebalancing.

To provide a benchmark and facilitate direct comparison, I also implemented a traditional LSTM model, training it on the preprocessed historical data. The LSTM's ability to capture temporal dependencies offered a useful contrast to the online learning models, highlighting the trade-offs between batch and incremental learning approaches. Currently I am focussing on fine tuning the models and preparing a comparative analysis.

3.3 Workflow Automation

Recognizing the importance of automation and reproducibility in data-driven projects, I replaced the earlier Airflow-based orchestration with a streamlined shell script. This script coordinates the sequential execution of all Python modules involved in data collection, preprocessing, feature engineering, and modeling. The switch to a shell script simplified the workflow, reduced overhead, and made the system easier to deploy and maintain on local or cloud environments.

To further enhance reliability, I began integrating basic monitoring and alerting mechanisms. These include status checks after each pipeline stage and notifications in case of failures or anomalies. While these features are still in their early stages, they lay the groundwork for a more robust and production-ready system.

Chapter 4: Discussion of Results and Objectives Met

This chapter provides a detailed reflection on the progress made so far in the project, highlighting the key objectives achieved and the practical outcomes realized.

4.1 Data Collection and Preprocessing

The data collection process has been successfully automated through custom Python scripts that reliably fetch and preprocess data from both financial and social media sources. Historical stock data is gathered using the yfinance API, ensuring comprehensive coverage of market activity of several years. On the social media front, Reddit posts and comments are collected from carefully selected subreddits that are relevant to financial markets and investment discussions.
Recognizing the inherent noisiness of social media data, especially from platforms like Reddit, the preprocessing pipeline incorporates several filtering mechanisms. These include focusing on top-voted comments to prioritize more credible and relevant content, as well as keyword filtering to capture posts that mention specific tickers or financial terms. This approach has significantly improved the quality of the sentiment data used in the model.

4.2 Feature Engineering

The project has successfully merged sentiment features derived from social media text with traditional financial indicators extracted from stock price data. Sentiment scores are generated using a combination of models, including FinBERT and FinVADER, and are aggregated to provide a balanced view of market sentiment. Alongside sentiment, a suite of technical indicators such as RSI, MACD, and Bollinger Bands have been computed to capture price trends and volatility. These diverse features are integrated into a cohesive dataset that serves as the input for the modeling phase, enabling the models to leverage both market data and alternative sentiment signals.

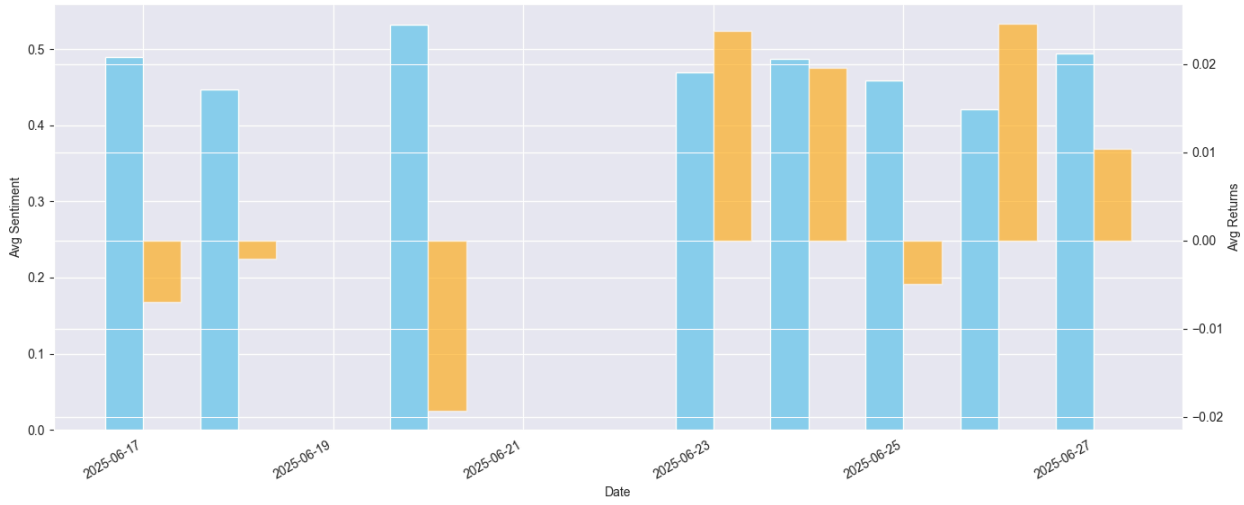

Figure 4: Average Daily Sentiment vs. Returns for META

This figure illustrates the relationship between average daily sentiment scores (derived from Reddit posts and comments) and daily stock returns for META over a recent period. Sentiment is shown on the left y-axis (blue bars), while returns are shown on the right y-axis (orange bars). The visualization highlights how shifts in online sentiment may correspond to changes in stock performance, providing insight into the potential predictive value of alternative data sources in portfolio management.

4.3 Model Implementation

The online learning models developed, including the SGDRegressor and a custom Online Gradient Descent with Momentum (OGDM) algorithm, are fully operational and integrated within the automated workflow. These models are designed to adapt continuously to new data, reflecting the dynamic nature of financial markets. The implementation has demonstrated the feasibility of using online learning for portfolio rebalancing, with evaluation strategy still under progress.

4.4 Workflow Automation

To streamline the project execution and simplify deployment, the entire workflow is orchestrated through a custom shell script. This decision was made after initial experimentation with Apache Airflow, which, while powerful, introduced complexity and overhead that were not ideal for the current project scale and resource constraints. The shell script approach has proven effective in managing the sequential execution of data collection, preprocessing, feature engineering, and model training steps, ensuring reproducibility and ease of maintenance.

4.5 Challenges and Solutions

One of the primary challenges encountered is the noisy nature of social media data, which can introduce volatility and reduce the reliability of sentiment signals. To mitigate this, the project employs targeted data collection from dedicated financial subreddits and prioritizes top-voted comments, which tend to be more informative and less prone to spam or irrelevant content. Additionally, combining sentiment scores from multiple models helps to smooth out individual model biases and improve overall sentiment accuracy. These strategies collectively enhance the robustness of the sentiment features and contribute to more stable model performance.

Chapter 5: Challenges and Future Work

5.1 Noisy Sentiment Data

One of the most persistent challenges in this project has been the inherent noisiness of social media sentiment, particularly from platforms like Reddit. Unlike structured financial data, Reddit posts and comments are highly unstructured, often filled with slang, sarcasm, and off-topic discussions. This makes it difficult to extract reliable sentiment signals that can be confidently used in financial modeling. To address this, I have focused on advanced aggregation strategies and careful feature engineering. For example, I prioritize data from dedicated financial subreddits and filter for posts and comments that receive higher upvotes, as these are more likely to reflect genuine community sentiment. Additionally, by combining the outputs of multiple sentiment models (FinBERT and FinVADER), I aim to smooth out individual model biases and reduce the impact of outliers. While these steps have improved data quality, extracting truly actionable signals from noisy sentiment remains an ongoing area of experimentation and refinement.

5.2 Deployment and Scalability

As the project matures, attention is shifting toward making the system more accessible and scalable. One of the key next steps is to develop a user-friendly interface that allows end users to interact with the portfolio rebalancing system without needing to understand the underlying codebase. This could take the form of a simple web dashboard or command-line tool that presents key metrics, model predictions, and alerts in an intuitive way. In parallel, I am exploring deployment options that will allow the system to run reliably on cloud platforms or other scalable environments. This involves addressing issues such as resource allocation, job scheduling, and data storage, ensuring that the system can handle larger datasets and more frequent updates as it scales.

5.3 Alert System and Model Comparison

Another important area of future work is the implementation of a robust alert system. This system will monitor the health of the data pipeline and model predictions, providing timely notifications in the event of failures, anomalies, or significant changes in portfolio recommendations. The goal is to ensure that users are always aware of the system's status and can respond quickly to any issues.

Additionally, a comprehensive comparison between the online learning models and traditional approaches is planned. While initial experiments suggest that online learning models offer greater adaptability, it is important to rigorously benchmark their performance against established methods such as LSTM networks and other batch learning algorithms. This comparison will focus not only on predictive accuracy but also on explainability, computational efficiency, and real-world applicability. The insights gained will guide further model tuning and inform recommendations for future enhancements.

5.4 Explainability and Model Transparency

As the project advances toward more sophisticated and potentially higher-stakes applications, the explainability of model decisions becomes increasingly important. In financial portfolio management, stakeholders—including end users, analysts, and regulators—need to understand not just what the model predicts, but also why it makes those predictions. This is particularly relevant when combining traditional financial features with alternative data sources such as social media sentiment, where the link between input and output can be less intuitive.

To address this, a key area of future work will be the integration of explainable AI (XAI) techniques into the modeling pipeline. Tools such as SHAP and LIME can be used to quantify and visualize the contribution of each feature—whether it is a technical indicator or a sentiment score—to the model's output. By applying these methods, it will be possible to identify which factors most influence rebalancing decisions and to detect potential biases or overreliance on noisy sentiment signals.

In addition to feature-level explanations, the project will also explore generating user-friendly summaries and visualizations that communicate model reasoning in a transparent manner. This may include interactive dashboards that allow users to inspect how changes in sentiment or market conditions impact portfolio recommendations.

Ultimately, prioritizing explainability will not only increase user trust and facilitate compliance with regulatory standards, but also provide valuable insights for ongoing model improvement and risk management. As the system evolves, regular audits and documentation of model behavior will be incorporated into the workflow to ensure that transparency remains a core principle of the project.

Chapter 6: References

[1] Adams, T., Ajello, A., Silva, D., & Vazquez-Grande, F. (2023). More than Words: Twitter Chatter and Financial Market Sentiment. Finance and Economics Discussion Series 2023-034. Board of Governors of the Federal Reserve System. The Fed - More than Words: Twitter Chatter and Financial Market Sentiment

[2] Borrageiro, G. F. (2023). Online Learning in Financial Time Series (Doctoral dissertation, University College London). Online learning in financial time series - UCL Discovery

[3] Banholzer, N., Heiden, S., & Schneller, D. (2019). Exploiting investor sentiment for portfolio optimization. Business Research, 12, 671–702. https://doi.org/10.1007/s40685-018-0062-6

[4] Korab, P. (2024). The Meme Stock Frenzy: Origins and Implications. SSRN Electronic Journal. The Meme Stock Frenzy: Origins and Implications

[5] Moquist, V. A. (2014). Dynamic Asset Allocation and Algorithmic Trading (Master's thesis, Copenhagen Business School). Dynamic Asset Allocation and Algorithmic Trading

[6] Rey, M. (2023). Rebalanced Portfolio Optimization. SSRN Electronic Journal. http://dx.doi.org/10.2139/ssrn.4526656

[7] Zhang, Y., & Skiena, S. (2024). Financial Sentiment Analysis: Techniques and Applications. ACM Computing Surveys, 57(2), 1-36. Financial Sentiment Analysis: Techniques and Applications | ACM Computing Surveys

[8] Arsenault, P.-D., Wang, S., & Patenaude, J.-M. (2024). A Survey of Explainable Artificial Intelligence (XAI) in Financial Time Series Forecasting. arXiv preprint arXiv:2407.15909. [2407.15909] A Survey of Explainable Artificial Intelligence (XAI) in Financial Time Series Forecasting