Final Project Report: Investor Sentiment in Stock Price Forecasting and Policy Analysis: A Machine Learning Approach

Author: Yiwei Zhang
STATS201: Introduction to Machine Learning for Social Science
Instructed by Prof. Luyao Zhang

Background

Stock market prediction has been a longstanding challenge in finance, with researchers exploring various quantitative models to anticipate price movements. Traditional methods, such as autoregressive integrated moving averages (ARIMA), GARCH models, and factor-based approaches, have formed the foundation of stock price forecasting (Sidi, 2020). Nabipour et al. (2020) employed decision tree, bagging, random forest, adaptive boosting (AdaBoost), gradient boosting, and eXtreme gradient boosting (XGBoost), and artificial neural networks (ANN), recurrent neural network (RNN), long short-term memory (LSTM) and found that LSTM shows more accurate results with the highest model fitting ability. However, the complexity of market behavior—driven by macroeconomic indicators, investor psychology, and market sentiment—has led to the rise of machine learning and deep learning models for stock prediction. In particular, Ren et al. (2018) assert that investor sentiment analysis is a critical factor in forecasting stock trends, with financial texts from social media and news influencing investor behavior. Moreover, the rise of natural language processing techniques has allowed for better extraction of investor sentiment and emotions, which can provide leading indicators for market movements (Lee et al., 2023). They propose the StockEmotions dataset—which contains fine-grained investor sentiment data from StockTwits—to integrate stock market prices and investor sentiment into stock prediction models.

However, previous studies lack a comprehensive approach that integrates both stock price forecasting and causal inference to examine the impact of financial policies on investor sentiment and market behavior. While prior research has explored sentiment-driven stock prediction, it remains limited in its ability to quantify the causal effects of major policy interventions on sentiment and stock prices. I propose this study to leverage the StockEmotions dataset (Lee et al., 2023), which spans 2019 and 2021, along with historical stock price data, to first develop machine learning-based stock price forecasting models, assessing whether integrating granular investor sentiment score would enhance predictive accuracy compared to models relying purely on historical price data. Then, this study applies a Regression Discontinuity (RD) design (detailed in the appendix) to estimate the causal impact of the Federal Reserve's emergency COVID-19 intervention on March 16, 2020, on investor sentiment. This dual approach will provide both predictive insights into stock market fluctuations and rigorous causal evidence on how monetary policies influence investor psychology, bridging the gap between financial forecasting and behavioral finance.

Related Works:

Stock market prediction has been widely studied using various modeling techniques, ranging from traditional time-series approaches to advanced machine learning and deep learning methods. For instance, Sidi (2020) explored time-series stock similarity to enhance predictive accuracy, demonstrating that incorporating related stocks during training improves forecasting performance compared to models trained on isolated stock data. Wang et al. (2021) developed an intraday trading strategy using Markowitz portfolio optimization and Multilayer Perceptron (MLP), validating the profitability of ML-based strategies in China's stock market under T+1 trading restrictions. Li et al. (2024) further introduced MASTER, a Market-Guided Stock Transformer, which captures momentary and cross-time stock correlations using a Transformer-based architecture, outperforming traditional methods in stock price forecasting.

Recent research has also incorporated investor sentiment analysis as a key factor influencing market movements. Alvarez-Gonzalez et al. (2021) investigated text-based emotion detection and evaluated the effectiveness of GoEmotions and Vent datasets, finding that human-perceived emotions often differ from those expressed by writers, which presents challenges in financial sentiment modeling. Abdul-Mageed and Ungar (2017) developed EmoNet, a Gated Recurrent Neural Network (GRU)-based model trained on a large-scale Twitter dataset, achieving state-of-the-art performance in fine-grained emotion detection. Cortis et al., (2017) focused on financial sentiment analysis in microblogs and news articles, highlighting the importance of domain-specific sentiment classification in stock market prediction. Lee et al. (2023) introduced StockEmotions, a novel dataset for financial sentiment and emotion classification, incorporating StockTwits investor discussions to examine the role of emotions in market fluctuations. Rather than aggregated level public emotions from the financial news (Maqbool et al., 2023), this dataset contains granular company specific emotions that could be more precise for forecasting. In this study, I will assess whether combining textual sentiment, emotional indicators, and numerical stock data improves time-series forecasting, particularly during periods of heightened volatility.

Research question

How does the integration of investor sentiment enhance stock price prediction accuracy, and what is the causal impact of the Federal Reserve's March 16, 2020, COVID-19 policy intervention on market sentiment?

Application scenario:

The dataset is derived from financial market. Unlike traditional financial sentiment datasets, StockEmotions provides a richer, fine-grained emotional taxonomy derived from social media investor discussions (Lee et al., 2023). Combining this data with historical stock price data allows for interdisciplinary analysis bridging the gap between financial forecasting and behavioral finance. By integrating machine learning-based stock price forecasting with causal inference through Regression Discontinuity Design, it provides a novel framework for assessing both predictive trends and policy-driven market shifts. Adding on to previous studies about correlation-based sentiment analysis,

this research establishes a causal link between monetary policy interventions and investor sentiment, offering deeper insights into how financial markets react to crisis-driven policy changes.

Methodology

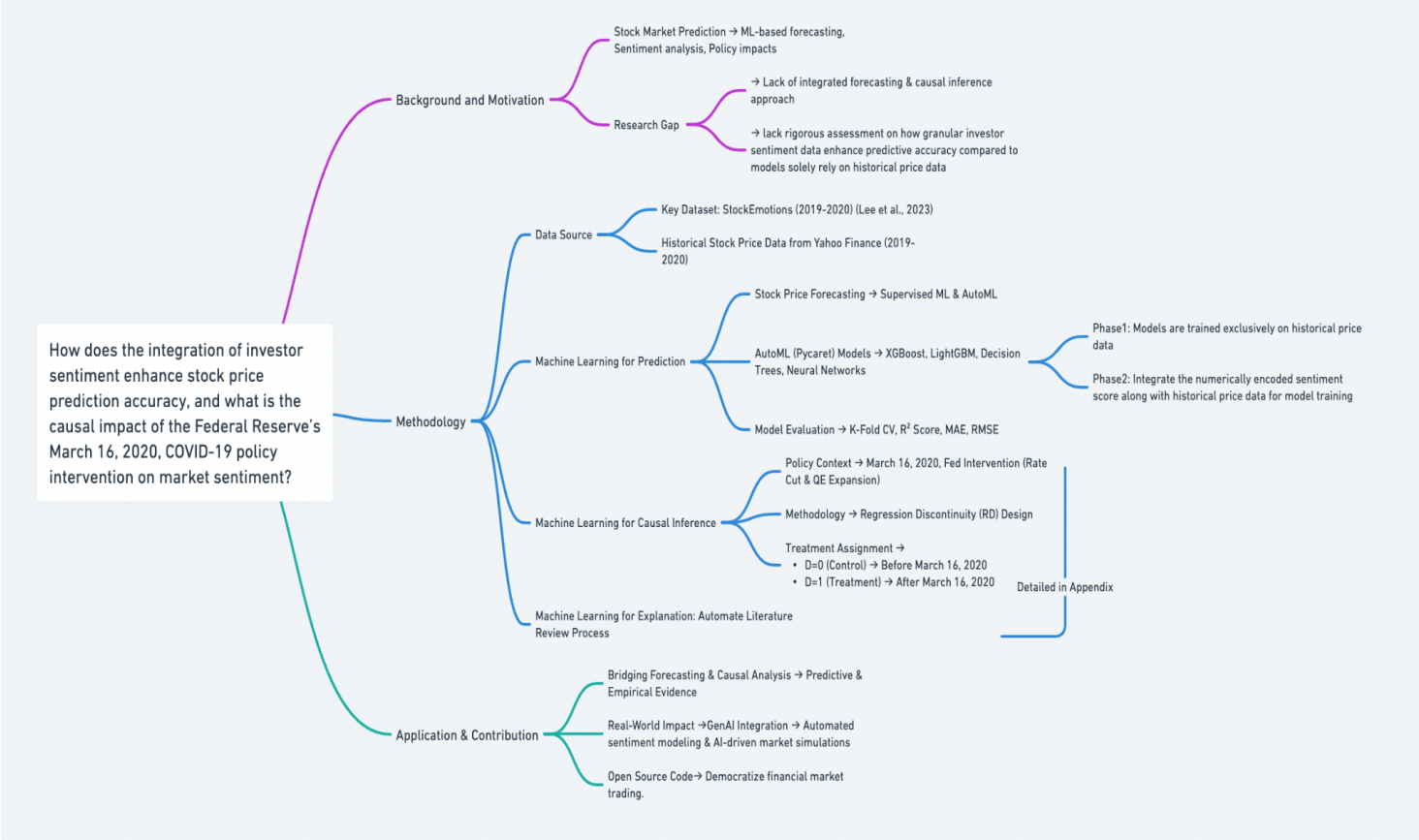


Figure 1: Illustration of wholistic Research Design and Methodology (This graph is hand-made using Whimsical software).

Data preprocessing and EDA:

The primary dataset consists of historical stock prices for multiple companies, each stored as a separate CSV file. To create a comprehensive dataset, this study applies the following preprocessing steps:

- Merging Multiple Stock Price CSV Files: Each company's stock prices are stored in individual CSV files. These files are merged into a single structured dataset: "merged_stock_prices.csv", where each row corresponds to a specific stock ticker, date, and price information. This step allows multi-stock analysis, enabling cross-company comparisons and ensures consistent structure for model training and visualization.
- EDA: This study visualizes closing price movements over time for all companies in the dataset. Furthermore, this research computes short-term (10 day) moving

average and long-term (50 day) moving average, as it allows financial analysis to smooth price fluctuations and highlight underlying trends.

As illustrated in Figure 1, the study uses machine learning for prediction as the main method and included the causal inference part in the appendix.

Main Method: Machine Learning for Prediction

1. Predicting Stock Price and Feature Engineering

Forecasting stock market movements is a complex problem due to the inherent volatility and non-linearity of financial data (Nabipour et al., 2020). Traditional econometric models often struggle to capture intricate patterns within stock price fluctuations. To address these challenges, this study leverages Supervised Machine Learning (ML) techniques, incorporating an Automated Machine Learning (AutoML) framework to systematically evaluate multiple predictive models and select the most optimal one. This approach ensures that the model selection process is data-driven, efficient, and reproducible, reducing potential bias in manually tuning hyperparameters or arbitrarily selecting algorithms.

Feature Engineering

To enhance predictive power, key feature engineering techniques were applied:

- Daily Returns: Computed as the percentage change in closing price.
- Moving Averages (MA): Short-term (10-day) and long-term (50-day) averages to capture trend dynamics.
- Lagged Returns: Previous day's return (Lagged_Return) was included as an input feature to capture short-term momentum effects.
- Sentiment Score:
  - Emotion Encoding: The original "train_stockemo.csv" dataset contained 12 categorical emotion labels in the emo_label column. To transform these qualitative labels into numerical values, a custom mapping was applied based on their anticipated sentiment impact on stock movements. They were mapped to numerical sentiment scores based on their expected market impact. Positive emotions such as *Optimism* and *Excitement* were assigned higher scores (+3), while negative emotions like *Panic* and *Depression* received lower scores (-3). Mildly positive emotions (*Belief*, *Surprise*) were assigned +1, and moderately negative emotions (*Anger*, *Disgust*, *Confusion*) were assigned -1 to -2. Neutral sentiments (*Ambiguous*) were mapped to 0 to reflect market uncertainty.
  - Integration of Stock Price and StockEmotions datasets: To establish a direct relationship between sentiment and stock performance, the processed sentiment scores were merged with historical stock price data. The merging was performed on both date and ticker to ensure proper alignment.

After feature construction, missing values (arising from rolling-window calculations) were removed, and the dataset was structured to align with PyCaret's AutoML framework.

2. AutoML Approach: Model Selection Using PyCaret

Rather than manually selecting a specific predictive model, this study employs PyCaret's automated model selection pipeline. This approach systematically evaluates multiple supervised machine learning models, including: Linear Models: Ordinary Least Squares Regression, Ridge Regression, Lasso Regression. Tree-Based Models: Decision Trees, Random Forest, Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGBoost), LightGBM. Ensemble Learning: Stacking and Blending techniques. Neural Networks: Fully Connected Feedforward Networks.

- To assess the impact of sentiment information on stock price prediction, the AutoML pipeline was applied in two phases.
- In the first phase, the model was trained exclusively on historical price data, incorporating features such as daily returns, lagged returns, and moving averages (MA-10, MA-50). This served as a baseline model, capturing technical patterns in stock price movements without considering market sentiment.
- In the second phase, the same AutoML pipeline was employed with an additional sentiment-based feature: the numerically encoded sentiment score, derived from stock-related textual data. Therefore, the model is able to account for both quantitative market trends and qualitative investor sentiment.

3. Model Training and Evaluation

Once the best-performing model is identified, it is trained on 80% of the dataset, while 20% is used for testing. The following cross-validation and evaluation techniques are employed: K-Fold Cross-Validation (K=5), R^2 Score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Logarithmic Error (MSLE).

4. Prediction Method Justification

The AutoML approach was selected for the following reasons:
Firstly, for automated model benchmarking: It ensures objective selection of the best predictive model instead of relying on arbitrary choices.
Secondly, for scalability and reproducibility: The method can be easily adapted to different financial instruments and market conditions.

Results:

This section presents the results of the two-phase predictive analysis. The AutoML pipeline systematically selected the best-performing models in both phases. In the first phase, the model was trained solely on historical stock price data to establish a baseline, and the best model selected was the linear regression model. In the second phase, sentiment scores were integrated to assess the impact of investor emotions on stock price forecasting, and the best model selected was the Extra Trees Regressor. We find that the phase 2 model significantly outperform purely price-based model (phase 1 model), demonstrating that granular investor sentiment serves as a valuable leading indicator in financial forecasting.

Phase 1: For this part, we use only historical price data. The **linear regression model** was identified as the best fit for prediction.

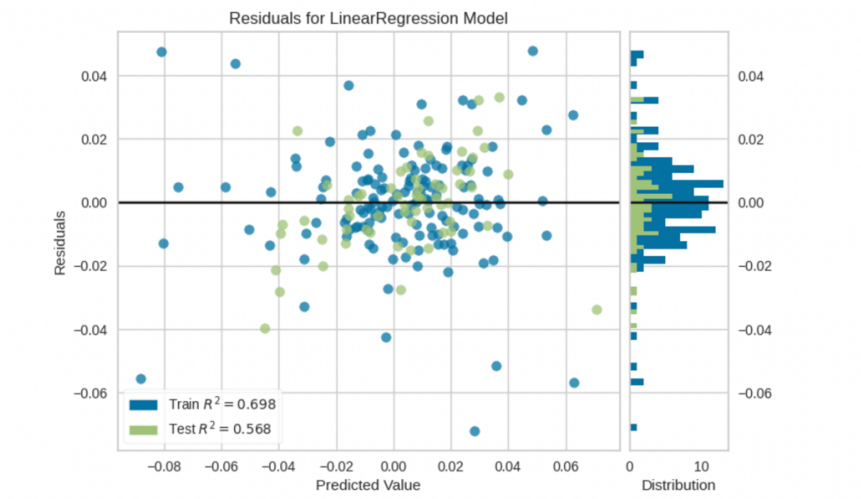


Figure 2: The residual plot for the linear regression model.

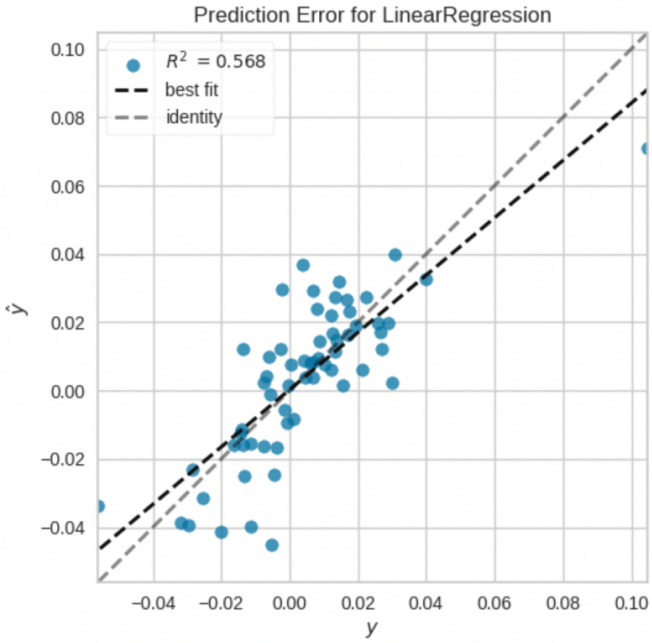


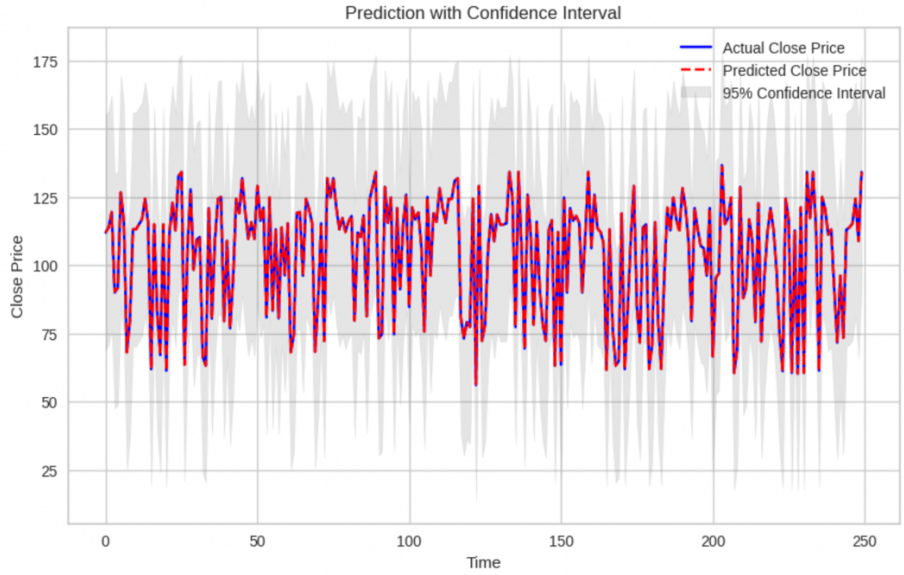
Figure 3: The prediction error plot for the linear regression model.

Figure 2 shows the training R^2 (0.698) is higher than the test R^2 (0.568), meaning the model performed better on training data than test data, indicating mild overfitting. The residuals are roughly centered around zero, with no obvious skew or severe heteroskedasticity. The prediction error plot (Figure 3) shows the best-fit line (black dashed line) deviates slightly from the identity line (gray dashed line), indicating some systematic prediction error. Points spread away from the identity line, particularly for

larger values, implying higher prediction error for extreme values.
Model performance metrics:

- Mean Absolute Error (MAE): 0.0117
- Mean Squared Error (MSE): 0.0003
- Root Mean Squared Error (RMSE): 0.0166
- R-squared (R^2): 0.568

Phase 2: For this part, the encoded investor sentiment data was incorporated alongside historical stock price data to evaluate its contribution to predictive performance. The AutoML pipeline identified the **Extra Trees Regressor** as the best-performing model for this integrated approach.




```
Mean Absolute Error (MAE): 0.05306456970214984
Mean Squared Error (MSE): 0.02292076648697909
Root Mean Squared Error (RMSE): 0.15139605836011416
R-squared (R2): 0.9999520421588884
```

Figure 4: The prediction result with confidence interval of the Extra Trees Regressor model.

Figure 5: The residuals plot of the Extra Trees Regressor model.

Figure 4 demonstrates that the actual and predicted closing prices align closely, demonstrating a high degree of accuracy in forecasting. The R^2 value improved significantly to 0.9, suggesting that the integrated model captures nearly all the variance in stock price movements. The error metrics (MAE, MSE, and RMSE) are also notably lower than in phase one. Therefore, we find that models trained with sentiment metrics has significantly larger prediction power compared to models trained solely on historical prices. Figure 5 shows the actual residuals (green line) fluctuate around zero, with certain spikes where the model fails to anticipate sudden shifts, likely due to external macroeconomic shocks or high-impact news events during the 2019-2021 covid highly volatile period.

Model performance metrics:

- Mean Absolute Error (MAE): 0.0531
- Mean Squared Error (MSE): 0.0229
- Root Mean Squared Error (RMSE): 0.1514
- R-squared (R^2): 0.99995

Intellectual Merit:

Contribution to existing literature:

This research makes significant academic contributions by integrating machine learning-based stock price forecasting with causal inference through Regression Discontinuity (RD) Design, offering a dual approach to understanding stock market dynamics. By incorporating the StockEmotions dataset, a unique resource containing fine-grained investor emotions and time-stamped financial discussions, the study systematically quantifies sentiment and demonstrates its predictive value within an automated machine learning pipeline, which is a novel combination in financial econometrics and behavioral finance.

Additionally, by incorporating GenAI and state-of-the-art machine learning techniques,

this research aligns with the frontier of financial technology, offering new methodologies for policy evaluation, algorithmic trading, and market behavior modeling.

Limitation and Future Research Extension

For the prediction model, this study has several limitations which suggest directions for future research:

1. **Potential overfitting in phase 2 model (sentiment integrated extra tree regressor model)**: The high R square value is 0.9. Since stock market movements are influenced by random and exogenous factors, such a high explanatory power may indicate model oversensitivity to training data rather than true generalizability.
2. **Limited generalizability across different market conditions**: This study focuses on 2019–2021, a period of high volatility due to COVID-19, and the finding suggests a high predictive power of investor sentiment. However, the model may not perform as well in normal market conditions where sentiment may have less predictive power. Future research should evaluate model performance across different time periods, including bullish, bearish, and stable markets, to assess robustness.
3. **The sentiment aggregation score could oversimplify market psychology**: The emotion-to-score mapping assigns fixed numerical values to emotions (e.g., optimism = +3, panic = -3), assuming linear effects on stock prices. Future research could apply context-aware sentiment analysis, such as: sentiment embeddings from deep learning models and market-dependent sentiment weighting (e.g., using rolling correlations).
4. **The Regression Discontinuity design for policy intervention (detailed in the appendix) is inherently weak due to the time-based cutoff.** There are numerous confounders that could influence the results. As a result, this part did not yield statistically significant findings in the study. Future research should focus on more robust causal inference methods for stock price prediction, potentially leveraging alternative indicators or instrumental variables to better isolate the causal impact of financial policies on market behavior.

Practical Impacts

Industry Application:

The insights from this research have broad applications across finance, policymaking, and AI-driven financial decision-making. For institutional investors and hedge funds, sentiment-driven stock forecasting models can enhance algorithmic trading strategies, improving risk assessment during volatile periods. Regulatory bodies and central banks can use the causal analysis framework to evaluate the effectiveness of monetary policies on market psychology, helping to design interventions that stabilize financial markets.

Social Benefits:

This study's findings could justify the idea that early detection of market sentiment shifts can help investors and financial institutions anticipate volatility, reducing the risk of sudden market crashes and contributing to crisis preparedness. Moreover, the integration of individual-level sentiment into prediction models could democratize financial market trading. Additionally, the open-source data and code could reduce information asymmetry and allow retail investors to make more informed decisions.

AI Governance and Ethical Considerations:

This study utilizes open-source machine learning tools (PyCaret, Scikit-Learn) and also provides open-source code, making sentiment-based stock prediction accessible to a broader audience. The methodology can be replicated and deployed in educational settings, fintech applications, and public financial literacy programs, promoting inclusivity in AI deployment. It is significant to ensure the public accessibility of financial forecasting models, in order to foster inclusivity in AI-driven financial decision-making and prevent AI from being a tool exclusively for elite market participants. Connecting to the field trip experience, combining technology with art and finance can both promote inclusivity by making complex art/financial insights more accessible, interpretable, and engaging, empowering individuals. Regardless of their art/financial expertise, individuals could make informed decisions in an increasingly AI-driven economy.

This research aligns with SDG9: Industry, Innovation, and Infrastructure by bridging AI, finance, and social impact. It showcases a scalable machine learning approach that can be applied across industries, including macroeconomic forecasting, risk assessment, and crisis intervention strategies. From a long-term perspective, it promotes responsible AI deployment in fintech, ensuring that machine learning applications in finance are ethical, transparent, and for the collective good. However, the high energy consumption and carbon emissions associated with the large-scale deployment of AI models could pose challenges to environmental sustainability, highlighting the need for further research into energy-efficient AI solutions.

Data and Code Availability Statement

The data and code supporting the findings of this study are available at the following GitHub repository:

https://github.com/Rising-Stars-by-Sunshine/Yiwei_Zhang_Final

Data Availability

The processed stock price and Stock Emotions datasets are available in the repository under the /data directory. If there are any access restrictions (e.g., due to licensing or privacy concerns), please refer to the repository's README for more details on data usage and permissions.

Code Availability

All scripts and codes used for explanatory data analysis are publicly available in the repository under the /code directory.

Reproducibility

To ensure reproducibility, all dependencies and setup instructions and documented

and can be found in the repository's README.md.

Bibliography

Abdul-Mageed, Muhammad, and Lyle Ungar. "EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Regina Barzilay and Min-Yen Kan, 718–28. Vancouver, Canada: Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/P17-1067.

Ashfaq, Nazish, Zubair Nawaz, and Muhammad Ilyas. "A Comparative Study of Different Machine Learning Regressors For Stock Market Prediction." arXiv, April 14, 2021. https://doi.org/10.48550/arXiv.2104.07469.

Cortis, Keith, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. "SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News." In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, edited by Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, 519–35. Vancouver, Canada: Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/S17-2089.

Chen, Zhenghua, Min Wu, Alvin Chan, Xiaoli Li, and Yew-Soon Ong. "Survey on AI Sustainability: Emerging Trends on Learning Algorithms and Research Challenges [Review Article]." *IEEE Computational Intelligence Magazine* 18, no. 2 (May 2023): 60–77. https://doi.org/10.1109/MCI.2023.3245733.

Lee, Jean, Hoyoul Luis Youn, Josiah Poon, and Soyeon Caren Han. "StockEmotions: Discover Investor Emotions for Financial Sentiment Analysis and Multivariate Time Series." arXiv, February 6, 2023. https://doi.org/10.48550/arXiv.2301.09279.

Li, Tong, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. "MASTER: Market-Guided Stock Transformer for Stock Price Forecasting." arXiv, December 23, 2023. https://doi.org/10.48550/arXiv.2312.15235.

Maqbool, Junaid, Preeti Aggarwal, Ravreet Kaur, Ajay Mittal, and Ishfaq Ali Ganaie. "Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach." *Procedia Computer Science*, International Conference on Machine Learning and Data Engineering, 218 (January 1, 2023): 1067–78. https://doi.org/10.1016/j.procs.2023.01.086.

Nabipour, M., P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and Shahab S. "Deep Learning for Stock Market Prediction." *Entropy* 22, no. 8 (August 2020): 840. https://doi.org/10.3390/e22080840.

Ren, Rui, Desheng Dash Wu, and Tianxiang Liu. "Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine." *IEEE Systems Journal* 13, no. 1 (March 2019): 760–70. https://doi.org/10.1109/JSYST.2018.2794462.

Sidi, Lior. "Improving S&P Stock Prediction with Time Series Stock Similarity." arXiv, February 8, 2020. https://doi.org/10.48550/arXiv.2002.05784.

Wang, Q., Y. Zhou, and J. Shen. "Intraday Trading Strategy Based on Time Series and Machine Learning for Chinese Stock Market." arXiv, March 24, 2021.

https://doi.org/10.48550/arXiv.2103.13507.

Appendix

Par1: Machine Learning for Casual Inference: Evaluating the Impact of Federal Reserve's COVID-19 Intervention on Investor Sentiment: A Regression Discontinuity Design Approach.

1.1 Policy Context

The COVID-19 pandemic triggered unprecedented financial market volatility, prompting urgent policy responses. On **March 15, 2020**, the Federal Reserve implemented an emergency rate cut, lowering the federal funds rate to near zero and launching a $700 billion quantitative easing (QE) program. This intervention aimed to stabilize financial markets and restore investor confidence. However, its immediate effect on investor sentiment remains an open question.

This study employs Regression Discontinuity (RD) Design to causally estimate the effect of the Fed's emergency intervention on investor sentiment, leveraging the StockEmotion dataset, which contains time-stamped investor sentiment data from 2019-2020.

1.2 Research Design

We apply RD to study how the Federal Reserve's emergency COVID-19 intervention affected investor sentiment, by leveraging March 16, 2020, as the cutoff date for treatment assignment.

- Policy: Federal Reserve's March 15, 2020, emergency intervention (rate cut & QE expansion).
- Implementation Threshold: The date of intervention (March 16, 2020), when markets opened following the policy announcement.
- Outcome Variable (Y): Investor Sentiment, measured via:
  Senti_label: Categorized as bullish, bearish, or neutral, derived from StockTwits financial text.
  Emo_label: Captures emotional intensity (e.g., excitement, fear).
- Cutoff Point for Treatment Assignment:
  Cutoff Date (Z): March 16, 2020
  Investors whose sentiment was recorded before March 16, 2020, belong to the control group (D = 0).
  Investors whose sentiment was recorded on or after March 16, 2020, belong to the treatment group (D = 1).

1.3 Results:

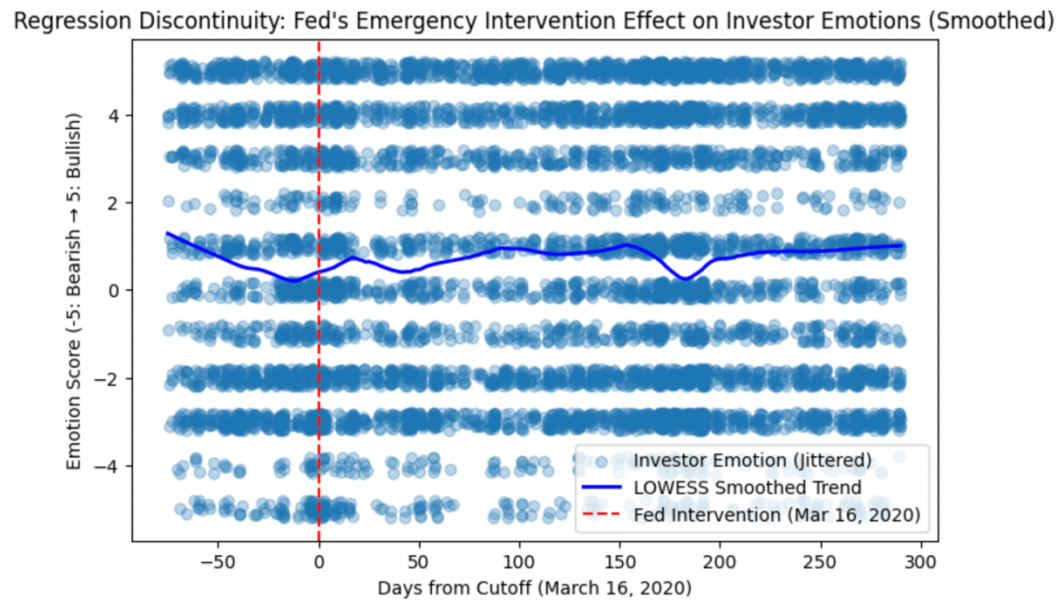


Figure 6: Regression discontinuity results with a time span of a whole year

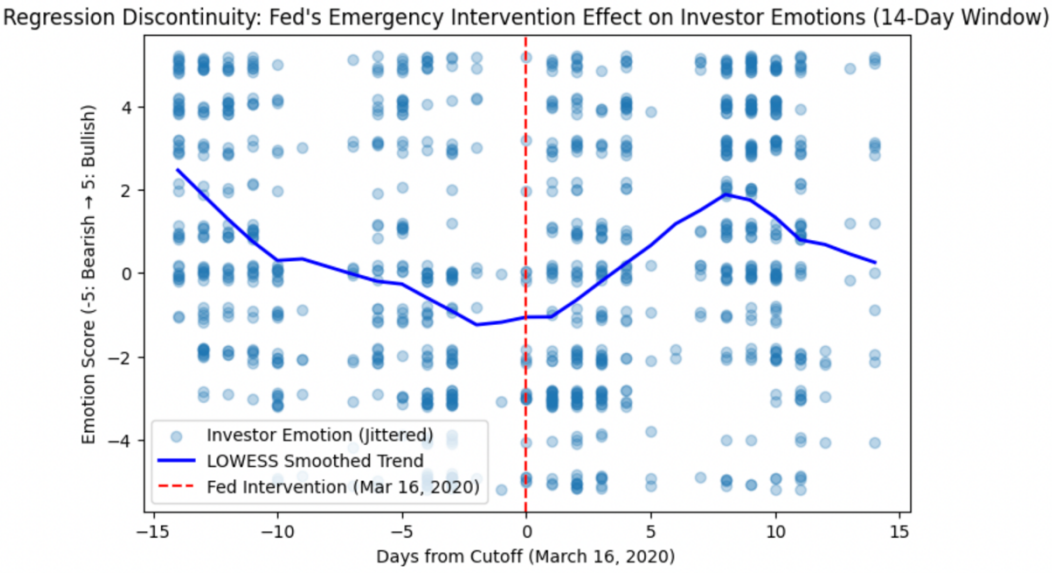


Figure 7: Regression discontinuity results with a time span of 14-day window

Part 2: Machine Learning for Explanation: Automated Literature Review using arXiv API.

This study implements an automated literature review pipeline that: Queries arXiv's API using search terms like "stock market prediction AND sentiment analysis". It also extracts key metadata (e.g., title, abstract) and saves results to a structured dataset (literature_data.csv) for further analysis. By generating the word cloud graph and literature network graphs, it substantially reduces manual effort in this section and provide a quantitative overview of research trends in ML-based stock prediction.

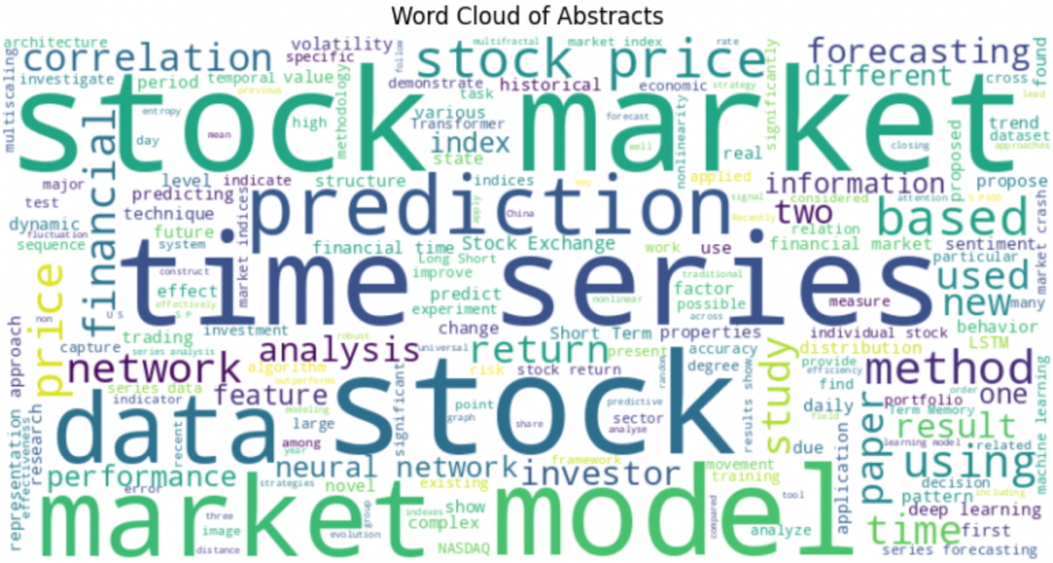


Figure 8: Word Cloud of Abstracts from Queried Relevant arXiv literature

Figure 9: Literature Review to Identify the most influential terms in the network using centrality measures.