Analyzing impact of twitter sentiment on stock market dynamics using spectral clustering and deep learning

Rahul Gupta, Vagmin Viswanathan

Department of Mathematics, Dartmouth College

Abstract

Modern social media platforms have become influential forums for discussing and disseminating public opinion on topics like the financial markets. Previous literature has extensively demonstrated the impact of investor sentiment on stock price movements. This study seeks to further characterize the underlying structure of the relationship between time series data of equity price metrics and sentiment streams during the pandemic. We aggregated web scraped financial tweets and employed Twitter Sentiment Analysis techniques.We then constructed a network and employed spectral clustering to explore structures within the data. Then, to ascertain the predictive value of our sentiment metrics, we utilized deep learning techniques. The findings of this research provide insights into the dynamics of social media sentiment in financial forecasting, offering a new perspective on market analysis.

Introduction

Background

The stock market is a complex system influenced by economic indicators, corporate performance, geopolitical events, and investor sentiment. Stock price movements reflect the collective actions of buyers and sellers, driven by expectations for future earnings and growth. The Efficient Market Hypothesis (EMH) [1] posits that stock prices fully reflect all available information. Since 2020, social media platforms and no-fee retail trading apps have amplified the impact of investor sentiment on financial markets.

Figure: Example tweet

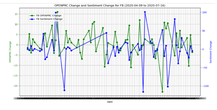

Figure: Distribution of tweets across companies

Platforms like X (formerly Twitter) have become significant sources of real-time news, influencing investor behavior and affecting stock prices. Research using Twitter Sentiment Analysis (TSA) techniques and deep learning has constructed time series of investor sentiment, which can be applied to financial market prediction.

Data collection

We constructed our sentiment time series by collecting Twitter data on financial markets during the Covid-19 pandemic, focusing on high market volatility and social media activity. Due to web scraping restrictions on X, we used existing datasets, notably Taborda et al.'s dataset of around a million finance-related tweets from April 9 to July 16, 2020 [2]. We correlated sentiment with stock market performance using historical data from the Wharton Research Data Services (WRDS) database [3], including daily prices, volume, SIC codes, and other financial metrics for NASDAQ companies.

Models and Methods

Structure of Data

Spectral Clustering

To understand the relationship between our Twitter sentiment time series and stock market movements we first conduct a χ^2 test, where we found statistically significant correlations. To visualize what companies are impacted by sentiment in similar ways, we construct an adjacency matrix using Pearson correlations. We then map this with a Gaussian kernel to network edges. Following Luxberg [4] we apply spectral clustering using the symmetrized Laplacian to this network. We match these clusters to economic sectors with an Adjusted Rand Index (ARI). We additionally consider a rolling window of time frames to reduce the impact of overarching market dynamics. Then, we analyze the entire stock time series in the same manner, applying Principal Components Analysis (PCA) to reduce the dimensionality of the data for computational efficiency.

Correlation

$$\rho_{S_A,S_B} = \frac{\sum_{i=1}^{n}(S_{A,i} - \bar{S}_A)(S_{B,i} - \bar{S}_B)}{\sqrt{\sum_{i=1}^{n}(S_{A,i} - \bar{S}_A)^2}\sqrt{\sum_{i=1}^{n}(S_{B,i} - \bar{S}_B)^2}}$$

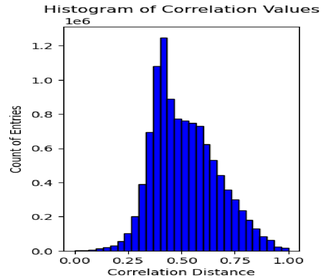

Figure: Correlation Histogram

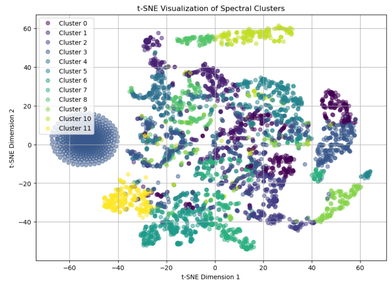

Figure: t-SNE Visualization of clusters

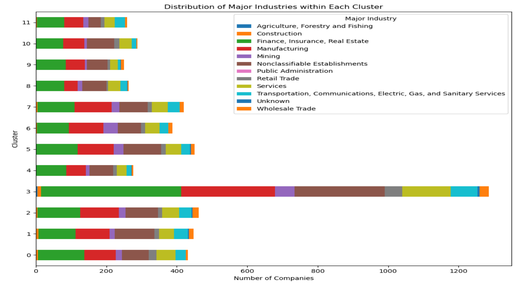

Figure: Distribution of industries within clusters

Machine Learning Analysis

We utilized Long Short-Term Memory (LSTM) neural networks to assess whether Twitter sentiment can predict stock prices for heavily discussed companies. LSTMs are suitable for time series forecasting, learning long-term dependencies. We trained LSTM models with features from Twitter sentiment scores and traditional stock market metrics to forecast future stock prices.

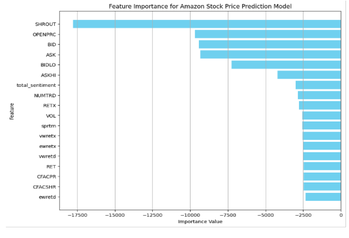

Figure: Feature Importance for Amazon

Feature Importance

Feature importance is used to understand the impact of variables, including sentiment scores, on stock price predictions. For LSTM, permutation feature importance quantifies the change in prediction error after permuting feature values. First, train the model and measure its baseline MSE ($\text{MSE}_{\text{baseline}}$). For each feature, permute its values, re-evaluate the model, and record the new MSE (MSE_f). Feature importance is calculated as:

$$\text{Importance}(f) = \text{MSE}_f - \text{MSE}_{\text{baseline}}$$

A larger discrepancy indicates higher importance. This method is computationally intensive but considers both direct and indirect effects.

Results and Analysis

Our sentiment analysis revealed significant variations in public sentiment towards different companies, correlating these with major events and stock price movements. We investigated "sentimental" stocks, where sentiment scores significantly predict price movements, by analyzing the top 50 stocks by tweet count using an LSTM model.

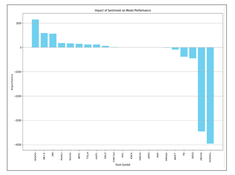

Figure: Sentimental Stocks

The histogram shows that sentiment scores were crucial for GOOGL, AMZN, QQQ, and MRNA. Moderna's sentiment-driven movements align with its vaccine development during COVID-19. Tech giants GOOGL and AMZN were influenced by sentiment due to their roles during lockdowns.

QQQ, a NASDAQ-100 index ETF, showed significant sentiment impact, reflecting the importance of technology companies during the pandemic. These findings highlight the value of social media sentiment analysis in financial forecasting and the need for a company-specific approach.

.

Conclusion

For the time period considered, we found significant correlations between investor sentiment and stock prices, particularly clustered together for tech stocks. The variable outcome of stock market predictions using machine learning suggest company-specific factors influence the effectiveness of sentiment as an indicator. These findings suggest that for sentimental stonks market observers can glean useful price signals.

Future Directions

Our research shows the significant impact of Twitter sentiment on stock prices. We plan to extend our study with a larger tweet dataset and include sentiment data from Reddit, TikTok, Instagram, and Mastodon. We also seek to pair real time stock and options data to better capture price movements due to sentiment. This could lead to the development of behavioral economic models capturing interplay of market forces and human psychology.

References

[1] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," in *The Journal of Finance*, vol. 25, no. 2, pp. 383-417, May 1970. DOI: 10.2307/2325486.

[2] B. Taborda, A. de Almeida, J. C. Dias, F. Batista, and R. Ribeiro, "Stock Market Tweets Data," IEEE Dataport, Apr. 15, 2021.

[3] Wharton Research Data Services, "WRDS," Wharton School of the University of Pennsylvania.

[4] U. von Luxburg, "A Tutorial on Spectral Clustering," Max Planck Institute for Biological Cybernetics, Technical Report No. TR-149, Aug. 2006.

[5] R. Braun, G. Leibon, S. Pauls, and D. Rockmore, "Partition Decoupling for Multi-gene Analysis of Gene Expression Profiling Data," 2011.