Predicting Stock Market Movements through Social Media Sentiment Analysis: A Data-Driven Approach to Understanding Public Perception and Financial Markets

Aryan Bhardwaj & Tyler Gomez Riddick

June 3, 2024

1 Research Question

How does public perception of companies as reflected on social media platforms relate to their stock price movements, and how can sentiment analysis of social media data be utilized to predict trends in stock market behavior for these companies?

2 Motivation

The motivation for exploring the relationship between public perception on social media and stock price movements is multifaceted. By utilizing sentiment analysis and real-time data from social media platforms, we aim to uncover valuable insights into market dynamics and develop predictive models that can predict stock market movement. The potential to enhance understanding of market behavior provides a strong impetus for this project.

3 Data

For this project, we will be pulling data from a variety of sources. In order to gauge public perception of companies of social media, we will be scraping the data from social media websites. The main social media sites we will be looking at are Reddit, X, LinkedIn, and Instagram. For all of these websites, we want to explore both current and historical posts to investigate past and current company sentiment.

For Reddit, we will be using the Reddit API via PRAW (Python Reddit API Wrapper). PRAW is a Python package that facilitates pulling data from Reddit into an easily-digestible Python-compatible format. We will primarily be examining posts in specific subreddits pertaining to the companies we are analyzing. The documentation for PRAW can be found here: PRAW Documentation. The Reddit API requires a Reddit account

The platforms X, LinkedIn, and Instagram all have APIs, but we will not be using either of them in this research. The X API has several levels of varying pricing, but the Free version disallows reading posts. The LinkedIn API is severely limited in its capabilities, and does not allow pulling posts. Thus, we will be scraping data directly from both websites. For X, we plan on using Playwright as outlined in this article from ScrapFly. For LinkedIn, we will primarily utilizing Selenium as touched on in this Medium article by Hugo Torché. This is what we plan to use to scrape Instagram data as well. We may also use Selenium for X web scraping, though this depends on how well we are able to implement these scrapers. For each of these websites, an account is needed in order to read the posts. We will most likely set up dummy accounts to accomplish this.

For this capstone, we will also be examining financial data. We will pull these data primarily from Yahoo Finance which gives us access to both historical and current financial performance data and stock prices. This data can be downloaded directly from the website as a CSV or JSON file. These data will be connected to the sentiment analysis through date and relevant company.

4 Methodology

4.1 Statistical Thinking

Statistical thinking is important to understand the relationship between social media sentiment and stock price movements. This would involve:

Descriptive Statistics: Summarizing the central tendency, dispersion, and distribution of both sentiment scores and stock price changes.

Correlation Analysis: Using Pearson or Spearman correlation coefficients to quantify the strength and direction of the relationship between sentiment scores and stock price movements.

Hypothesis Testing: Conducting tests (e.g., t-tests, ANOVA) to determine if there are significant differences in stock price movements based on different levels of sentiment.

4.2 Data Visualization

Data visualization techniques will be used to explore and present the data in a clear and insightful manner.

Time-Series Plots: Visualizing the sentiment scores and stock prices over time to identify trends, patterns, and anomalies.

Scatter Plots: Displaying the relationship between sentiment scores and stock price changes.

Heatmaps: Showing the correlation matrix between various features such as sentiment scores, trading volumes, and stock prices.

Interactive Dashboards: Using Power BI to create dashboards that allow for interactive exploration of the data.

4.3 Data Engineering

Data engineering is crucial for collecting, processing, and preparing the data for analysis.

Data Collection:

X (Twitter) Web Scraper: Collect tweets mentioning selected companies, ensuring to gather metadata such as timestamp, retweets, and likes.

Reddit API & PRAW: Collect posts and comments mentioning selected companies, and gathering metadata (just like the Twitter web scraper) like timestamp and upvotes.

Instagram and LinkedIn Web Scrapers: Collect posts regarding the select companies, and, like the above, gathering relevant metadata.

Stock Market Data: Collect historical stock price data, including open, close, high, low, and trading volume.

Data Cleaning:

Twitter Data: Remove noise by filtering out URLs, mentions, hashtags, and non-English tweets. Standardize text by converting to lowercase, removing stopwords, and performing tokenization.

Reddit, Instagram, & LinkedIn Data: Filter out URLs, mentions, hashtags, and non-English posts for sentiment analysis. Also, standardizing the text, removing stopwords, and performing tokenization.

Stock Data: Ensure consistency by aligning stock price data with corresponding tweet timestamps.

Feature Engineering:

Sentiment Scores: Use sentiment analysis models to assign sentiment scores (positive, negative, neutral) to tweets.

Aggregated Metrics: Create daily or weekly sentiment averages and other features such as moving averages and trading volume changes.

4.4 Machine Learning

Machine learning techniques will be used to build predictive models that leverage sentiment data to forecast stock price movements.

Sentiment Analysis: Utilize pre-trained NLP models like VADER, TextBlob, or BERT to analyze the sentiment of tweets. Aggregate sentiment scores to create daily sentiment indicators for each company.

Predictive Modeling:

Time-Series Forecasting: Employ models such as ARIMA, Prophet, or LSTM to predict future stock prices.

Feature Selection: Incorporate sentiment scores, historical stock prices, and other relevant features into the models.

Model Training and Evaluation: Train models on historical data and evaluate their performance using metrics like MAE, RMSE, and MAPE. Perform cross-validation to ensure robustness.

Model Interpretation: Use techniques like SHAP (SHapley Additive exPlanations) values to interpret model predictions and understand the contribution of sentiment scores to stock price movements.

4.5 Ethical Concerns

Addressing ethical concerns is essential to ensure responsible use of data and analysis techniques.

Privacy and Data Protection: Adhere to data protection regulations (e.g., GDPR) by anonymizing user data and ensuring secure data storage.

Bias and Fairness: Assess and mitigate potential biases in sentiment analysis models and ensure fairness in predictions.

Transparency and Accountability: Maintain transparency in data collection methods, model selection, and interpretation of results. Provide clear documentation and reports in final deliverable.

Impact on Stakeholders: Consider the potential impact of predictions and ensure that the research is conducted with integrity and respect for all parties involved.