Stock Movement Prediction Using Social Media Data

@abhi

https://github.com/letsbegincode/stock_trends_prediction.git

Abstract

This report presents an in-depth analysis of predicting stock movements using sentiment analysis of social media data. By leveraging *FinBERT*, a transformer-based model specifically fine-tuned for financial sentiment, we combine this with traditional financial indicators to develop a comprehensive dataset for machine learning. The report addresses several key challenges, including sentiment aggregation across multiple posts, handling missing data, and deduplication of content. Various machine learning models were evaluated to determine the most effective approach, with the *Random Forest* model emerging as the best performer due to its high accuracy and robustness against noisy data. Additionally, the report discusses feature engineering methods used to extract meaningful insights from both social media sentiment and financial metrics. It also explores future directions for improving the model, including the potential for regression-based stock price prediction, integration of diverse data sources, and advanced model architectures such as transformers for end-to-end prediction. By incorporating both sentiment data and financial time series, this framework has the potential to provide valuable predictions for stock market movements.

1 Introduction

The financial market is increasingly influenced by public sentiment, with social media platforms playing a pivotal role in shaping investor behavior. Platforms like Reddit, with large, active communities discussing financial trends, offer a rich source of real-time sentiment data. This study focuses on exploring the correlation between sentiment expressed in social media discussions and stock price movements, specifically for high-profile companies such as Tesla, Apple, and Amazon.

The primary goal of this research is to investigate whether sentiment analysis of social media content, particularly posts from subreddits such as *r/stocks*, *r/wallstreetbets*, and *r/investing*, can provide valuable insights into stock trends. To achieve this, we integrate sentiment data with traditional financial indicators such as stock price movements, trading volume, and other market metrics. By utilizing advanced machine learning techniques, including the FinBERT transformer model fine-tuned for financial sentiment analysis, we aim to enhance the accuracy of stock trend predictions. This combination of sentiment analysis and financial data could help identify patterns and offer a more nuanced understanding of stock market dynamics, providing investors with actionable insights for decision-making.

1.1 Data Collection Process

The data for this project was collected from two primary sources: Reddit for sentiment data and Yahoo Finance for stock market data. Each data source was utilized to capture distinct information that contributes to predicting stock movements.

1.1.1 Stock Data Collection from Yahoo Finance

To gather the necessary stock data for Tesla, Apple, and Amazon, the `yfinance` library was employed. `yfinance` is a Python package that allows easy access to historical market data, such as stock prices and trading volumes, directly from Yahoo Finance. The stock data collected included daily values of the following features:

- **Open Price:** The price at which the stock opened for trading on a given day.
- **Close Price:** The price at which the stock closed for trading on a given day.
- **High Price:** The highest price reached by the stock during the trading day.
- **Low Price:** The lowest price reached by the stock during the trading day.
- **Volume:** The total number of shares traded during the day, reflecting market activity.

The data was retrieved for specific time periods corresponding to the Reddit posts. By aligning the stock data with daily sentiment data from Reddit, we aimed to establish correlations between market activity and social media sentiment. For each stock, the `yfinance` library was queried using the stock's ticker symbols (e.g., `TSLA` for Tesla, `AAPL` for Apple, and `AMZN` for Amazon) and the date range relevant to the Reddit data collection. This allowed for the retrieval of historical stock prices and trading volumes, which were then cleaned, preprocessed, and merged with the sentiment data for model training.

1.1.2 Sentiment Data Collection from Reddit

The sentiment data was sourced from Reddit, specifically from subreddits such as `r/stocks`, `r/wallstreetbets`, and `r/investing`, which are known for discussions about stock market trends and financial news. The `PRAW` (Python Reddit API Wrapper) library was utilized to interact with Reddit's API and collect posts containing discussions related to Tesla, Apple, and Amazon.

The process began by specifying relevant queries for each of the three companies, targeting specific keywords such as ”Tesla”, ”Apple”, and ”Amazon” within the posts' titles and bodies. Data was collected from posts over a set period, ensuring that the sentiment data aligned with the stock data timeline. Each post retrieved from the targeted subreddits contained both textual content and metadata, including the post's creation time and user engagement metrics (such as upvotes and downvotes).

Once the data was collected, the sentiment of each post was determined using the `FinBERT` model, a transformer-based model fine-tuned specifically for financial sentiment analysis. `FinBERT` classifies posts into three categories: Positive, Neutral, or Negative, based on the sentiment expressed towards the company. The sentiment scores were further weighted by the upvotes received on each post, with higher upvotes indicating greater confidence in the sentiment. This approach helped aggregate sentiment over time, ensuring that the most influential discussions (those with higher engagement) had a larger impact on the overall sentiment for each day.

By analyzing posts from these subreddits, it was possible to capture public sentiment around each stock on a daily basis, which was then matched with the corresponding stock market data. The aggregation of sentiment was performed using a weighted average, considering the sentiment of each post and the number of upvotes as weights. This sentiment aggregation approach allowed for a comprehensive view of the overall market sentiment for Tesla, Apple, and Amazon, which was subsequently used as a feature for machine learning models.

1.2 Challenges and Solutions

- **Challenge: Identifying Valid Channels and Queries**
  Reddit contains numerous off-topic or irrelevant channels, making it crucial to filter for reliable finance discussions. To address this, a curated list of trusted subreddits was developed, and company-specific queries were refined to ensure the relevance of the collected data.

- **Challenge: Missing Titles and Text Columns**
  Many posts lacked content in either the *title* or *text* fields, limiting their utility for sentiment analysis. To overcome this, missing text fields were supplemented with titles when available, and posts with both fields empty were excluded from the analysis.

- **Challenge: Duplicate Entries**
  Duplicate or cross-posted content could distort sentiment scores. This was mitigated by deduplicating posts based on their *title* and *text* combinations, ensuring that each unique post was only considered once in the sentiment analysis.

- **Challenge: Aggregating Sentiments for a Single Day**
  On any given day, multiple posts about a single company might express conflicting sentiments. To address this, sentiments were aggregated using a weighted average, calculated as:

  $$\text{Aggregated Sentiment Score} = \sum(\text{Sentiment Score} \times \text{Post Score})$$

  Additionally, to refine this process further, the final sentiment of the day was determined by identifying the maximum sentiment score across all posts for that day. The optimal sentiment for the day was then computed as the average of these maximum scores, providing a more reliable representation of the overall sentiment.

- **Challenge: API Rate Limits**
  High-frequency requests to Reddit's API often led to throttling, disrupting data collection. To prevent this, a delay mechanism was incorporated into the requests, allowing for efficient data collection while ensuring compliance with API rate limits.

2 Feature Extraction

2.1 Sentiment Analysis Using FinBERT

A key innovation in this study was the use of FinBERT, a transformer model specifically fine-tuned for financial sentiment analysis. FinBERT is highly effective at identifying nuanced sentiment in finance-related text, making it ideal for classifying Reddit discussions. Sentiments were labeled as Positive, Neutral, or Negative, with accompanying confidence scores.

2.2 Features and Their Importance

The final dataset utilized a comprehensive set of features, each contributing significantly to the predictive power of the model. These features can be categorized into the following types:

- **Sentiment Features:**
  - **Sentiment Label:** This feature, derived from the FinBERT model, classifies each discussion as either Positive, Neutral, or Negative. The sentiment label helps capture the emotional tone of the posts and plays a crucial role in predicting stock price movement based on public sentiment.
  - **Sentiment Score:** This numerical value reflects the model's confidence in the assigned sentiment label. A higher sentiment score indicates stronger confidence in the sentiment classification, providing a quantifiable measure of sentiment strength.
- **Financial Features:**
  - **Open, High, Low, Close Prices:** These features track the stock's daily performance, capturing key indicators of market movement and volatility. The Open, High, Low, and Close prices provide a comprehensive view of daily market activity, essential for predicting price fluctuations.
  - **Volume:** This feature represents the total trading volume of a stock on a given day. It serves as a proxy for market activity and investor engagement, with higher volumes often signaling greater market interest and potential for price movement.
- **Temporal Features:**
  - **Date:** The date feature ensures the alignment of sentiment data with the corresponding financial metrics. By linking sentiment scores to specific trading days, this feature helps establish the temporal relationship between public sentiment and stock performance.
- **Combined Features:**
  - **Aggregated Sentiment Score:** This feature aggregates sentiment data for each day, providing a weighted average sentiment score that reflects the overall sentiment towards a company. The aggregated sentiment score is a crucial feature as it combines the influence of multiple posts into a single value, which is then used to predict stock trends.

3 Target Variable

The target variable was designed to predict directional price changes:

$$\text{Price Change} = \text{Close}_{t+1} - \text{Close}_t$$

- **1 (Up):** Indicates a positive price change.
- **-1 (Down):** Indicates a negative price change.
- **0 (Neutral):** No significant price movement.

This classification focuses on predicting trends rather than precise values.

4 Model Evaluation and Performance

4.1 Model Comparison and Insights

Several machine learning models were trained and evaluated to identify the most effective algorithm for predicting stock movements based on social media sentiment. The following table presents the accuracy and the best hyperparameters for each model tested:

Model	**Accuracy (%)**	**Best Parameters**
Logistic Regression	76.30	C=0.1, Solver=liblinear
Support Vector Machine	86.50	C=1, Kernel=rbf
Gradient Boosting	88.11	n_estimators=200, learning_rate=0.1
LightGBM	88.50	n_estimators=200, max_depth=7
Random Forest	90.00	n_estimators=200, max_depth=15

Table 1: Model Performance Comparison

As seen from the table, Logistic Regression had the lowest accuracy at 76.30%, followed by Support Vector Machine at 86.50%. Gradient Boosting (88.11%) and LightGBM (88.50%) showed improved performance, while Random Forest achieved the highest accuracy of 90.00%. The models with higher accuracy, particularly Random Forest, LightGBM, and Gradient Boosting, demonstrated better capacity in handling complex data relationships. Hyperparameter tuning was essential to achieve these results, especially for the more complex models like Gradient Boosting and Random Forest.

4.2 Model Selection and Justification

The Random Forest model was selected for this study based on several key advantages that made it particularly well-suited for the task of stock movement prediction:

- **High Accuracy:** The Random Forest model demonstrated a high accuracy of 90% on the test set, outperforming other models in terms of predictive performance. This high accuracy made it the most reliable model for capturing the complex patterns in stock movement.

- **Robustness to Noisy and Imbalanced Data:** Financial data, particularly social media sentiment, can often be noisy and imbalanced. The Random Forest algorithm, by aggregating multiple decision trees, is inherently robust to such noise and able to handle class imbalances effectively. This ensures the model remains stable and generalizes well even in the presence of outliers or skewed data distributions.

- **Strong Interpretability and Non-linear Modeling:** Unlike some other machine learning models, Random Forest provides an intuitive way to interpret feature importance, making it easier to understand the factors influencing stock movement predictions. Additionally, Random Forest is capable of modeling complex, non-linear relationships between sentiment, financial metrics, and stock price movements, which is crucial for capturing the dynamics of the financial markets.

5 Future Expansions

Several potential improvements can enhance the current framework for stock movement prediction based on social media sentiment:

- **Regression Models:** Extending the framework to **regression models** could allow the prediction of **exact price changes**, not just directional movements. This would provide more granular insights and better accuracy. Models like **Linear Regression**, **SVR**, or **LSTMs** can be used to predict continuous stock price values, improving decision-making for investors.

- **Diverse Data Sources:** Integrating additional data from **Twitter**, **financial news websites**, and **economic indicators** (like interest rates or inflation) could increase model robustness. Real-time sentiment from Twitter and breaking news from platforms like **Bloomberg** can enhance predictions and better capture market reactions.

- **Advanced Models:** Exploring **transformer-based architectures** such as **BERT** or **GPT** could boost prediction accuracy. These models can capture intricate patterns in textual data and be fine-tuned for financial contexts to enhance both sentiment analysis and market trend forecasting.

- **Multilingual Support:** Implementing **multilingual sentiment analysis** would enable the model to capture global market sentiment. Using tools like **mBERT** could help analyze sentiment from non-English sources, improving the model's effectiveness for international stock predictions.

- **Time Series Analysis:** Incorporating **time series analysis** alongside sentiment data could improve the model's performance by accounting for trends and patterns in historical stock prices. Techniques like **ARIMA** or **LSTM-based time series forecasting** could be employed to predict future stock movements based on historical data and sentiment evolution over time.

6 Conclusion

This study demonstrates the potential of leveraging advanced natural language processing models, such as FinBERT, in combination with machine learning techniques to predict stock movements based on social media sentiment. By using sentiment analysis from platforms like Reddit and integrating financial data, we have shown that sentiment plays a significant role in forecasting stock trends.

Among the models evaluated, the Random Forest model stood out, achieving the highest accuracy of 90% on the test set. Its ability to handle noisy, imbalanced data and model complex, non-linear relationships between sentiment and stock performance made it the most suitable choice for this task. The model's strong interpretability also allowed for valuable insights into the key factors driving stock price changes.

While the current framework focuses on classification-based predictions of stock direction (up, down, or neutral), future work will aim to extend this model to predict continuous price changes through regression models. Additionally, efforts will be made to incorporate data from other platforms, such as Twitter, as well as financial news sites, to enrich the feature set and improve model performance. Advanced techniques, including transformer-based models like GPT, may also be explored to enhance prediction accuracy further. Furthermore, the model could be expanded to support multilingual sentiment analysis, allowing it to capture global sentiment trends and their impact on stock movements.

In conclusion, this study highlights the promising intersection of social media sentiment analysis and stock market prediction, offering new opportunities for predictive modeling in the financial domain.

References

- Coyne, S., et al. (2017). Forecasting Stock Prices Using Social Media Analysis.
- Mehta, R., et al. (2012). Sentiment Analysis and Influence Tracking Using Twitter.
- Skuza, M., & Romanowski, A. (2015). Sentiment Analysis of Twitter Data for Stock Prediction.
- Yang, Y., et al. (2020). FinBERT: Financial Sentiment Analysis Using Pretrained Language Models.