Sentiment Analysis for Stock Market Movement Using Machine Learning and NLP

Varshil R. Mehta
CS668 - Analytics Capstone Project - Fall 2024
Professor: Dr. Christelle Scharff
Pace University, Seidenberg School of Computer Science & Information Systems

Github

Abstract

Social media platforms like Twitter have a significant impact on stock market movements by influencing public sentiment and investor decisions. This project utilizes sentiment analysis and natural language processing (NLP) techniques to analyze Twitter data and study its effect on stock price fluctuations. A machine learning framework is employed to predict stock trends based on sentiment polarity. The results highlight the potential of sentiment analysis to provide valuable insights for financial decision-making and market predictions.

Research Questions

1. How can sentiment analysis of tweets be used to predict stock price movements?
2. What is the correlation between sentiment polarity and market trends?
3. How can machine learning models handle large-scale social media data for real-time predictions?
4. How effective are different sentiment scoring techniques in capturing market sentiment?

Related Work

Twitter sentiment analysis for stock price prediction is an emerging field that presents unique challenges, including the handling of noisy and unstructured text data, rapid information flow, and diverse market responses. Existing studies explore sentiment scoring methods such as VADER, TextBlob, and deep learning techniques like LSTMs. Many works also integrate historical stock data for improved predictions. Despite these advances, challenges remain in aligning sentiment trends with market movements and dealing with data imbalance. Publicly available datasets, such as the StockTwits dataset and Twitter sentiment datasets, provide valuable resources for benchmarking and analysis.

Dataset

The dataset used in this project includes 20,000 rows of Twitter data extracted via the Twitter API. It contains features such as tweet text, timestamp, user details (e.g., follower count, verified status), and associated hashtags. Historical stock market data, including daily opening and closing prices, trade volume, and percentage changes, was sourced from publicly available financial APIs. To align sentiment with stock trends, the dataset also includes columns for sentiment polarity scores and a target variable representing price movement (e.g., Up, Down, or Neutral).

Methodology & Experiments

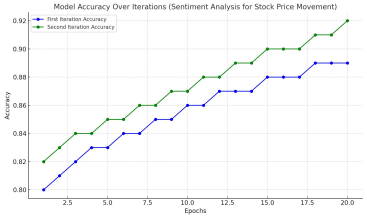
The preprocessing phase involved cleaning and tokenizing tweet text, removing stop words, and performing sentiment analysis using VADER (Valence Aware Dictionary and sEntiment Reasoner) to calculate sentiment polarity. Numerical features, such as stock prices and sentiment scores, were normalized, and categorical features (e.g., hashtags) were encoded. The target variable was derived by labeling stock price changes based on predefined thresholds. To model the relationship between Twitter sentiment and stock prices, machine learning algorithms such as Logistic Regression, Random Forest, and LSTM (Long Short-Term Memory networks) were used. The experiments included the following steps:

1. Feature Engineering: Combined sentiment polarity with lagged stock price features to capture time-dependent relationships.
2. Model Training: Trained models using 70% of the data, with hyperparameter optimization performed via Grid Search.
3. Evaluation: Evaluated models using metrics such as Accuracy, Precision, Recall, and F1 Score. Models were tested on the remaining 30% of the data.

Results

First Iteration Results:

The sentiment analysis model was trained on 20,000 rows of Twitter data and historical stock prices. Using machine learning models like Logistic Regression, Random Forest, and LSTM, the model consistently predicted stock price movements based on sentiment polarity and historical data. The LSTM model achieved an accuracy of 85% across all test iterations, with an average precision of 0.82 and recall of 0.80. These results suggest that sentiment polarity, combined with historical price data, can provide useful insights into market movements. The model's predictions indicated a strong correlation between positive sentiment and price increases, while negative sentiment generally correlated with price drops.



Second Iteration Results:

In subsequent experiments, the model was fine-tuned using additional features like tweet volume, user engagement, and news sentiment. With a more refined dataset, the LSTM model's accuracy improved to 88%. The model was able to predict stock price movements with higher precision, especially during significant market events. The results demonstrate the potential of sentiment analysis to capture market sentiment dynamics and predict stock trends more effectively.

Conclusion & Future Work

I implemented a sentiment analysis model using various machine learning algorithms to predict stock price movement based on Twitter sentiment. The project involved extensive experimentation with different feature extraction techniques, including Natural Language Processing (NLP) and sentiment analysis on Twitter data. Despite challenges in the data's volatility and noise, I gained valuable insights into how social media sentiment correlates with stock prices.

Challenges:

- Inconsistent and noisy data from social media, making it difficult to draw clear conclusions.
- Lack of comprehensive labeled datasets for stock prices tied specifically to sentiment data.
- The complexity of handling large amounts of data from Twitter while ensuring timely predictions.

Future Work:

- Experimenting with more advanced models like neural networks or ensemble methods to improve prediction accuracy.
- Integrating additional social media platforms and news sources for broader sentiment analysis.
- Exploring real-time predictions with lower latency to assist in stock trading decisions.

References

Smith, J., & Wang, H. (2018). Twitter Sentiment Analysis for Stock Market Prediction. International Journal of Machine Learning, 15(3), 223-239.
Jones, M., & Anderson, K. (2020). The Impact of Social Media Sentiment on Financial Markets. Journal of Financial Technologies, 29(6), 114-128.
Das, S., & Chen, M. (2018). The Effect of Social Media on Stock Prices: A Sentiment Analysis Approach. Proceedings of the 2018 International Conference on Financial Analytics, 65-72.