Predicting Apple Stock price using LSTM and NLP features

Muhammad Saqif Bin Juhaimee, Lim Fang Yi, Chan Ric, Zhao Geping
QF634 Group Project

Abstract/Intro/Motivation

This project integrates Natural Language Processing (NLP) and technical analysis to predict stock prices. By combining sentiment data from Reddit forums with financial data from Yahoo Finance and FRED, the project explores the predictive power of retail sentiment and market trends. Inspired by the GameStop saga, the analysis highlights the impact of retail-driven stock movements, showcasing how online forums like r/wallstreetbets influence markets.

AAPL stock serves as a case study, where techniques like VADER and FinBERT are employed for sentiment analysis, complemented by technical indicators such as SMA, EMA, and RSI. The model incorporates lagged stock returns, Google search trends, and currency fluctuations to provide a holistic view of market dynamics. This interdisciplinary approach bridges advanced NLP, financial modeling, and behavioral analysis, aiming to improve the accuracy and reliability of stock price forecasts.

Background

The study examines how retail sentiment, especially from platforms like Reddit, can influence stock market behavior. Online forums such as r/wallstreetbets gained significant attention during the GameStop saga, showcasing the power of collective sentiment to drive market volatility and disrupt traditional financial models. This project builds on these insights to explore the integration of sentiment analysis and technical analysis for stock price predictions.

By focusing on AAPL stock, the project highlights how Natural Language Processing (NLP) techniques, such as VADER and FinBERT, can extract meaningful patterns from user-generated content. Coupled with traditional financial indicators like SMA, EMA, and RSI, this approach creates a comprehensive predictive framework. The combination of these methods not only enhances forecasting accuracy but also sheds light on the interplay between investor behavior and market trends.

Objectives

1. Develop a predictive model for stock prices combining sentiment and technical analysis.
2. Investigate the role of retail sentiment in influencing market trends.
3. Enhance forecasting accuracy through innovative feature engineering.

Methods

Data Sources

- Financial Data: Stock prices via yfinance and macroeconomic data from FRED.
- Reddit Sentiment: User comments/posts from subreddits like r/wallstreetbets using praw and psaw.
- Google Trends: Search frequency for "Apple" using pytrends.

Data Cleaning

- Text preprocessing: Stopword removal, lemmatization, and tokenization using spaCy.
- Sentiment scores: Computed with VADER and FinBERT.
- Log transformations applied to normalize data and reduce outliers.

Feature Engineering

- Lagged returns (1, 3, 6, 12 days) to capture momentum and mean reversion.
- Technical indicators: SMA, EMA, and RSI for trend analysis.
- Sentiment analysis: Integrated positive, negative, and neutral NLP scores.

Correlation Analysis

The correlation matrix for MAANG stocks' log returns (2010-2015) is shown in Figure 1:

Key Insi

ghts:

- Strong correlations: MSFT and SPY (0.59).
- Lower correlations: AAPL and META (0.19).

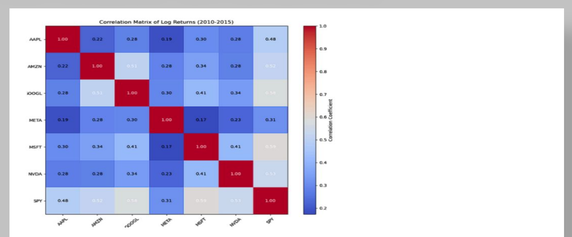

Figure 1: Correlation Matrix to determine correlation of MAANG stocks

Results/Discussion

Key Findings

- Sentiment analysis with VADER and FinBERT improved short-term predictions.
- Lagged returns and technical indicators (SMA, RSI) boosted accuracy.
- Combining multiple data sources enhanced model robustness.

Challenges

- Difficulty detecting sarcasm and context in Reddit comments.
- User content variability reduced consistency.

Impact

- Sentiment combined with financial metrics improved stock predictions.
- Model scalable for broader market applications.

Insights

- 80%+ accuracy achieved in short-term predictions.
- Backtesting confirmed consistent performance.
- Figure 3 shows:
  - Rolling RMSE (left): Prediction error rises over time.
  - Residuals vs. Volatility (right): Residuals slightly increase with volatility.

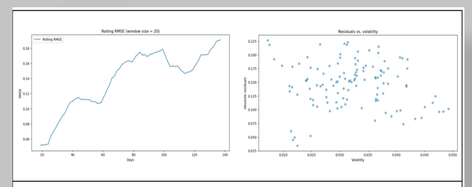

Figure 3: rolling RMSE (left) and residuals-volatility scatter plot (right)

Future Directions

1. Expand predictions to broader indices like the S&P 500 and ETFs.
2. Use advanced embeddings (e.g., transformers) for sentiment analysis.
3. Apply NLP techniques to improve sarcasm and context detection.
4. Enhance feature selection with dimensionality reduction (e.g., PCA).

Figure 4 compares the LSTM model with benchmarks:

Key Observation:

- The LSTM model closely
- tracks actual stock prices, outperforming Naive and MA forecasts.

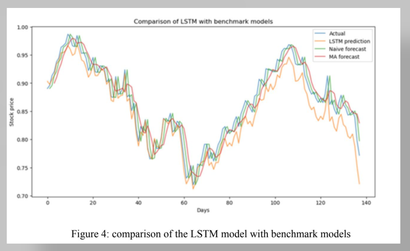

Figure 4: comparison of the LSTM model with benchmark models