Project Report: Analyzing the Impact of News Sentiment on Apple (AAPL) Stock Price

Introduction

This project investigated the relationship between financial news sentiment and Apple's (AAPL) stock performance. Leveraging historical stock prices and news articles, we applied a lexicon-based sentiment analysis to quantify daily news sentiment. This sentiment data was then integrated with stock returns and technical indicators. Through correlation and linear regression analysis, we aimed to determine if news sentiment significantly impacts AAPL's daily stock price movements.

Project Phases and Activities

The project followed a structured data science workflow, encompassing several key phases:

1. Data Acquisition: We sourced historical daily stock price data for Apple (AAPL) and a dataset containing financial news articles related to AAPL. These formed the foundational inputs for our analysis.

2. Data Preprocessing & Feature Engineering:

a) For Stock Data: Raw stock prices were cleaned by converting date formats, and new features like daily log returns and simple moving averages (SMA_10, SMA_20) were calculated to capture price momentum.

b) For News Data: News article timestamps were parsed and converted to a standardized date format. Crucially, raw news content was prepared for sentiment analysis by handling missing values and ensuring consistent column naming.

c) Sentiment Feature Engineering: A lexicon-based sentiment analysis was performed on each news article's content. These individual article sentiment scores were then aggregated to daily average sentiment (`Daily_Avg_Lexicon_Sentiment`) and daily article counts (`Num_Articles_That_Day`), serving as key features for our model.

3. Data Integration: The pre processed historical stock price data and the aggregated daily news sentiment data were merged into a single, comprehensive dataset. This allowed for joint analysis of market movements and news sentiment on corresponding dates.

4. Analysis & Modelling:

a) Correlation Analysis: We explored the linear relationships between various financial metrics (like stock returns and prices) and the newly engineered sentiment features (`Daily_Avg_Lexicon_Sentiment`, `Num_Articles_That_Day`).

b) Linear Regression Modelling: A statistical model was built to predict daily stock returns (`Log_Return`). This model incorporated both sentiment metrics and technical indicators to assess their individual and combined predictive power.

5. Evaluation & Interpretation: The developed model was evaluated using statistical measures (p-values, R-squared) to determine the significance and strength of relationships. The findings were then interpreted to draw conclusions regarding the impact of news sentiment on stock prices, which are presented in detail in the following sections.

Relevance of This Project in Today's World

- Information Overload: In an era of instant news and massive data generation, systematically processing and deriving insights from vast amounts of unstructured text data like news articles is critical for decision-making. This project demonstrates a method for doing so.
- Rise of Algorithmic and Quantitative Trading: A growing portion of market activity is driven by algorithms. Projects like this contribute to understanding how alternative data (like sentiment) can be incorporated into these algorithms for competitive advantage, even if direct linear links aren't always found.
- Behavioral Finance: While traditional finance assumes rational markets, behavioral finance acknowledges that human emotions and biases can influence investor decisions and market movements. News sentiment is a proxy for collective market emotion, making this study relevant to understanding these behavioral aspects.
- Data-Driven Decision Making: The project underscores the importance of empirical, data-driven analysis in finance, moving beyond intuition to build models and test hypotheses rigorously.
- Market Efficiency Debate: It contributes to the ongoing academic and practical debate about market efficiency – specifically, whether all public information, including subjective news sentiment, is immediately and fully reflected in stock prices. Our findings suggest that, for this specific method, it's not linearly priced in.

In essence, this project navigates the complex intersection of big data, natural language processing, and financial markets, aiming to uncover hidden patterns that could potentially offer a deeper understanding of stock price dynamics in the modern world.

Results

1. **Relationship Between AAPL Stock Price and News Article Sentiment**
   Based on the correlation matrix derived from our analysis:

- Sentiment vs. Stock Returns: The correlation coefficient between `Daily_Avg_Lexicon_Sentiment` and `Log_Return` was 0.063. Similarly, the correlation between `Num_Articles_That_Day` (number of articles) and `Log_Return` was 0.038. These values are very close to zero.
- Sentiment vs. Absolute Stock Price: The correlation between `Daily_Avg_Lexicon_Sentiment` and the `Close` price was 0.016, and for `Num_Articles_That_Day` it was -0.009.
- These extremely low correlation coefficients indicate that there is virtually no discernible linear relationship between the average daily sentiment of news articles (or the volume of articles) and either the daily changes (returns) or the absolute price levels of Apple's stock during the analyzed period.

2. **Affect of News Articles on the Movement of Stock Price**

- To assess if news articles (through their sentiment) affect stock price movement, we analyzed the results of a linear regression model where `Log_Return` (representing stock price

movement) was the dependent variable, and `Daily_Avg_Lexicon_Sentiment`, `SMA_10`, and `SMA_20` were the independent variables.

- Statistical Significance of Sentiment: The p-value (`Pr(>|t|)`) for `Daily_Avg_Lexicon_Sentiment` in the regression model was 0.518. Since this p-value is significantly higher than the conventional threshold of 0.05, we conclude that `Daily_Avg_Lexicon_Sentiment` is not a statistically significant linear predictor of daily stock returns for AAPL. In simpler terms, based on this model, the daily average sentiment of news articles does not reliably explain or predict changes in Apple's stock price.

- Overall Model Context: While the `Daily_Avg_Lexicon_Sentiment` itself was not significant, the overall model's Adjusted R-squared was 0.8097, indicating that a large portion of the variance in `Log_Return` is explained by the model. However, this strong explanatory power is predominantly attributed to the technical indicators (`SMA_10` and `SMA_20`), both of which showed high statistical significance (very low p-values). This suggests that the stock's own past price movements (represented by Simple Moving Averages) are far more influential in linearly predicting its daily returns than the lexicon-based news sentiment in this analysis.

- In summary, based on our linear analysis, news articles, specifically their aggregated daily average sentiment, do not appear to significantly affect the movement of Apple's stock price in a direct linear fashion within the scope of this project.

Performance of Sentiment Analysis

Sentiment analysis in this project was performed using a lexicon-based approach, specifically inspired by the Loughran-McDonald financial sentiment lexicon. The process involved the following steps:

- Lexicon Definition: A pre-defined list of positive and negative words (representative of the Loughran-McDonald lexicon, tailored for financial contexts) was used. Each word in the lexicon was assigned a polarity (positive or negative).

- Text Pre-processing: For each news article's content, the text was converted to lowercase, and then tokenized (broken down into individual words). Punctuation was removed to ensure accurate word matching.

- Sentiment Scoring per Article: For every news article, the script iterated through its words. Each word was checked against the defined positive and negative word lists.

a) A count of positive words was accumulated for the article.
b) A count of negative words was accumulated for the article.
c) A sentiment score for each article was then calculated using a simple formula: (Number of Positive Words - Number of Negative Words) / (Total Number of Positive + Negative Words). This normalized score typically ranges from -1 (most negative) to +1 (most positive).

- Daily Aggregation: After individual article sentiment scores were calculated, these scores were aggregated to a daily level. For each day, the average (`Daily_Avg_Lexicon_Sentiment`) and sum (`Daily_Sum_Lexicon_Sentiment`) of all article sentiment scores published on that day were computed. The total number of articles (`Num_Articles_That_Day`) was also counted. This daily aggregated sentiment was then used for correlation and regression analysis with the stock price data.

This lexicon-based method provides a quantitative measure of sentiment by identifying and counting emotionally charged words, offering a straightforward way to gauge the overall tone of news coverage related to AAPL.

Further Scope of Expansion of This Project

- Advanced Sentiment Analysis: Utilize machine learning or deep learning models (e.g., Transformer-based models like BERT or FinBERT) for more nuanced sentiment extraction, which can capture context, sarcasm, and negations more effectively than simple lexicons. Explore event-driven sentiment analysis to focus on the impact of specific, high-magnitude news events rather than daily averages.

- Broader Data Sources: Incorporate sentiment from social media platforms (e.g., Twitter, Reddit financial forums) to capture retail investor sentiment. Analyze sentiment from earnings call transcripts, analyst reports, or regulatory filings (e.g., 10-K, 10-Q) which often contain specific financial language.

- Complex Modelling Techniques: Investigate non-linear relationships between sentiment and stock prices using machine learning models (e.g., Random Forests, Gradient Boosting) that can capture complex patterns. Explore time-lagged effects to see if sentiment from previous days impacts current stock movements. Implement Granger causality tests to determine if sentiment "causes" stock returns or vice-versa. Integrate macroeconomic factors (e.g., interest rates, inflation) into the models.

- Portfolio Management & Strategy: Develop and back test sentiment-driven trading strategies to evaluate their profitability. Expand the analysis to a portfolio of stocks across different sectors to generalize findings.

This Project and Its Potential Expansions Can Benefit Various Stakeholders

- Individual Investors: Can gain a better understanding of how news and information might influence stock prices, helping them make more informed (or at least more aware) investment decisions beyond just technical charts.

- Financial Analysts and Researchers: Provides a foundational empirical study for further academic research into market efficiency, behavioral finance, and the complex interplay between information, emotion, and asset prices.
- Algorithmic Trading Firms: While our current linear model showed limited direct impact, such projects are crucial starting points for developing sophisticated quantitative trading strategies that integrate alternative data like news sentiment.
- Data Scientists and Machine Learning Engineers: Demonstrates a real-world application of data collection, processing, natural language processing (NLP), and statistical Modelling in the financial domain.
- Financial News Providers: Can potentially refine their news delivery or categorization based on insights into what type of news content truly moves markets.