Investigating the Relationship Between Daily News Sentiment and Dow Jones Volume

By Ibsa Barisa

Introduction

This report presents an analysis of the potential relationship between the sentiment of the top news stories of the day and the trading volume of the Dow Jones Industrial Average (DJIA). The study seeks to answer the question of whether daily news sentiment has any influence on stock market behavior, specifically the volume of shares traded in the DJIA.

Data Collection

The analysis draws on two primary datasets:

1. The *RedditNews.csv* dataset, which contains top news headlines ranked by their popularity on Reddit. Each date in the dataset has 25 associated headlines.
2. The *DJIA_table.csv* dataset, which features daily data from the DJIA, including "Open", "High", "Low", "Close", "Volume", and "Adj Close". This data is downloaded directly from Yahoo Finance.

The datasets span the years 2008 to 2016 and were retrieved from Kaggle.com, originally compiled by Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1.

Data Preprocessing

Before the data analysis phase, preprocessing steps were undertaken to ensure the quality and usability of the data:

1. **DJIA_table.csv:**
    - Headers were promoted to clarify the data structure.
    - Column distribution was analyzed to check for normal spread and identify outliers.
    - Column quality checks were performed to identify errors and empty values.
    - Data types were converted to the appropriate formats.
    - The data was sorted in ascending order by date.
    - Decimal points were rounded to appropriate places.
    - Volume data was divided by one million for easier readability.
    - The day name was extracted from the date column for future analysis.
2. **RedditNews.csv:**
    - Headers were promoted for data structure clarity.
    - Data types were converted to appropriate formats.
    - Errors and blank rows were removed.
    - The news was sorted and indexed by date and importance.
    - Text was cleaned and trimmed for analysis.

Outliers were treated using the 1.5 IQR rule and the normalization of data was checked using histograms, ensuring a foundation for accurate statistical testing.

Exploratory Data Analysis

Summary statistics were generated, along with box and whisker plots to visualize data spread. Relationships between various variables were also identified using visualizations.

Data Analysis

The analysis began by establishing the following hypotheses:

- Null Hypothesis (H0): There is no relationship between the sentiment of the top news stories of the day and the volume of the Dow Jones.
- Alternative Hypothesis (H1): There is a relationship between the sentiment of the top news stories of the day and the volume of the Dow Jones.

Sentiment Analysis

A Python script was employed to analyze the headlines and categorize each as either positive, neutral, or negative.

The two datasets were then merged using a left join on dates, followed by a vlookup to input the sentiment of the day into the combined dataset.

Hypothesis Testing

A t-test was conducted to compare trading volumes on days when news sentiment was negative and non-negative (positive or neutral). The aim was to determine whether there was a statistically significant difference in trading volume mean between these two groups. The level of significance was set at $p = 0.05$. The resulting p-value was 0.04945.

Interpretation

The p-value (0.04945) is less than the set significance level (0.05), indicating statistical evidence against the null hypothesis. This suggests that there is indeed a relationship between the

sentiment of top news stories and DJIA trading volume. However, it's important to note that correlation does not imply causation. Although this evidence suggests a relationship, it does not confirm whether one variable causes changes in the other.

Further testing could reveal more about this relationship. Additional analyses that could be considered include correlation analysis, causality tests, time series analysis, factor analysis, segmentation analysis, and sentiment classification. Furthermore, extending the analysis to other markets, like the NASDAQ or the S&P 500, could provide valuable insights.

Communication

To make the analysis more accessible and actionable, a pivot table was created, grouping data by dates and calculating total and average volumes for both negative and non-negative days. Two slicers were added to this table for easier navigation: one for date and one for sentiment (negative/non-negative). Additionally, a visualization was created to make observations easier.

Conclusion

The data analysis suggests a statistically significant relationship between the sentiment of the top news stories of the day and the trading volume of the DJIA. However, further research and testing are required to uncover more details of this relationship, its causal nature, and the extent of its applicability across different markets.