Stock Sentiment Analysis: Predicting Stock Movements Using News Headlines

Author: Gursimran Sujlana

Abstract

This research investigates the effectiveness of using news headlines to predict stock price movements. By employing a RandomForestClassifier, the study aims to determine the sentiment of news articles and their impact on stock prices. The results demonstrate a significant correlation between the sentiment of news headlines and stock price movements, achieving an accuracy of 85.19%.

1. Introduction

The stock market is influenced by various factors, including news events. Sentiment analysis of news headlines can provide valuable insights into predicting stock price movements. This study explores the application of machine learning techniques, particularly RandomForestClassifier, to predict stock price movements based on the sentiment derived from news headlines.

2. Literature Review

Previous studies have shown the impact of news on stock prices. Traditional approaches include sentiment analysis using lexicon-based methods and machine learning techniques. Recent advancements in Natural Language Processing (NLP) and machine learning have enabled more accurate sentiment analysis, providing better predictive capabilities.

3. Methodology

3.1 Dataset

The dataset comprises news headlines and stock price labels. Each entry includes:

- Date: The date of the news.
- Label: Indicates the stock price movement (0 for low, 1 for high).
- Top1 to Top25: News headlines for the day.

Dataset Statistics:

- Total entries: 4101
- Features: 27 (Date, Label, Top1 to Top25)
- Label distribution:
  - High stock price (1): 2166
  - Low stock price (0): 1935

3.2 Data Preprocessing

The dataset was split into training and testing sets based on the date. Headlines were preprocessed to remove special characters and punctuations. CountVectorizer was used to convert the headlines into a matrix of token counts, specifically bigrams.

3.3 Model Training

A RandomForestClassifier was trained using the preprocessed headlines and corresponding labels from the training set. The model parameters were optimized to achieve the best performance.

3.4 Evaluation Metrics

The model's performance was evaluated using:

- Confusion Matrix
- Accuracy
- Precision
- Recall
- F1-Score

4. Results

4.1 Confusion Matrix

- True Positives (TP): 184
- True Negatives (TN): 138
- False Positives (FP): 48
- False Negatives (FN): 8

4.2 Accuracy

The model achieved an accuracy of 85.19%.

4.3 Classification Report

Metric	Class 0 (low stock price)	Class 1 (high stock price)	Accuracy	Macro Avg	Weighted Avg
Precision	0.95	0.79	-	0.87	0.87
Recall	0.74	0.96	-	0.85	0.85
F1-Score	0.83	0.87	-	0.85	0.85
Support (count)	186	192	378	378	378

Key Metrics:

- **Class 0 (low stock price)**:
  - Precision: 0.95
  - Recall: 0.74
  - F1-Score: 0.83
  - Support: 186
- **Class 1 (high stock price)**:
  - Precision: 0.79
  - Recall: 0.96
  - F1-Score: 0.87
  - Support: 192

5. Discussion

The model demonstrated high accuracy in predicting stock price movements based on news headlines, particularly in identifying high stock prices. The high precision for low stock prices indicates the model's effectiveness in minimizing false positives. However, the recall for low stock prices suggests potential improvements in identifying all low stock price events.

6. Conclusion

This study highlights the potential of using news headlines for stock sentiment analysis. The RandomForestClassifier provided robust predictions with an overall accuracy of 85.19%. Future work could explore more advanced NLP techniques and other machine learning models to further enhance predictive performance.

7. Future Work

Future research could involve:

- Incorporating additional features such as the volume of news articles, sentiment scores, and financial indicators.
- Experimenting with deep learning models, such as LSTM or BERT, for improved text processing and sentiment analysis.
- Conducting a comparative analysis with other machine learning algorithms.

References

- [1] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.
- [2] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139-1168.
- [3] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.