Leveraging Machine Learning for Financial Sentiment Analysis

James Calnan

Introduction and Research Question

This study explores the application of machine learning techniques for financial sentiment analysis, focusing on the impact of news sentiment on stock market trends. The primary research question is: "How can machine learning be utilised to accurately classify the sentiment of financial news and predict stock market trends?".

Sentiment analysis, a crucial aspect of natural language processing, leverages machine learning to interpret and classify subjective information in text data. In financial markets, where news sentiment can significantly influence stock trends, the ability to accurately analyse and predict these trends using machine learning presents a valuable tool for investors and analysts.

Machine Learning Model Selection and Justification

The study employed a strategic selection of machine learning models, each chosen for its unique strengths in the context of financial sentiment analysis. The models included:

- **Supervised Models** like *Logistic Regression* and *Random Forest Classifier*, renowned for their efficiency in binary classification, and *KNeighborsClassifier*, effective in small dataset scenarios.
- **Advanced Ensemble Models**, such as *GradientBoostingClassifier* and *XGBClassifier*, were crucial for managing complex patterns and imbalanced datasets.
- **Specialised Text Classification Models,** including *MultinomialNB* and *SGDClassifier*, were selected for their proficiency in handling large-scale text data.

This diverse range of models was chosen to balance computational efficiency and accuracy, specifically tailored to the challenges of interpreting nuanced financial news sentiment.

Data Preparation Justification and Challenges

Data preparation was an important step for this study. The initial step involved merging datasets from Kaggle's Financial Sentiment Analysis [2] and Sentiment Analysis for Financial News [3], creating a comprehensive base for analysis. Critical to the study's success was encoding sentiments into binary values, streamlining the classification task. Text data preprocessing, including lowercasing, punctuation, and number removal, was done to normalise the data. Balancing the dataset was considered but ultimately decided against due to the limitations in the volume of training data.

Application of Machine Learning Techniques

The study applied a variety of supervised machine learning classifiers to the dataset, focusing on model training and evaluation. Key performance metrics such as accuracy, precision, recall, and the F1 score were employed to assess each model's effectiveness, with detailed results in Table 1. Additionally, confusion matrices were utilised to gain insights into the predictive capabilities of each model. This approach facilitated a comprehensive comparison of model performances, highlighting strengths and weaknesses in financial sentiment analysis. The models were then saved as pipelines, enhancing their efficiency and scalability for future applications.

Application to Unseen Data

Models were tested on unseen data from Kaggle's Massive Stock News Analysis DB for NLPBacktests [1], integrating sentiment scores with Yahoo Finance's historical stock data. An **Ensemble Method** was also employed, selecting the most common sentiment across all models for each date, thus enhancing prediction accuracy in real-world scenarios.

Trading Simulation and Model Edge

The study incorporated a simulated trading strategy to evaluate the practical applicability of sentiment analysis in trading scenarios. This simulation used the models to interpret financial news sentiment and make corresponding trading decisions (buy or sell). The simulation aimed to mirror real-world trading conditions, testing the models across various stocks. The performance on Morgan Stanley stock using the models is detailed in Table 2, and the portfolio value throughout the simulation is shown in Figure 1.

Despite all models showing a trading edge above 50%, Merck & Co's portfolio performance was less than optimal, as seen in Figure 1 and Table 2. This could be attributed to suboptimal position sizing and timing of entry and exit in trades.

The trading edge metric was used to measure the models' predictive accuracy against random guessing. An edge above 50% across all models indicated their effectiveness in performing better than guessing, demonstrating their potential to provide actionable trading insights. This not only underscores the models' utility in automated trading systems but also affirms the relevance of machine learning in financial sentiment analysis.

Reflection

Model Selection: Balancing Advantages, Disadvantages, and Trade-offs

The study's model selection balanced the need for accuracy, simplicity, and real-world applicability in the context of financial sentiment analysis. Simpler models like **LogisticRegression** and **KNeighborsClassifier** offered speed and interpretability but showed limitations in capturing nuanced market trends, as evidenced in the Morgan Stanley stock trading simulation. While achieving higher accuracy, more sophisticated models such as **GradientBoostingClassifier** and **XGBClassifier** required careful calibration to manage their complexity, as reflected in their trading performance metrics. Models like **DecisionTreeClassifier** and **RandomForestClassifier** struck a balance, demonstrating strong ROI but facing challenges with complex financial data. The study's approach to model selection was thus characterised by navigating these trade-offs, aiming to optimise the blend of computational efficiency and accuracy in analysing financial sentiments.

Acknowledgement of Research Limitations

Several limitations need to be discussed. For example, the data may not fully capture the breadth of financial sentiments, potentially leading to biases. Additionally, the models' performance could vary with different datasets, and the generalizability of the findings might be limited. Data availability from the Massive Stock News Analysis DB for NLPBacktests [1] was limited; for the stock Merck & Co, there was only availability for 53% of trading days, suggesting limitations in the trading simulation and model edge calculations. Future research could address these limitations by incorporating a wider range of data sources and exploring more advanced machine learning techniques that can better handle the complexities and nuances of financial news sentiment analysis.

Conclusion

This research underscores the potential of machine learning in financial sentiment analysis, showing how these techniques can offer valuable insights into market trends. However, it also highlights the limitations and the need for ongoing improvements in model sophistication and data collection methods.

Model	F1 Score	Recall	Precision	Accuracy
RandomForestClassifier	0.943	0.981	0.908	0.919
SGDClassifier	0.935	0.941	0.929	0.910
LogisticRegression	0.927	0.963	0.893	0.895
KNeighborsClassifier	0.923	0.969	0.881	0.889
XGBClassifier	0.912	0.969	0.862	0.872
MultinomialNB	0.895	0.885	0.905	0.857
GradientBoostingClassifier	0.890	0.991	0.808	0.831
DecisionTreeClassifier	0.873	0.981	0.786	0.803

Table 1: Model Performance Metrics

Model	Price vs Portfolio (MS)	Edge (MS)	Price vs Portfolio (MRK)	Edge (MRK)
finbert	70.20%	61.12%	-105.94%	56.89%
DecisionTreeClassifier	49.63%	56.57%	-83.95%	54.08%
ensemble	43.92%	55.73%	-79.67%	54.15%
RandomForestClassifier	43.90%	55.58%	-90.31%	54.97%
KNeighborsClassifier	43.60%	51.48%	-124.82%	50.79%
LogisticRegression	41.91%	55.35%	-116.49%	52.23%
GradientBoostingClassifier	39.86%	56.19%	-74.21%	54.49%
SGDClassifier	5.47%	54.37%	-110.51%	50.99%
MultinomialNB	-3.81%	53.83%	-128.85%	50.72%
XGBClassifier	-8.65%	56.57%	-91.29%	54.76%
Random	-21.76%	48.83%	-109.14%	49.51%

Table 2: Consolidated Performance Metrics for Morgan Stanley and Merck & Co Inc Stocks

References

[1] Daily financial news for 6000+ stocks. https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests. Accessed: [20/11/2023].

[2] Financial sentiment analysis dataset. https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis. Accessed: [20/11/2023].

[3] Sentiment analysis for financial news dataset. https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news. Accessed: [20/11/2023].

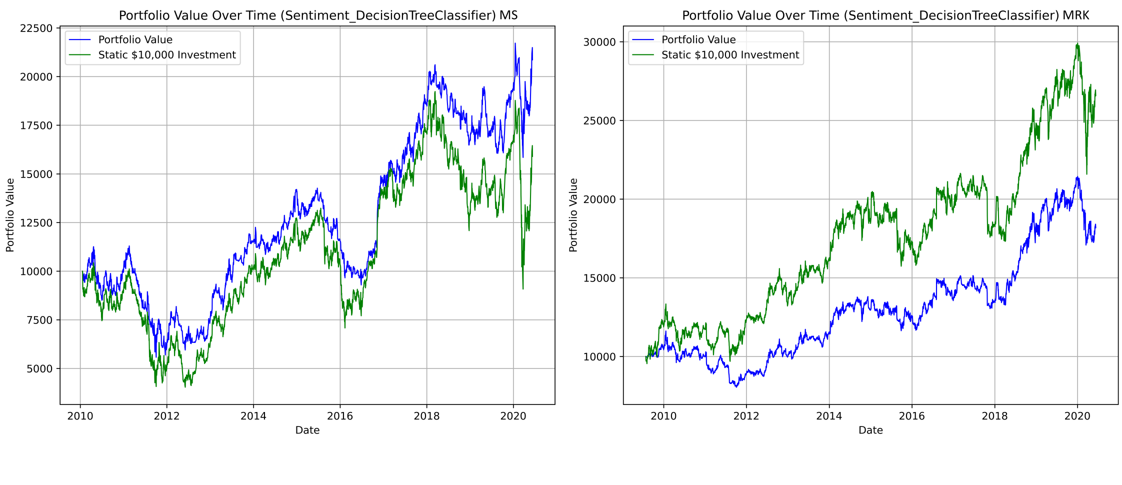

Figure 1: Trading performance graphs for MS and MRK