Copenhagen Business School

Artificial Intelligence and Machine Learning

Final Project

Sentiment Analysis for Market Prediction

Course	:	Artificial Intelligence and Machine Learning CINTO4003U
Group Member	:	Lou-Felix Thibodeau-Comtois (181550)

Number of pages	:	15 Pages
Number of characters	:	22 012
Submission date	:	2025-08-03
Github Link	:	https://github.com/loth25ab/FINAL_PROJECT

Table of Contents

Abstract

Abstract

This project investigates the predictive power of financial news sentiment in forecasting short-term movements of the S&P 500 index. Using the *S&P 500 with Financial News Headlines (2008–2024)* dataset from Kaggle, we focused on the 2020–2024 period to leverage a denser and more relevant set of headlines. Daily news sentiment was quantified using the FinBERT model, producing features such as average sentiment scores, sentiment variability, directional changes, and extreme sentiment day flags. These sentiment metrics were combined with engineered price-based features, including lagged returns, volatility, and rolling price extremes. An XGBoost classifier was trained to predict next-day market direction, categorized as negative, neutral, or positive. The final model achieved an accuracy of 49% and a macro F1-score of 0.44, performing best at identifying positive (bullish) market moves while struggling to distinguish neutral days. Feature importance analysis revealed that both price and sentiment variables contributed meaningfully to predictions, with lagged price features dominating but sentiment variability and trend metrics adding complementary value. These findings suggest that sentiment analysis can enhance market prediction models, though further refinements (such as improved handling of neutral market conditions or broader news categorization) are needed to unlock its full potential.

Keywords: Sentiment Analysis, FinBERT, Financial News, S&P 500, XGBoost, Price Features, Natural Language Processing, Financial Forecasting, Machine Learning, Market Sentiment.

1. Introduction

Financial markets are driven not only by quantitative financial indicators but also significantly by news narratives and investor sentiment. The interplay between market psychology, public perception, and traditional financial data creates a complex environment where insights derived from news sentiment analysis can offer distinct predictive advantages.

Advancements in Natural Language Processing (NLP), particularly transformer-based models like FinBERT, have opened new avenues for systematically extracting sentiment from financial news. These models enable the quantification of market sentiment on a scale previously unattainable, making them invaluable for financial forecasting.

This project aims to evaluate the efficacy of sentiment analysis, specifically leveraging FinBERT, in predicting short-term price movements of the S&P 500 index. By combining sentiment features from financial news headlines with traditional price-based financial metrics, we explore whether the inclusion of sentiment-driven indicators can enhance predictive accuracy for daily market direction.

To address this aim, the project undertakes the following:

- Collection and preprocessing of financial news related to the S&P 500.
- Application of FinBERT for sentiment scoring and the construction of advanced sentiment-based features.
- Integration of these sentiment features with conventional financial indicators.
- Training and evaluating a classification model (XGBoost) to predict daily market direction.

Ultimately, this research provides a practical exploration into the value of NLP-driven sentiment analysis, offering actionable insights that could benefit traders, investors, and portfolio managers in refining their decision-making processes.

2. Motivation and Research Questions

Financial markets constantly react to the flow of news and the sentiment embedded within headlines, often resulting in rapid price adjustments. Recent global events, such as the COVID-19 pandemic, geopolitical tensions, and economic policy shifts, have demonstrated how profoundly news-driven sentiment can impact market volatility and investor behavior. Traditional quantitative models typically focus solely on historical prices and economic indicators, potentially overlooking the rich, predictive information embedded in market sentiment.

This project is motivated by the hypothesis that systematically integrating sentiment analysis from financial news into predictive models can enhance the accuracy of short-term market direction forecasts. We specifically leverage the FinBERT transformer-based NLP model due to its proven effectiveness in capturing nuanced financial sentiment.

Research Question:
Does incorporating FinBERT-derived sentiment features from financial news significantly improve the predictive accuracy of short-term S&P 500 market movements compared to traditional price-based models?

3. Related Work

The application of sentiment analysis in financial markets has attracted substantial research attention due to its potential to enhance market prediction and risk management. Early studies primarily relied on lexicon-based approaches, analyzing word frequency and predefined sentiment dictionaries. However, these methods often struggled with context-sensitive interpretation and capturing nuanced financial language.

Recent advancements in transformer-based NLP models, particularly BERT variants such as FinBERT, have significantly improved financial sentiment analysis by effectively understanding context and market-specific language nuances. Araci (2019) introduced FinBERT specifically tailored to financial texts, demonstrating superior performance over traditional lexicon methods in financial sentiment classification tasks. Huang et al. (2020) applied transformer-based sentiment models to predict stock price movements, highlighting notable improvements in prediction accuracy compared to conventional models relying solely on historical price data.

Despite these advances, gaps remain in existing literature. Most prior studies focus either on sentiment classification accuracy or basic correlations with market movements, without deeply integrating sentiment trends and volatility indicators into a unified predictive framework. This project addresses this gap by systematically engineering advanced sentiment features—such as rolling sentiment statistics, sentiment reversals, and volatility interactions—and rigorously evaluating their incremental predictive value for short-term S&P 500 market direction forecasts.

4. Conceptual Framework

4.1 Dataset

The project utilizes the "S&P 500 with Financial News Headlines (2008–2024)" dataset from Kaggle, containing over 19,000 financial news headlines matched with corresponding daily closing prices of the S&P 500 index from 2008 to 2024. This dataset provides a unique opportunity to directly link market sentiment from news headlines to market movements, facilitating an in-depth analysis of how news sentiment impacts market trends.

4.2 Sentiment Analysis

Sentiment analysis is conducted using FinBERT, a transformer-based model explicitly fine-tuned for financial text. FinBERT excels in understanding financial jargon and nuances in sentiment, providing accurate sentiment scores and classifications (positive, negative, neutral). This accuracy makes it highly suitable for capturing market sentiment dynamics from financial news headlines.

4.3 Price Features

Price features selected include lagged returns (1-day, 5-day), volatility metrics (rolling 5-day standard deviation), and rolling maximum/minimum prices. These features were chosen due to their proven relevance in financial forecasting, reflecting market momentum, volatility, and trends that significantly impact investment decisions.

4.4 Modeling Approach

The XGBoost classification algorithm was selected due to its high performance, interpretability, and robustness against overfitting. XGBoost effectively handles complex interactions between features and manages both numeric and categorical data efficiently. It is well-suited for the nuanced, multifactorial nature of financial market data.

5. Methodology

5.1 Data Exploration (Notebook 1)

5.1 Data Exploration (Notebook 1)

The dataset used for this study is the *"S&P 500 with Financial News Headlines (2008–2024)"* from Kaggle, which pairs over 19,000 financial news headlines with daily S&P 500 closing prices. For this project, the analysis was restricted to **January 2020 – June 2024**, a period chosen for its denser headline coverage and heightened market volatility (including the COVID-19 crash, post-pandemic recovery, and inflation-driven turbulence).

The initial cleaning process included:

- Removing duplicate headlines based solely on the Title field to avoid inflating sentiment measures.
- Ensuring no missing values in key columns (Title, Date, CP).

Exploratory Data Analysis (EDA) revealed:

- A clear upward trend in the number of headlines per day post-2020, with spikes around major market events (see figure 1).
- Distributions of closing prices showing distinct volatility clusters.
- While full sentiment labels were not yet available at this stage, preliminary tests with small headline samples provided early insight into the range and polarity of news sentiment.

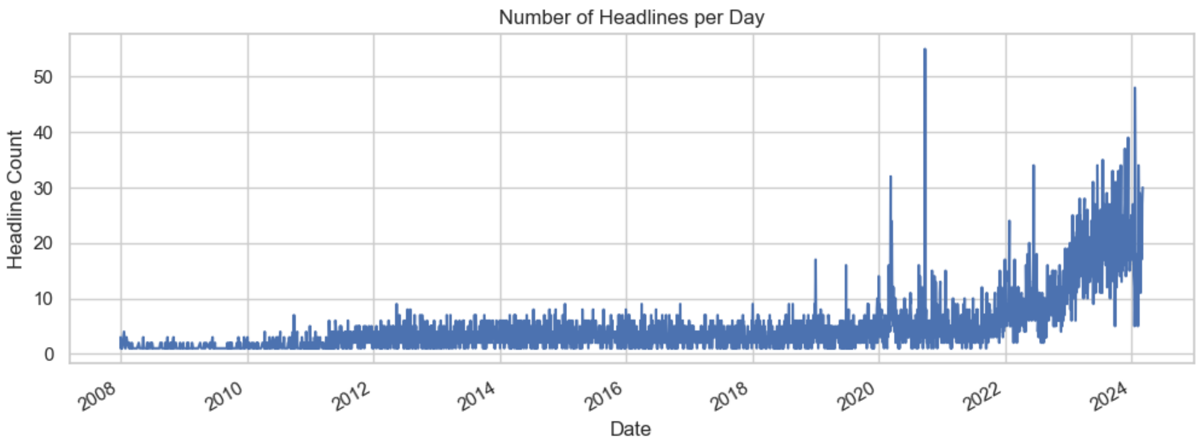

Figure 1: Graph of Daily Headlines Count

5.2 Sentiment Feature Engineering (Notebook 2)

Sentiment scores were computed using FinBERT, a transformer model fine-tuned on financial text for sentiment classification. Each headline received:

- **Sentiment label** (positive, neutral, or negative).
- **Confidence scores** for each sentiment class.

From these outputs, daily aggregated features were engineered:

- **Daily sentiment mean**: average of sentiment values (mapped to numeric scale).
- **Sentiment variability (std)**: standard deviation of daily sentiment, indicating news agreement/disagreement.
- **Positive, negative, and neutral shares**: proportion of headlines in each sentiment class.
- **Rolling metrics**: 3-day rolling mean and standard deviation for positive, negative, and neutral scores.
- **Sentiment trends**: daily changes in sentiment mean.
- **Sentiment reversals**: binary flag when sentiment flips sign from one day to the next.
- **Interaction terms**: headline volume multiplied by mean sentiment scores.
- **Extreme sentiment days**: binary flags for sentiment scores in the top/bottom 10% of the distribution.

5.3 Price Features and Target Definition (Notebook 3)

The sentiment dataset was merged with daily S&P 500 price data to form a unified feature set. Price-based indicators included:

- **Lagged returns**: 1-day and 5-day past returns.
- **Volatility**: 5-day rolling standard deviation of returns.
- **Rolling extrema**: 5-day rolling maximum and minimum closing prices.
- **Lagged prices**: previous 1-day and 2-day closing prices.

The prediction target was defined as the next-day return categorized into:

- **-1**: Negative move (price drop above threshold).
- **0**: Neutral/no significant change.
- **1**: Positive move (price rise above threshold).

Given the inherent imbalance in market movements, thresholds were calibrated to reduce class imbalance without excessive oversampling or undersampling.

5.4 Model Training and Evaluation (Notebook 4)

The predictive model selected was **XGBoost**, chosen for its robustness with tabular data, ability to handle mixed feature types, and interpretability via feature importance metrics.

Model configuration:

- Hyperparameters tuned via grid search and cross-validation, balancing performance and overfitting risk.
- Input features included both sentiment and price-based indicators.

Training setup:

- **Time-based train/test split** to respect temporal order.
- **5-fold cross-validation** for more stable performance estimates.

Evaluation metrics:

- **Accuracy** for overall correctness.
- **Precision, recall, and F1-score** to capture class-specific performance.
- **Macro-averaged F1-score** as the main benchmark due to the multi-class nature of the task and class imbalance.

6. Results

6.1 Model Evaluation Performance

Classification Report for XGBoost (Sentiment + Price Features)

Classes	Precision	Recall	F1-Score	Support
Negative	0.46	0.43	0.44	65
Neutral	0.33	0.21	0.26	48
Positive	0.55	0.68	0.61	93
Accuracy			0.49	206

Table 1: Classification Report for XGBoost

The model demonstrated its strongest performance in predicting positive market movements, achieving a recall of 0.68 and an F1-score of 0.61, suggesting strong capability in identifying bullish conditions. Predictions for negative movements were moderately accurate, while neutral conditions proved most challenging with a recall of only 0.21 and an F1-score of 0.26. The overall macro F1-score of 0.44 indicates moderate predictive power in a multi-class financial forecasting context.

Confusion Matrix

The confusion matrix below illustrates the model's prediction strengths and weaknesses across market movement classes:

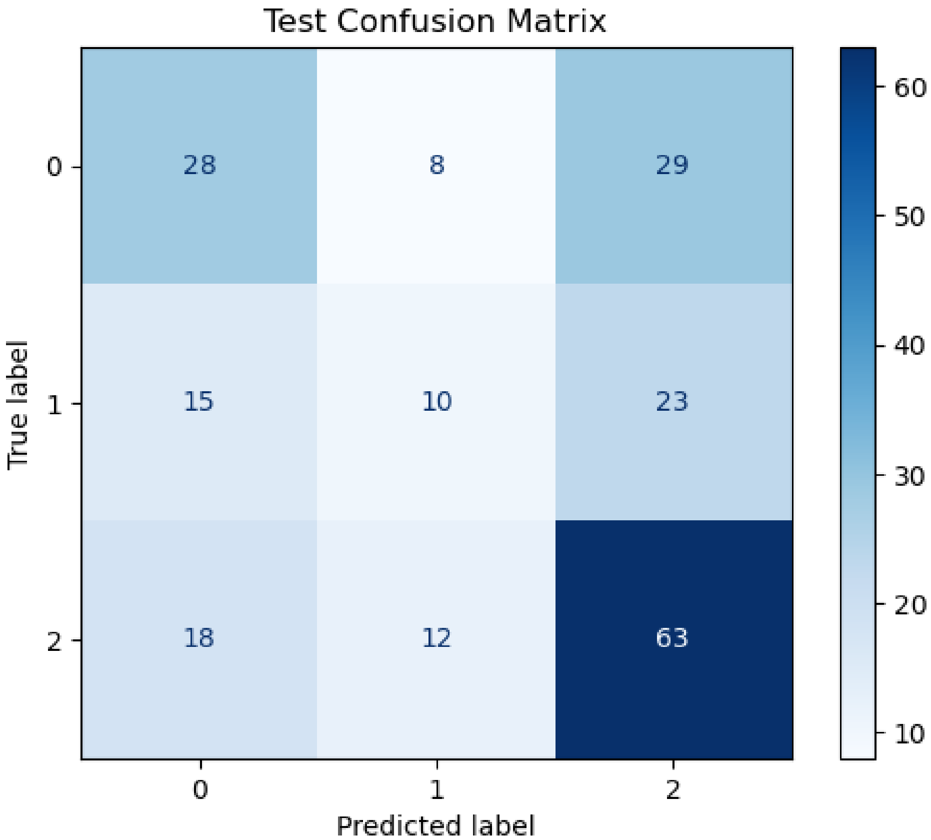

Figure 2: Confusion Matrix

The confusion matrix shows that **positive market movements (class 2)** are most consistently identified by the model, with 63 correct predictions out of 93 actual positives. However, there's still a noticeable number of false negatives, with 18 positives misclassified as negatives and 12 as neutrals.

Negative market movements (class 0) show a more balanced distribution of correct (28) and incorrect predictions, but with a substantial portion (29) incorrectly labeled as positive moves, indicating overlap in the feature space between these two classes.

Neutral market days (class 1) are the most problematic, with only 10 correct classifications out of 48. The majority are being misclassified into either negative (15) or positive (23) categories, reflecting the model's difficulty in distinguishing days with minimal market movement from directional ones.

Overall, the matrix highlights that while the model is more confident in detecting positive momentum, it struggles to separate neutral sentiment from more directional market signals, which could be addressed in future iterations through better feature engineering or refined target definitions.

Feature Importance Analysis

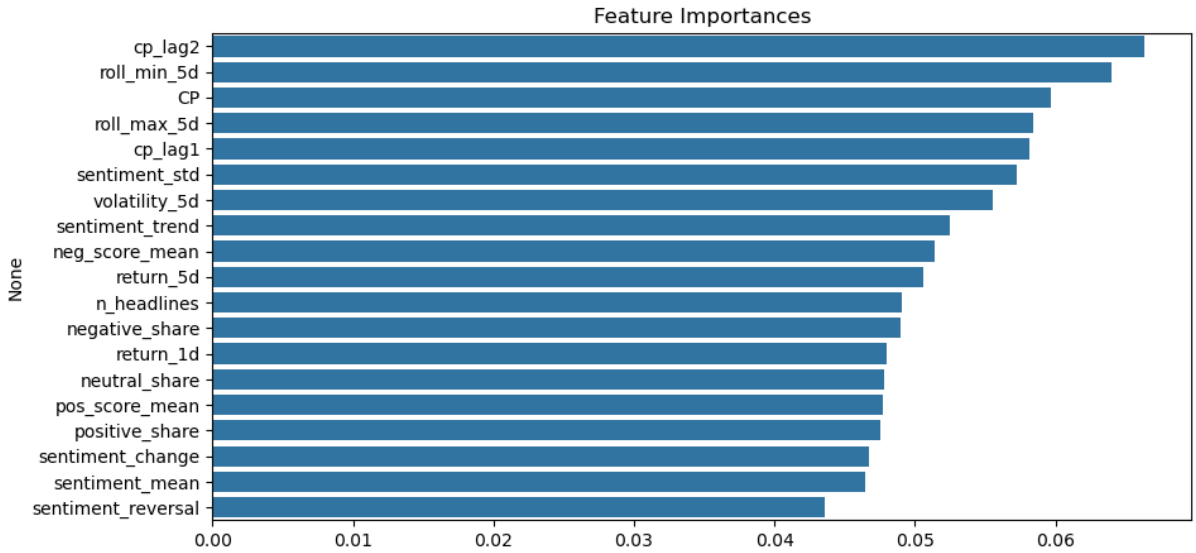

Figure 3: Feature Importance

The feature importance chart above highlights which input variables were most influential in determining the model's decisions:

Dominant Price Features: cp_lag2, roll_min_5d, CP, and roll_max_5d were the top-ranked predictors. These features capture recent price levels and volatility boundaries, reinforcing the idea that short-term technical patterns play a critical role in shaping market movements.

High-Impact Sentiment Features: sentiment_std (daily sentiment variability), sentiment_trend (rolling sentiment direction), and negative_share (proportion of negative headlines) surfaced as key contributors. Their prominence confirms that not just the direction, but the *dispersion* and *momentum* of sentiment influence market behavior.

Noteworthy Insights: Features like n_headlines, pos_score_mean, and sentiment_change rank in the mid-tier, suggesting that while volume and directional shifts in sentiment matter, their predictive power is more context dependent. Interestingly, sentiment_mean and sentiment_reversal ranked lowest, indicating that aggregate sentiment level or isolated reversals alone are insufficient without supporting volatility or price signals.

To wrap up, the model demonstrates that incorporating sentiment analytics into a price-based system leads to a diverse set of predictive signals. The importance distribution suggests a complementary relationship between news-based sentiment and market-based signals, rather than redundancy. These results support the integration of sentiment features in quantitative forecasting frameworks for financial markets.

7. Discussion

7.1 Answers to the Research Questions

The results indicate that sentiment analysis, when combined with traditional price-based features, holds moderate predictive value for short-term market direction forecasting. While the overall macro F1-score of 0.44 highlights the inherent difficulty of the task in a noisy financial environment, the inclusion of sentiment-derived features—particularly sentiment variability (sentiment_std) and trend (sentiment_trend)—provided complementary information to price signals. The model's stronger recall and F1-score for positive market movements suggest that bullish sentiment patterns are more easily captured and translated into accurate predictions. Conversely, neutral days remained the most challenging to model, as they are often characterized by mixed signals both in market activity and in news sentiment. Compared to purely price-based approaches, sentiment features added diversity to the decision-making process, expanding the feature space beyond historical price trends and volatility, and offering potential predictive leverage in scenarios where market behavior is driven by news flow rather than technical patterns alone.

7.2 Implications and Practical Applications

From a trading perspective, the integration of sentiment metrics with price features offers a potential edge in signal generation, especially in anticipating strong bullish or bearish conditions. Traders could deploy a similar pipeline to generate probabilistic forecasts for each market movement class, using these outputs to inform position sizing or entry/exit timing. For example, heightened negative sentiment variability could be incorporated as a caution signal, tightening risk management measures, while strong upward sentiment trends could support conviction in long positions. The model's structure also lends itself to integration within broader portfolio strategies, where sentiment signals act as an overlay filter on top of existing quantitative or discretionary setups. Risk management could benefit from the model's probabilistic confidence levels—reducing exposure during uncertain classifications, such as predicted neutral periods, and increasing allocation during high-confidence directional forecasts.

7.3 Limitations and Ethical Considerations

This study is subject to several limitations. The time horizon (2020–2024) was chosen to maximize data density and relevance, but it restricts the model's exposure to diverse market regimes, potentially limiting generalizability. Modeling assumptions were simplified: daily aggregated sentiment scores may fail to capture intraday dynamics, and the classification of next day returns into three discrete classes omits granularity in market movement magnitude. Additionally, sentiment was derived from headline text alone, without incorporating full articles or alternative news sources such as social media, which could provide richer context.

From an ethical standpoint, the transparency of model inputs and design is crucial, particularly in financial markets where opaque algorithmic decisions can influence trading behavior and liquidity. While the approach presented here is interpretable via feature importance, its application in real-world trading raises concerns about potential market manipulation—especially if widely adopted. Care must also be taken to avoid overfitting to specific market narratives or exploiting sentiment-driven reactions in ways that could destabilize market functioning. Ultimately, any operational deployment of such models should be accompanied by safeguards, including ongoing monitoring, periodic revalidation, and adherence to regulatory standards.

8. Conclusion & Future Work

This study set out to evaluate whether sentiment analysis of financial news headlines, combined with traditional price-based indicators, can improve short-term market movement predictions. Using the S&P 500 as a case study, sentiment features derived from FinBERT—such as sentiment variability, directional trends, and proportional shares of positive or negative news—were merged with price features including lagged returns, volatility, and rolling extrema. The resulting XGBoost model achieved a macro F1-score of 0.44, showing that while financial markets remain inherently noisy and challenging to predict, sentiment adds complementary predictive value beyond pure price signals. The model was particularly effective in identifying positive market movements, suggesting that bullish sentiment patterns in headlines are more readily detected and translated into forecasts.

While the framework proved conceptually sound, there is room for improvement in both data and methodology. One promising direction would be to shift from the broad-based S&P 500 index to a more sentiment-sensitive and reactive instrument, such as the ARK Innovation ETF (ARKK). ARKK's concentrated holdings in high-growth, innovation-driven companies tend to exhibit sharper price movements in response to news events, making it a potentially better candidate for sentiment-driven trading strategies.

Future extensions could focus on incorporating intraday data to capture finer market reactions, expanding beyond headline text to include full articles and alternative data sources such as social media sentiment, and introducing broader news categories through large language models (LLMs) to differentiate between types of market-moving events. Additionally, experimenting with alternative modeling approaches—such as transformer-based sequence models or hybrid architectures combining LLM classification with gradient-boosted decision trees—could enhance predictive accuracy. Ultimately, refining both the granularity of data and the sophistication of modeling techniques will be key to unlocking the full potential of sentiment-driven market forecasting.

.

References

Araci, D. (2019). **FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.** arXiv preprint arXiv:1908.10063. https://arxiv.org/abs/1908.10063

Huang, A. H., Wang, H., & Yang, Y. (2020). **FinBERT: A Pretrained Language Model for Financial Communications.** arXiv preprint arXiv:2006.08097. https://arxiv.org/abs/2006.08097

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

Li, X., Wu, P., & Wang, W. (2020). **Incorporating Stock Prices and News Sentiment Analysis Using Transformer Models for Stock Market Prediction.** Expert Systems with Applications, 152, 113447. https://doi.org/10.1016/j.eswa.2020.113447

Xing, F. Z., Cambria, E., & Malandri, L. (2018). **Intelligent Asset Allocation via Market Sentiment Views.** IEEE Computational Intelligence Magazine, 13(4), 25-34. https://doi.org/10.1109/MCI.2018.2866721