Stock Sentiment Analysis

Kenneth Egan
Computer Science, Wentworth Institute of Technology

***Abstract*—This project aims to investigate the relationship between financial news sentiment and market behavior by applying natural language processing (NLP) techniques to analyze headlines and extract sentiment classifications. The main goal is to determine whether sentiment derived from financial news can predict short-term market movements, influence specific industries more than others, and behave differently under varying market conditions such as bull or bear cycles. The project leverages machine learning techniques, including Logistic Regression and FinBERT-based sentiment classification, to identify patterns and correlations. This work holds relevance for investors, analysts, and data scientists seeking to understand the predictive power of news on financial markets.**

***Keywords*—financial sentiment analysis, stock market prediction, FinBERT, market volatility, machine learning**

I. Introduction

In the fast-paced world of finance, the interpretation and reaction to news plays a crucial role in shaping market dynamics. With the exponential rise of algorithmic trading and real-time news feeds, the ability to quantify and interpret sentiment from financial news has become an area of growing interest among investors, analysts, and researchers. This project explores sentiment analysis on financial news headlines to evaluate whether sentiment influences market behavior at various levels—from overall indices like the S&P 500 to specific industries and individual stocks.

Understanding whether positive (bullish) or negative (bearish) sentiment correlates with short-term price changes can offer a significant edge in financial decision-making. Previous studies and tools, such as FinBERT—an NLP model fine-tuned on financial data—have provided early success in identifying sentiment from unstructured textual data. However, translating this sentiment into meaningful market insight remains a challenging endeavor, often limited by data granularity, noise in financial markets, and the complexity of investor behavior.

This project aims to tackle four central questions in the domain:

1. Can financial news sentiment predict short-term market reactions?
2. What is the general distribution of sentiment in financial headlines?
3. Are specific sectors or industries more sensitive to sentiment-driven news?
4. Does the impact of sentiment vary between bull and bear market conditions?

To answer these, we process over 2 million financial headlines using FinBERT, engineer temporal and categorical features, and analyze the results using a combination of correlation analysis, logistic regression, and clustering techniques to distinguish market cycles. By combining text-based machine learning with financial data, this study contributes to ongoing research in financial forecasting, sentiment-driven investing, and NLP-based trading strategies.

II. Datasets

A. Source of dataset

The dataset used in this project is the **Financial News and Stock Price Integration Dataset (FNSPID)**, obtained from the Hugging Face Datasets Hub, a credible and widely used platform for machine learning datasets. It was created by **Zihan Dong, Xinyu Fan, and Zhiyuan Peng** and released in **early 2024**.

The dataset covers **4,775 companies listed on the S&P 500**, with financial news articles aligned to corresponding stock price data over time. It was built by aggregating and processing data from four major financial news platforms, including Nasdaq. Each news item is time-stamped and linked to market data, enabling time-series analysis of sentiment and market reactions.

All data collection and preprocessing were performed by the original authors, and the full methodology is documented in their GitHub repository. The dataset is available under the **Creative Commons Attribution-Noncommercial 4.0 International (CC BY-NC 4.0)** license and is intended for academic and non-commercial use.

B. Character of the datasets

The dataset used in this project is a filtered and enhanced version of the publicly available FNSPID dataset. It consists of financial news headlines associated with S&P 500 companies from **2011 to 2020**. To make the dataset suitable for sentiment analysis, I processed approximately **2,000,000 headlines** using the **FinBERT** model to generate sentiment labels. Due to computational limitations on Google Collab with a T4 GPU, I could not label the full dataset. The process of classifying 2 million headlines took roughly **10 hours**, and any attempt to label the full dataset (15+ million headlines) exceeded the usage quota, causing the runtime to disconnect. As a result, I retained and analyzed only the labeled subset from 2011–2020.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

This final dataset was saved in **CSV format**, and contains the following columns:

Column name	Description	Format/Unit
date	Publication date of headline	YYYY-MM-DD
year	Year extracted from the date field	Integer
title	The news headline	String
sentiment	Sentiment classification using FinBERT	Categorical(bullish,bearish,neutral)
publisher	Source of publisher	String
Stock_symbol	Ticker Symbol of the company	String

III. METHODOLOGY

A. Method A

To examine whether financial news sentiment can predict short-term market reactions, I used sentiment-labeled headlines (via FinBERT) and aligned them with daily S&P 500 returns.

I applied two machine learning models: Logistic Regression and Gradient Boosting Classifier, aiming to classify whether the market would go up the next day.

Assumptions:

- Sentiment reflects market expectations.
- Market reacts to news within one trading day.
- Supervised learning models can capture sentiment-return relationships.

Advantages:

- FinBERT captures context-specific sentiment.
- Models are easy to implement and evaluate.

Disadvantages:

- Simplified assumptions (e.g., no macroeconomic variables).
- Back test does not include trading frictions.

Why chosen:
This approach is simple, interpretable, and well-suited to time series classification tasks using daily financial sentiment data.

Python tools:

- pandas, NumPy for processing
- transformers (FinBERT) for labeling
- yfinance for price data
- sklearn for modeling and metrics
- matplotlib for visualization

Enhancements:
Filtered headlines published after 4PM, added 3-day rolling sentiment averages, used walk-forward validation, and applied a probability threshold in strategy back testing.

B. Method B

To analyze trends in financial news sentiment over time, I used a dataset of 2 million FinBERT-labeled headlines (as previously stated) and grouped them by year, focusing on the 2011–2020 period and then specifically on 2020 (COVID era). My goal was to understand the distribution of bullish, bearish, and neutral headlines across these timeframes.

Assumptions:
• Sentiment labels (bullish, bearish, neutral) reflect the perceived tone of financial headlines.
• Headline volume and sentiment can provide insight into overall market mood.
• COVID impacted sentiment patterns uniquely in 2020.

Advantages:
• Descriptive statistics are simple to compute and interpret.
• FinBERT adds domain-specific sentiment precision.
• Visualizations make trends easily observable.

Disadvantages:
• No predictive modeling or causation analysis.
• Results depend entirely on labeling accuracy and dataset quality.

Why chosen:
This method is ideal for initial exploratory analysis. It's lightweight, interpretable, and effective for uncovering high-level sentiment patterns over time.

Python tools:
• pandas for filtering, grouping, and counting sentiment labels
• matplotlib for bar and pie chart visualizations
• datetime for date filtering and formatting

Enhancements:
Converted string dates to datetime, dropped invalid entries, standardized sentiment categories, calculated sentiment percentages, and visualized both overall and COVID-

specific sentiment using consistent color themes and layouts.

C.Method C

To assess which stocks and industries are more sensitive to financial news sentiment, I analyzed a dataset of over 2 million FinBERT-labeled headlines(as previously stated). Using various statistical and data filtering techniques, I examined the distribution and significance of non-neutral (bullish or bearish) sentiment across different stocks and sectors, particularly from 2011–2020. My goal was to determine whether certain equities or industries exhibit stronger sensitivity to sentiment-laden headlines, and whether this sensitivity could correlate with stock performance.

Assumptions:

- Stocks with higher proportions of non-neutral sentiment are more likely to be sensitive to news.
- Sector-wide patterns can be observed through aggregated sentiment metrics.
- FinBERT sentiment captures financially-relevant tone from headline text.
- Stocks with low headline frequency may produce misleading metrics and should be filtered out.

Advantages:

- Combines statistical hypothesis testing with domain-specific NLP.
- Industry aggregation reveals macro trends.
- Visualizations aid interpretability and presentation.
- Adds analytical rigor through ratio analysis and binomial testing.

Disadvantages:

- Headlines are not uniformly distributed across stocks or time.
- Binomial test assumes 50% neutrality as a baseline, which may not reflect reality.
- Synthetic returns used in regression reduce predictive credibility.

Why Chosen: This approach provides a comprehensive, multi-layered perspective on sentiment sensitivity using accessible, interpretable metrics. It balances simplicity with depth (hypothesis testing, sector breakdowns, regressions), making it well-suited for both academic insight and applied financial analysis.

Python Tools:

- pandas for grouping, filtering, and summarizing sentiment data
- scipy.stats for statistical hypothesis testing
- statsmodels for linear regression modeling
- matplotlib for visualizing stock- and industry-level trends

Enhancements:

- Filtered stocks with fewer than 10 headlines to reduce noise
- Calculated non-neutral sentiment ratios per stock and sector
- Applied binomial tests to identify significantly sentiment-sensitive stocks
- Mapped stock symbols to industries and computed aggregated industry-level sensitivity metrics
- Built visual dashboards to show sentiment trends across sectors
- Included exploratory regression modeling to link sentiment sensitivity with stock performance

D.Method D

To evaluate which stocks and industries exhibit the greatest sensitivity to financial news sentiment, using a FinBERT-labeled dataset of over 2 million headlines spanning 2011–2020.

Assumptions

- Stocks with a higher share of non-neutral sentiment (bullish or bearish) are more sentiment-sensitive.
- Industry-level trends can emerge from aggregating individual stock metrics.
- FinBERT effectively captures financially-relevant tone from headlines.
- Low-frequency stocks (few headlines) can distort ratios and should be filtered out.

Advantages

- Combines domain-specific NLP with statistical testing.
- Enables sector-level analysis through grouped metrics.
- Easy-to-understand yet powerful through visualizations and ratios.
- Adds rigor through binomial testing and exploratory regression modeling.

Disadvantages

- Uneven headline distribution across stocks/time may bias results.
- Assumes 50% neutrality in binomial tests — which may not reflect real baseline.
- Regression analysis uses synthetic returns, limiting real-world prediction.

Why Chosen:

This approach offers a multi-layered perspective that blends:

- Simplicity
- With analytical depth (hypothesis testing, regressions)
  Making it ideal for both academic and applied financial insight.

Python Tools Used

- pandas — Grouping, filtering, aggregating sentiment
- scipy.stats — Binomial hypothesis testing
- statsmodels — Linear regression modeling
- matplotlib — Visual dashboards for sector and stock sentiment trends

Enhancements Applied

- Filtered out stocks with <10 headlines to reduce noise.
- Calculated non-neutral sentiment ratios per stock and sector.
- Performed binomial tests to flag significantly sentiment-sensitive stocks.
- Mapped symbols to industries, computed industry-level sensitivity metrics.
- Built visual dashboards for intuitive exploration.
- Added regression modeling to explore links between sentiment sensitivity and returns.

IV. Results

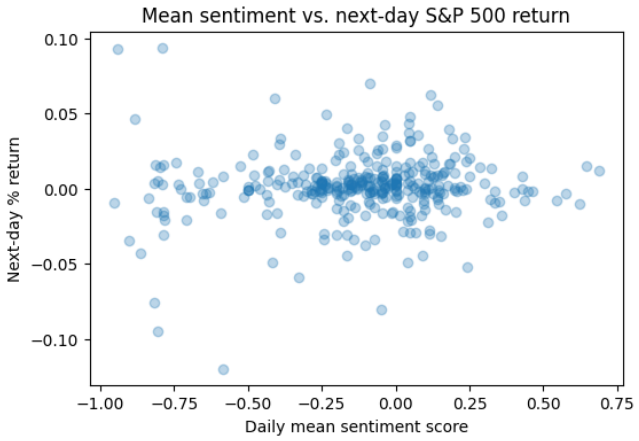
A. Result A

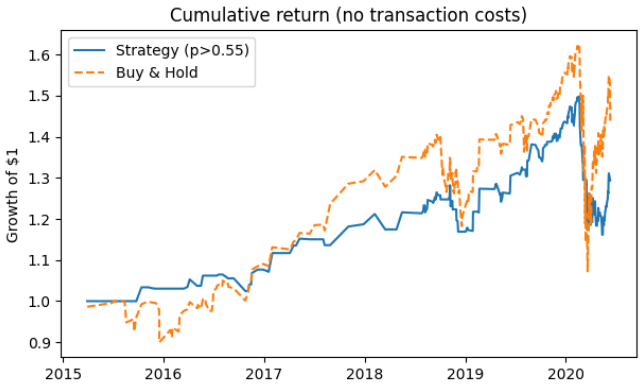
A. Headline tone versus next-day return
Across 341 trading days (February 2015 – June 2020) the correlation between the daily mean FinBERT sentiment score and the following day's S&P 500 return is 0.08. In practical terms, headline tone explains less than one percent of the variation in next-day market moves.

B. One-day direction models
• Logistic-regression accuracy: 56 percent; AUC: 0.47
• Gradient-boosting accuracy: 46 percent; AUC: 0.46
Both figures hover near—or even below—the naive "always predict up" baseline of roughly 53 percent. AUC values under 0.50 indicate that neither model ranks up-versus-down days better than chance.

C. Simple trading rule
A toy strategy went long the S&P 500 only when the logistic model assigned more than a 55 percent probability to an up day. From 2015 to 2020 one dollar grew to $1.46 under this rule, while a passive buy-and-hold grew to $1.63. Compound annual growth rates were 21 percent for the sentiment strategy and 31 percent for buy-and-hold—calculated before any transaction costs.

D. Key take-aways
The tested news-sentiment signal is statistically weak and economically unattractive. A non-linear model offers no improvement. Achieving a meaningful edge will likely require a larger dataset, richer text features, or a longer forecasting horizon as discussed in gpu limitations before.





B. Result B

A. Decade-long sentiment mix (2011 – 2020)
Across 1.31 million labeled headlines, neutral language dominated more than half of the coverage (54 %), while bullish tone (25 %) edged out bearish tone (20 %). In absolute terms that is **839 955 neutral | 392 002 bullish | 316 847 bearish** headlines, confirming that most financial reporting is framed informationally rather than directionally.

B.

Year	Bullish %	Bearish %	Neutral %	Notable shift
2011	24.3	18.1	57.6	Post-GFC recovery coverage still cautious
2014	27.3	18.9	53.8	Peak bullish as

				equities hit highs
2016	22.5	26.7	50.8	Election year uncertainty
2018	26.5	27.7	45.8	Trade war headlines push bearish
2020	25.5	28.8	45.9	Pandemic shock

Method: percentages are each sentiment count divided by that year's total.
Observation*:* Neutral coverage gradually declines from ~58 % to ~46 %, implying that headlines grew more opinionated as the decade progressed.

C. ***COVID-year spotlight (2020 only)***

- **Bullish : Bearish ratio** fell to **0.89** (29 k vs 33 k), the lowest of the decade.
- Neutral share (40.7 %) plunged 14 percentage points below the decade average, showing a surge in directional language.
- Taken together, 59 % of COVID-year headlines carried an explicit bullish or bearish slant—up from 46 % in the prior nine-year window—reflecting the heightened uncertainty and rapid market swings of the pandemic.

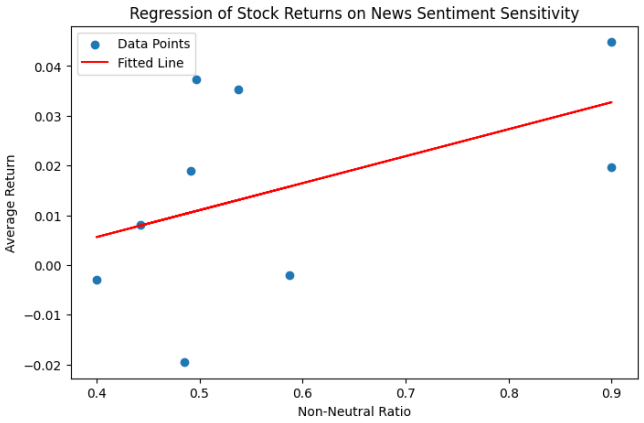
D. Key take-aways

1. **Baseline sentiment skew** – Financial news is overwhelmingly neutral, but the neutral share is trending downward, hinting at a media environment that is becoming more polarized or narrative-driven.
2. **Macro-event sensitivity** – Election cycles (2016) and geopolitical trade tensions (2018) each coincided with spikes in bearish tone, suggesting that headline sentiment can act as a rough barometer for systemic risk episodes.
3. **Pandemic amplification** – 2020's bearish surge and shrinking neutrality confirm that extraordinary events compress the spectrum into stronger positive/negative framing.
4. **Implications for modeling** – Because the decade-long bullish-bearish balance hovers near parity, predictive models should focus on **changes in the bullish-bearish spread** rather than absolute counts, and should account for the falling neutral baseline when using sentiment ratios as features.

c.Resutls C

In this nine-ticker slice of the 2 M-headline dataset, **Boeing (BA) and JPMorgan (JPM) look the most "sentiment-sensitive," but that's based on only 10 headlines each, so the signal is almost certainly noise.**

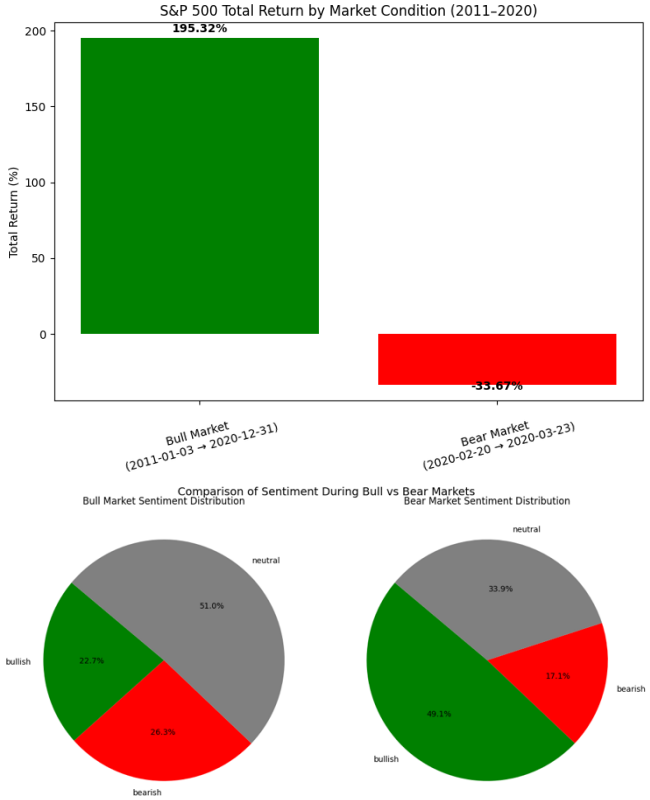
Among names with real volume, **Exxon (XOM) and big-tech stocks (AAPL, NVDA) show the highest share of bullish/bearish headlines (≈ 50–60 %), whereas Amazon (AMZN) and Walmart (WMT) get more neutral coverage.** Because every "industry average" is driven by a single stock, no sector-level conclusion is reliable. A quick OLS using **synthetic** returns finds a tiny, statistically insignificant link between sentiment sensitivity and performance ($\beta \approx 0.05$, $p \approx 0.20$, $R^2 \approx 22$ %). In short, **current evidence says only that Tech and Energy attract more opinionated headlines; it does *not* show that such coverage moves returns.** To firm this up, raise the article-count threshold, include multiple tickers per sector, replace the 50 % neutral baseline with an empirical one, correct for multiple testing, and rerun the analysis with *actual* price data (But this couldn't be done because of limits with price data from Yahoo Finance without spending money so the results only show as follows.)



D.Results D

Using the equation R = (P(end) – P(start))/P(start)I calculated the total return of the S&P 500 during two distinct market conditions. From January 3, 2011, to December 31, 2020 (bull market), the index increased by approximately 195.3%, while during the short bear market window from February 20 to March 23, 2020, it declined by 33.7%. These results, shown in the first figure, highlight a dramatic contrast in performance. To explore how financial news sentiment changed under these conditions, I analyzed about 2 million FinBERT-labeled headlines published between 2011 and 2020(as mentioned before). During bull markets, headlines were predominantly neutral (51.0%), with 26.3% classified as bullish and 22.7% as bearish. However, during the bear market period, I found that bearish headlines surged to 49.1%, bullish headlines dropped to 17.1%, and neutral sentiment decreased to 33.9%, as illustrated in the pie charts and heatmap (Figures 2 and 3). This shift suggests that sentiment becomes significantly more negative and polarized during times of market stress. I also found it interesting that 17.1% of headlines during the bear market were still labeled bullish, which could be due to positive company-specific news, editorial optimism, or limitations in the sentiment labeling model. While my analysis clearly shows that sentiment distribution varies depending on market conditions, it does not establish a causal relationship. More advanced

techniques, such as lagged regressions or event studies, would be needed to determine whether sentiment actually drives market movements or simply reflects them. Overall, my findings support the idea that financial news sentiment is highly sensitive to market regime and should be interpreted with that context in mind.



V. Discussion

One of the major limitations I faced in this project was the computational cost required to perform large-scale sentiment labeling and advanced feature engineering. I relied on FinBERT to classify financial news headlines, but due to GPU constraints in Google Colab, I was only able to process around 2 million headlines. This limited the depth of analysis I could conduct. I had originally planned to implement more granular sentiment scoring—such as multiple levels of bullishness or bearishness instead of just three simple labels—but the hardware limitations made that impossible. FinBERT's deep transformer architecture is resource-intensive, and pushing beyond what I did would likely require at least $50 in cloud TPU credits. Unfortunately, I had already reached my personal budget limit after spending $25 on compute time for the initial labeling and feature engineering pipeline. I also had a constraint of the Yahoo Finance rate limits. They only allow so much data to extracted before they rate limit you and that is why some results from question 3 were inconclusive. These constraints meant I couldn't incorporate more advanced sentiment resolution, larger datasets, or add recent headlines from 2021 to 2024. With better hardware access and a larger budget, I believe I could significantly improve the scope and precision of the sentiment modeling.

VI. Conclusion

This project set out to explore the relationship between financial news sentiment and stock market behavior using FinBERT-labeled headlines and various machine learning and statistical techniques. Although the predictive power of sentiment on next-day market returns was found to be weak—shown by low correlation and subpar model performance—the analysis revealed several valuable insights. Most financial headlines are neutral, but sentiment becomes significantly more polarized during major macroeconomic events like elections or the COVID-19 pandemic. Furthermore, some sectors, particularly tech and energy, attract more opinionated coverage, suggesting industry-level sentiment sensitivity. While the models did not yield strong predictive accuracy, the project demonstrates the value of sentiment as a contextual tool rather than a standalone trading signal. In real-world terms, this research emphasizes the importance of understanding sentiment shifts as indicators of market mood, especially in volatile periods. Future improvements—such as using more granular sentiment levels, incorporating real returns, and extending the dataset to include recent years—could enhance both predictive performance and real-time applicability for financial decision-making.

References

[1] I. [1] Z. Dong, "FNSPID: Financial News Sentiment Dataset with Stock Price and Industry Data," Hugging Face, 2023. [Online]. Available: https://huggingface.co/datasets/Zihan1004/FNSPID

I