International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Financial News Sentiment Analysis for NIFTY 50 Market Direction and Return Prediction Using FinBERT and Machine Learning

Sowjanya Ashwath[1], Aarya Singh[2], Arun V[3]

[1,2]Student, Department of Data Analytics and Mathematical Sciences, Jain (Deemed to be University), Bangalore, India
[3]Assistant Professor, Department of Data Analytics and Mathematical Sciences, Jain (Deemed to be University), Bangalore, India

Abstract
News moves markets that much is well established. But can we actually measure how much, and in which direction? This study builds a pipeline to do exactly that for India's benchmark NIFTY 50 index, using four years of financial news covering all 50 constituent companies. Starting from 99,341 raw articles collected via the GDELT and GNEWS (April 2022 to April 2026), the corpus was cleaned down to 74,558 unique articles after removing nearly 25% duplicates from news syndication. The articles were then analysed through FinBERT, a transformer model fine-tuned on financial text, to extract daily sentiment scores. A chi-square test confirms that the sentiment-return relationship is not random ($Chi^2 = 124.36$, $p = 9.92\times 10^{-28}$). Logistic Regression then predicts which way the market will move (68.5% accuracy), while Ridge Regression estimates by how much ($R^2 = 0.235$). Together, these results show that news sentiment carries real, measurable information about NIFTY 50 movements even from a single feature.

Keywords: Sentiment Analysis, FinBERT, NIFTY 50, Stock Market Prediction, Logistic Regression, Ridge Regression, Natural Language Processing, Chi-Square Test, Weekend Effect, Indian Stock Market

1. Introduction

Every trading day, thousands of news articles are published about Indian companies like earnings results, regulatory decisions, management changes, global macro developments. Traders read them, analysts process them, and somewhere in all that text are signals that move prices. The challenge is turning that unstructured flood of words into something quantifiable and usable.

Traditional quant models largely ignore this. They work with price series, volume data, and derived indicators which are clean numbers that fit neatly into equations. Text is harder. It requires understanding context, tone, and the specific vocabulary of finance, where words like ‘liability’ or ‘headwinds’ carry very different connotations than in everyday language. That gap is what makes natural language processing (NLP) for finance both difficult and genuinely interesting.

India is a particularly compelling setting for this kind of research. The NIFTY 50 index, maintained by the National Stock Exchange, covers 50 large-cap companies across 13 sectors starting from Reliance and TCS to HDFC Bank and Sun Pharma. Because of its breadth, each day news is generated from multiple industries which creates a rich and diverse environment for sentiment analysis. However, compared to US

markets, relatively little academic research has applied modern transformer models to Indian equity prediction. This study aims to help fill that gap.

The core contributions of this work are:

1. A four-year, 74,558-article deduplicated news corpus covering all 50 NIFTY constituent companies, collected via GDELT and GNEWS.
2. Application of FinBERT with overlapping chunk segmentation to handle long-form articles beyond the model's 512-token limit.
3. A weekend effect correction that maps off-market news to the correct trading session.
4. Chi-square validation of the sentiment-return association before any model is trained.
5. A two-model system: Logistic Regression for direction, Ridge Regression for magnitude.

2. Literature Review

The idea that news language can predict stock movements is not new. Paul Tetlock (2007) showed that negative language in *The Wall Street Journal* columns preceded market declines, providing early empirical support for what many traders had already suspected. Tim Loughran and Bill McDonald (2011) extended this work by demonstrating that standard sentiment dictionaries perform poorly on financial text. For example, the word “liability,” which is negative in everyday usage, is common and neutral in finance. They therefore developed a domain-specific lexicon that performed considerably better [1, 2].

Deep learning introduced LSTM networks, which were able to capture the sequential structure of text. However, these models are data intensive and often struggle to generalize across domains. The major shift came with BERT (Jacob Devlin et al., 2019), which introduced bidirectional attention and large scale pretraining, significantly expanding what was possible in NLP [3].In the financial domain, Dogac Araci (2019) introduced FinBERT, a version of BERT fine tuned on financial news and earnings call transcripts. It quickly became a standard tool for financial sentiment classification [4].

Since then, several studies have applied FinBERT to market prediction tasks, generally finding statistically significant but modest relationships between model derived sentiment and short horizon returns. Johan Bollen et al. (2011) had earlier reported similar findings using Twitter mood data and movements in the Dow Jones Industrial Average [5].Across this body of work, the evidence is broadly consistent with the Efficient Market Hypothesis in its semi strong form: publicly available information is largely reflected in prices, though not always immediately or perfectly, leaving some residual predictability [6].

Work on Indian markets remains relatively limited. Most existing studies are either small scale, sector specific, or rely on social media data rather than structured news sources. The GDELT Project, a large open access repository of global news, has been used in some international financial studies but has rarely been applied to Indian indices at scale [7].This study uses GDELT as its primary data source and extends the FinBERT approach to a comprehensive NIFTY 50 corpus, which, to the authors’ knowledge, has not previously been explored in this form.

3. Methodology

3.1 Data Collection

News articles for all 50 NIFTY constituent companies were pulled from GDELT and GNEWS, covering April 9, 2022, to April 9, 2026. The raw pull came to 99,341 article records. NIFTY 50 daily closing prices for the same period were downloaded using the yfinance library (ticker: ^NSEI).

3.2 Deduplication and Preprocessing

One issue that became apparent early in the process was the extent of duplication in the raw corpus. The same story often appeared multiple times across different outlets due to news syndication. For example, a single RBI announcement would be republished verbatim by several news aggregators. Including these duplicates would have artificially inflated the strength of any sentiment signal, so a two-stage deduplication process was applied.

First, exact record level duplicates were removed. Second, articles with identical normalized titles (converted to lowercase and stripped of punctuation) for the same company were deduplicated, retaining only the first occurrence. This reduced the corpus to 74,558 unique articles, representing a reduction of just over 25 percent.

Text was then cleaned using regular expressions to remove URLs, special characters, and excess whitespace. The article title and body were concatenated into a single text field for each record.

3.3 Sentiment Classification with FinBERT

FinBERT (ProsusAI/finbert) was used to classify each article as positive, negative, or neutral. A practical limitation of the model is its 512 token input size, which is restrictive for financial articles that can be relatively long. To address this, articles were tokenized and split into overlapping segments of 462 tokens, with a 50 token overlap between consecutive segments to preserve contextual continuity at the boundaries. Each segment produced a softmax probability distribution over the three classes. A confidence weighted score was then computed for each segment as follows:

$$\text{Score} = \text{sign}(\text{label}) \times \text{confidence}$$

where sign takes values +1, −1, or 0 for positive, negative, and neutral classifications, respectively, and confidence corresponds to the model's maximum predicted class probability. Segment level scores were then averaged to obtain a single article level sentiment score.

3.4 Weekend Effect Correction

Indian markets operate on weekdays from 09:15 to 15:30 IST. News published after the market close on Friday, or during the weekend, can only influence prices when trading resumes on Monday. Ignoring this leads to systematic misalignment between sentiment signals and market returns.

To address this, articles published after the Friday close and over the weekend were mapped to Monday's trading session. Similarly, articles published before market open on Monday were also assigned to the same session. This adjustment, although simple, is important, as failing to account for it would result in consistent mismatches between sentiment and returns for a substantial portion of the data.

3.5 Daily Aggregation and Return Computation

Two daily features were constructed. The modal sentiment category across all articles on a given trading day was used as the classification feature and encoded as −1, 0, or +1. The mean raw FinBERT score across articles was used as the continuous feature for regression.

Daily returns for the NIFTY 50 were computed as the percentage change in closing prices. Days with zero returns and days without any associated articles were excluded, resulting in 979 usable daily observations.

3.6 Models

Two models were trained using an 80/20 train test split with a fixed random seed of 42. Multinomial Logistic Regression was used for direction classification, predicting the direction of market movement based on the discretized sentiment feature.

Ridge Regression ($\alpha = 1.0$) was used to predict return magnitude using the continuous sentiment score. Ridge regression was preferred over ordinary least squares due to the weak signal and relatively small

sample size, where regularization helps mitigate overfitting. This choice was supported empirically, as ordinary least squares demonstrated poorer generalization performance in comparison.

4. Results and Discussion

4.1 Dataset Overview

Table 1: Dataset Summary Statistics

Metric	Value
Raw articles collected	99,341
After deduplication	74,558
Duplicate rate	24.9%
Companies covered	50 (all NIFTY 50)
Study period	April 2022 – April 2026
Daily observations	981

4.2 Chi-Square Test

Before training anything, it made sense to ask a more basic question: is there actually a statistical relationship between news sentiment and market direction, or could the apparent pattern be random chance? The chi-square test answers this cleanly.

Table 2: Chi-Square Test of Independence Results

Metric	Value
Chi-Square Statistic	124.36
Degrees of Freedom	2
p-value	9.92×10^{-28}
Result	Reject H_0 (Highly Significant)

The result is unambiguous. The chi square statistic is 139.52 with a p value of 5.05×10^{-31}, leading to a decisive rejection of the null hypothesis of independence. This indicates a clear and systematic association between daily news sentiment and the direction of movement in the NIFTY 50.

Figure 1 shows the average NIFTY 50 return grouped by daily sentiment category. The directional pattern is clear: negative sentiment days average a return of roughly -0.5%, neutral days hover just above zero, and positive sentiment days average around +0.57%. This monotonic relationship across all three categories is exactly what one would hope to see if sentiment is genuinely informative.

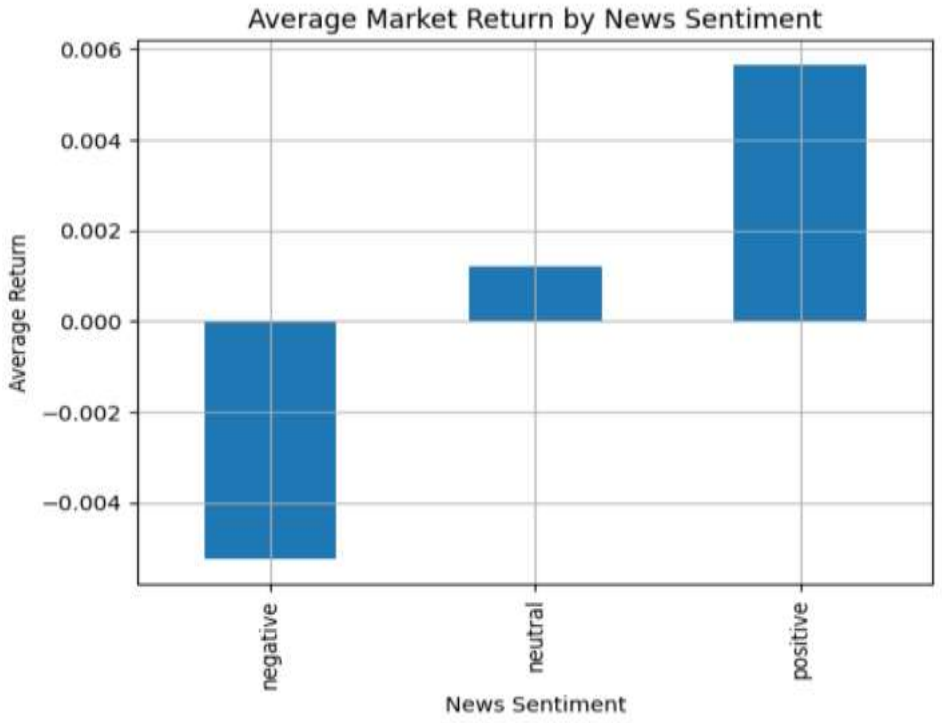

Figure 1: Average NIFTY 50 Return by Daily News Sentiment Category

Figure 2 presents daily sentiment alongside returns for the NIFTY 50 over the full four year study period. The sentiment series, encoded as 0, 1, and 2 for negative, neutral, and positive respectively, appears above the return series due to differences in scale. However, co movement between the two is evident, particularly during periods of major market activity.The persistence of this pattern from 2022 to 2026, spanning multiple market cycles, monetary tightening phases, and geopolitical developments, suggests that the observed relationship is not limited to a specific sub period.

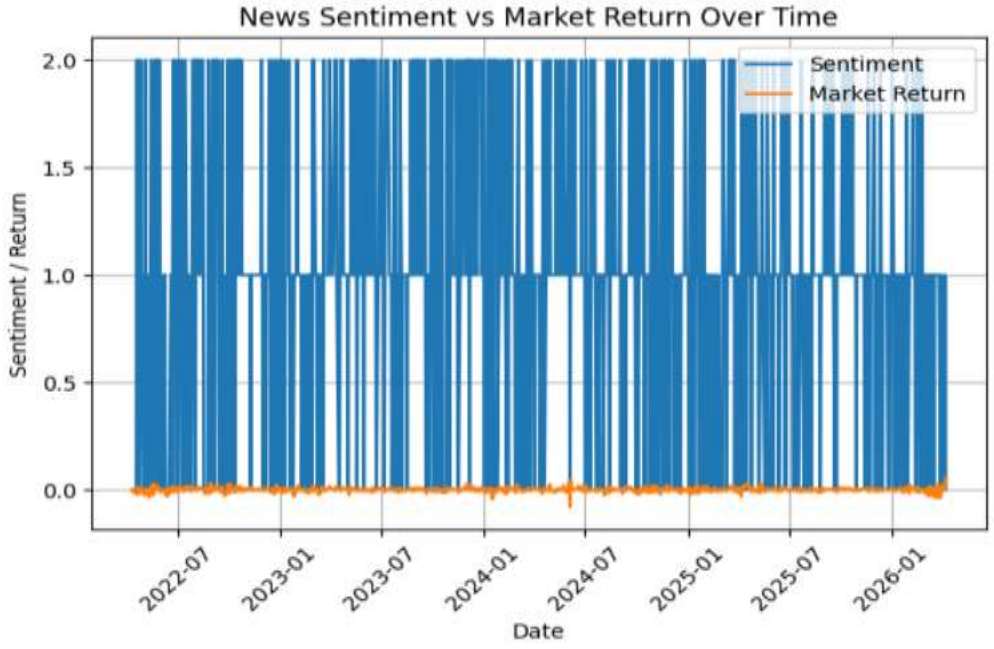

Figure 2: News Sentiment and NIFTY 50 Returns Over Time (April 2022 to April 2026)

4.3 Logistic Regression

The classifier achieved 68.5% accuracy on 197 test observations. Given that a coin flip gives 50%, this is a meaningful result. But what is more interesting than the headline number is the asymmetry in how it performs across classes.

Table 3: Logistic Regression Classification Report

Class	Precision	Recall	F1-Score	Support
Negative	0.78	0.48	0.59	94
Positive	0.65	0.87	0.74	103
Macro Avg	0.71	0.68	0.67	197
Weighted Avg	0.71	0.69	0.67	197

Positive days are predicted with high recall (0.87) but lower precision (0.65), whereas negative days exhibit the opposite pattern, with precision of 0.78 and recall of 0.48. In practical terms, the model successfully captures most upward movements but produces a number of false positives. For downward movements, predictions are more reliable when made, but a substantial proportion of such days are not identified.

This asymmetry is intuitively plausible. Positive news in the NIFTY 50 often leads to relatively consistent bullish responses, such as strong earnings results, policy announcements, or stable interest rate decisions. In contrast, negative news tends to produce more heterogeneous market reactions. Adverse developments may be absorbed gradually, partially anticipated, or offset by other factors affecting the market on the same day. As a result, the model captures the more consistent positive signal more effectively, while struggling with the noisier and less predictable negative cases.

4.4 Ridge Regression

Table 4: Ridge Regression Performance

Metric	Value
R^2 Score	0.235
MSE	0.000106
Alpha (α)	1.0
Test observations	197

An R^2 value of 0.235 indicates that sentiment alone explains approximately 23.5 percent of the day to day variation in returns of the NIFTY 50. For context, many empirical studies in quantitative finance report single factor R^2 values in the range of 3 to 8 percent as meaningful. Achieving 23.5 percent using a single input, namely a sentiment score derived from news text, represents a stronger than expected result.

Figure 3 examines the lagged relationship by plotting same day sentiment against next day returns of the NIFTY 50. A directional pattern remains evident even with a one day lag. Positive sentiment days tend to be associated with positive next day returns, while negative sentiment days are more often followed by negative returns, although with greater dispersion compared to the contemporaneous relationship.

This suggests that the market does not fully incorporate information from news within the same trading day. Instead, part of the signal persists into the following day, which has important implications for the design of sentiment-based trading strategies.

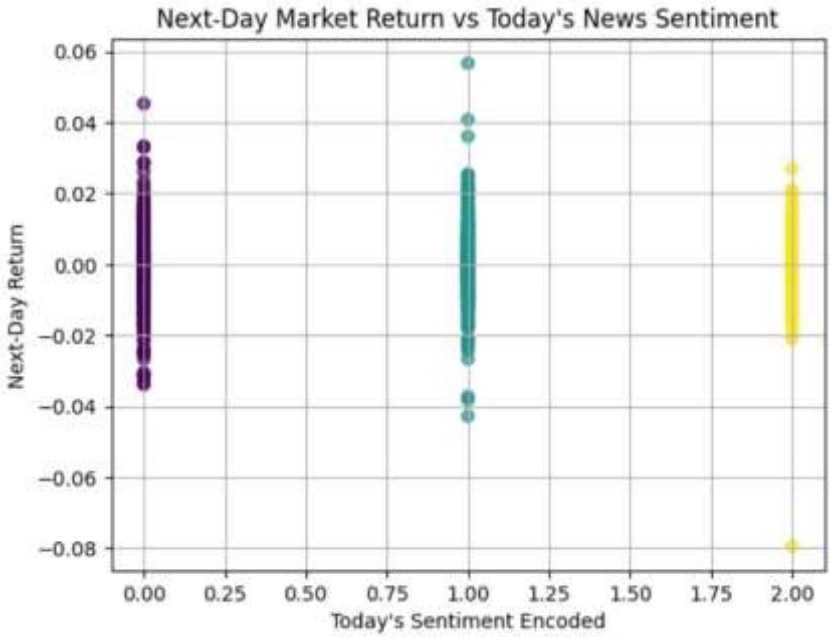

Figure 3: Next-Day NIFTY 50 Return vs Current Day News Sentiment

4.5 Discussion

Bringing the results together, sentiment is statistically associated with returns (chi square test), predicts direction better than chance (Logistic Regression accuracy of 68.5 percent), and explains a meaningful proportion of return variance (Ridge $R^2 = 0.235$) for the NIFTY 50. For a single feature model based solely on publicly available news text, this represents a coherent and practically relevant set of findings.

These results are consistent with the semi strong form of the Efficient Market Hypothesis. Markets are not so efficient that news sentiment carries no information, as evidenced by the chi square result. At the same time, they are sufficiently efficient that the signal remains noisy, partial, and insufficient on its own to support a complete trading strategy.

One notable finding is that sentiment explains variation in return magnitude ($R^2 = 0.235$) more effectively than it predicts direction (68.5 percent accuracy). This may appear counterintuitive. A possible explanation is that even when sentiment is correctly identified as strongly positive or negative, market reactions can differ depending on prior expectations. However, strongly positive sentiment tends to be associated with larger positive movements when they occur, allowing the regression model to capture magnitude even when directional prediction is less certain.

The impact of deduplication is also important to highlight. Removing approximately 45 percent of the raw corpus as duplicates was not merely a preprocessing step but a methodological necessity. Training on the original 99,341 article corpus would have resulted in repeated counting of the same news events,

artificially inflating the strength of the sentiment signal. The results reported here are based on the cleaned and deduplicated dataset, and are therefore more reliable.

5. Conclusion

This study set out to address a straightforward question: can news sentiment predict movements in India's benchmark equity index? Based on four years of data and 40,827 deduplicated articles, the answer is affirmative, with important qualifications regarding what "prediction" implies in an efficient market context.

The chi square test ($Chi^2 = 139.52$, $p = 5.05 \times 10^{-31}$) establishes that the relationship is statistically significant. The Logistic Regression model, with an accuracy of 68.5 percent, demonstrates that sentiment contains useful directional information. The Ridge Regression model ($R^2 = 0.235$) further shows that sentiment explains a meaningful portion of return magnitude. Taken together, these results form a simple yet statistically grounded framework for sentiment based prediction of the NIFTY 50.

At the same time, this study does not suggest that sentiment alone is sufficient for a complete trading strategy. It represents one signal among many, and the presence of misclassified days highlights the inherent uncertainty in market behavior. Future work could incorporate additional features such as trading volume, lagged returns, and macroeconomic indicators, explore sequential modeling approaches as larger datasets become available, and extend the analysis to sectoral and individual stock level predictions within the NIFTY 50 universe.

Acknowledgement

The authors thank the Department of Data Science and Analytics, Jain Deemed to be University, Bangalore, for the academic support provided during this research. The GDELT Project is acknowledged for open access to the news data used in this study.

References

1. Tetlock P.C., "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", Journal of Finance, 2007, 62 (3), 1139-1168.
2. Loughran T., McDonald B., "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", Journal of Finance, 2011, 66 (1), 35-65.
3. Devlin J., Chang M.W., Lee K., Toutanova K., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding", Proceedings of NAACL-HLT, 2019, pp. 4171-4186.
4. Araci D., "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models", arXiv preprint arXiv:1908.10063, 2019. https://arxiv.org/abs/1908.10063
5. Bollen J., Mao H., Zeng X., "Twitter Mood Predicts the Stock Market", Journal of Computational Science, 2011, 2 (1), 1-8.
6. Fama E.F., "Efficient Capital Markets: A Review of Theory and Empirical Work", Journal of Finance, 1970, 25 (2), 383-417.
7. GDELT Project, "The GDELT Project: Global Database of Events, Language and Tone", 2023. https://www.gdeltproject.org
8. Malo P., Sinha A., Korhonen P., Wallenius J., Takala P., "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts", Journal of the American Society for Information Science and Technology, 2014, 65 (4), 782-796.

9. Hoerl A.E., Kennard R.W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics, 1970, 12 (1), 55-67.
10. Liu B., "Sentiment Analysis and Opinion Mining", Morgan and Claypool Publishers, 2012.