International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Comparative Analysis of NLP Frameworks FinBERT & VADER in Predicting Market Sentiment of NIFTY 50

Anjani Shivraj

Army Public School Noida, Uttar Pradesh, India

Abstract: ***Financial news headlines constitute a critical source of real-time information for investors, analysts, and policymakers operating in equity markets. Sentiment analysis of such headlines provides valuable insights into market psychology and supports data-driven investment decision-making. This paper presents a comparative analysis of two sentiment analysis frameworks, namely Valence Aware Dictionary for sEntiment Reasoning (VADER), a lexicon-based model, and FinBERT, a domain-specific transformer model pre-trained on financial corpora, applied to financial news headlines pertaining to the NIFTY 50 index of the National Stock Exchange (NSE) of India. Headlines were collected from leading Indian financial news platforms and manually classified into three sentiment categories: positive, negative, and neutral, from a financial perspective. Following standard text preprocessing procedures including tokenization, stop word removal, and lemmatization, the headlines were processed through both models. Model performance was evaluated using four metrics: accuracy, sensitivity, specificity, and neutral specificity. Furthermore, this study investigates the challenges encountered by both models when processing Indian financial terminology, market-specific abbreviations, and the linguistic nuances characteristic of NIFTY 50 news headlines. Statistical hypothesis testing is employed to validate the findings. Experimental results reveal that neither model demonstrates uniform superiority across all evaluation metrics. FinBERT achieves higher accuracy and sensitivity, reflecting its capacity to leverage contextual understanding derived from domain-specific pre-training, while VADER demonstrates comparatively stronger performance in neutral specificity, attributed to its structured lexicon-based classification approach. These findings underscore the importance of metric-specific model selection in financial sentiment analysis applications.***

Keywords: Sentiment *analysis, VADER, FinBERT, NIFTY 50, Financial news*

1. Introduction

The rapid expansion of India's capital markets has fundamentally transformed the landscape of retail investment and financial information consumption. The registered investor base at the National Stock Exchange of India (NSE) nearly tripled between March 2020 and March 2024, reaching 9.2 crore, potentially translating into approximately 20 percent of Indian households channelling their savings into financial markets. Market capitalisation on the NSE rose from USD 2 trillion in 2017 to USD 5 trillion within six months from December 2023, reflecting one of the most rapid expansions among major global exchanges. Over the past five years, the NIFTY 50 and NIFTY 500 indices have delivered strong annualised returns of 15 percent and 18 percent respectively. This unprecedented growth in retail participation has correspondingly intensified the demand for timely and accurate financial information through digital news platforms.

Financial news headlines serve as a primary channel through which retail investors form opinions, assess market conditions, and make investment decisions (Kolbitsch and Maurer, 2006; van Ooijen et al., 2019). The concise nature of such headlines makes them particularly suited to automated sentiment analysis, a branch of Natural Language Processing (NLP) that identifies and classifies emotional tone in textual data (Al-Qablan et al., 2023; Shayaa et al., 2018). In the context of equity markets, sentiment analysis of financial news has demonstrated considerable potential in gauging investor sentiment, identifying market trends, and supporting predictive modelling of index movements (Dahal et al., 2023; Nemes and Kiss, 2021; Das et al., 2021; Koukaras et al., 2022). Prior research has specifically examined the relationship between news sentiment and NIFTY 50 index movements, applying lexicon-based scoring to quantify the directional influence of financial news on market behaviour.

Despite the growing body of research in financial sentiment analysis, significant gaps remain in the literature with respect to emerging market contexts. The majority of existing studies have focused on well-developed markets, employing datasets from sources such as Twitter, FOMC minutes, or US financial news platforms (Das et al., 2021; Al-Shabi, 2020; Kim et al., 2024; Pano and Kashef, 2020). Comparatively little attention has been directed toward evaluating sentiment tools on Indian financial news, which presents distinct linguistic characteristics including market-specific terminology, domestic regulatory acronyms, and macroeconomic narratives particular to an emerging economy (Dahal et al., 2025; Srivastava et al., 2022).

This paper addresses this gap by presenting a comparative evaluation of VADER and FinBERT applied to NIFTY 50 financial news headlines. VADER is a rule-based lexicon model requiring no training data, making it computationally efficient and interpretable (Hutto and Gilbert, 2014; Bonta et al., 2019; Al-Natour and Turetken, 2020). FinBERT, introduced by Araci (2019), is a language model based on BERT that, even with a smaller training set and fine-tuning only part of the model, outperforms state-of-the-art machine learning methods on financial sentiment datasets. FinBERT is pre-trained on the TRC2-financial corpus comprising approximately 1.8 million Reuters news articles, enabling it to internalise the lexicon, syntax, and thematic content typical of financial texts. However, while FinBERT achieves

stronger performance in detecting positive and negative sentiments, its accuracy on neutral sentiment classification remains an area requiring further improvement. The central question this study seeks to answer is whether domain-specific pre-training confers a measurable advantage over lexicon-based methods when applied to Indian equity market news (Araci, 2019; Mujahid et al., 2023; Saha et al., 2022).

The comprehensive vision of this study is illustrated in Figure 1. Financial news headlines are first collected and manually labelled into positive, negative, and neutral categories. The data then undergoes standard preprocessing procedures including tokenization, lemmatization, and stop word removal. Both VADER and FinBERT are subsequently applied to generate sentiment predictions, which are evaluated using four performance metrics: accuracy, sensitivity, specificity, and neutral specificity. Statistical hypothesis testing using the Wald test is employed to validate all findings. The primary contributions are: (a) a systematic comparison of VADER and FinBERT on manually labelled NIFTY 50 headlines; (b) evaluation across four performance metrics; and (c) statistical validation through hypothesis testing. The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 overviews both models. Section 4 defines performance metrics. Section 5 describes the experimental design. Section 6 presents the discussion and conclusion.

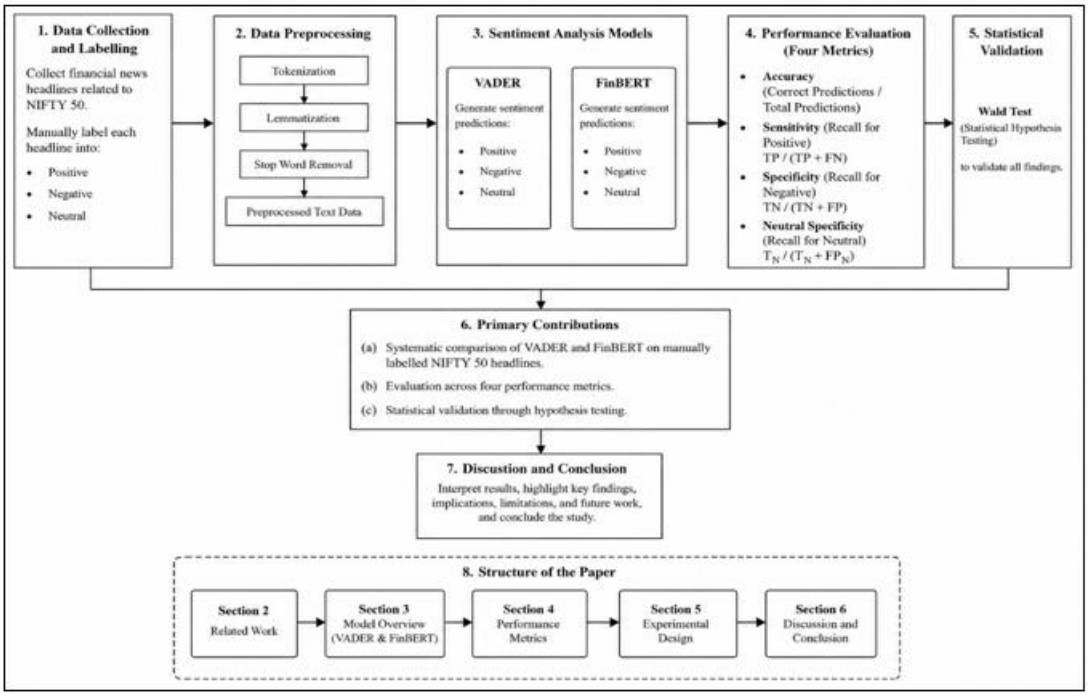

Figure 1: Schematic diagram of the proposed research framework

2. Literature Survey

In recent years, sentiment analysis applied to financial news has emerged as a widely studied approach for understanding market behaviour and supporting investment decision-making. This section reviews prior studies organised around three themes: the use of lexicon-based methods, the application of transformer-based models, and comparative evaluations across sentiment analysis tools.

2.1 Lexicon-Based Sentiment Analysis in Financial Contexts

VADER has been extensively employed across financial sentiment analysis studies owing to its computational efficiency and interpretability (Hutto and Gilbert, 2014; Bonta et al., 2019). Das et al. (2021) conducted a comparative evaluation of seven sentiment analysis tools across datasets sourced from Twitter, Facebook, Economic Times headlines, and stock market news articles, finding that VADER achieved the highest overall positive sentiment score of 56.63 percent among lexicon-based tools. Al-Shabi (2020) evaluated five prominent sentiment analysers including VADER, SentiWordNet, SentiStrength, Liu and Hu opinion lexicon, and AFINN-111, concluding that VADER achieved the highest accuracy in classifying positive and negative sentiments on Twitter data. Srivastava et al. (2022) compared lexicon-based models including VADER and AFINN-111 against machine learning approaches, finding that VADER outperformed AFINN-111 with an accuracy of 88.7 percent compared to 86.0 percent. Researchers have also combined VADER with other techniques such as Long Short-Term Memory networks to create hybrid models for improved stock price forecasting, and have modified VADER by incorporating financial lexicons to enhance its performance specifically in the financial domain. A study specifically focused on the Indian stock market applied VADER to Twitter data related to the State Bank of India, finding that sentiment analysis alone achieved an accuracy of approximately 60 percent in predicting stock price direction. However, the inherent limitation of VADER as a general-purpose lexicon designed for social media text means it may struggle with formally written financial news, where sentiment is conveyed through domain-specific terminology rather than emotionally expressive language (Al-Qablan et al., 2023; Dahal et al., 2025).

2.2 Transformer-Based Models in Financial Sentiment Analysis

The introduction of FinBERT by Araci (2019) marked a significant advancement in domain-specific financial sentiment analysis. FinBERT is a language model based on BERT that, even with a smaller training set and fine-tuning only part of the model, outperforms state-of-the-art machine learning methods on financial sentiment analysis datasets. Subsequent research has explored hybrid approaches combining FinBERT embeddings with lightweight classifiers or feeding transformer outputs into downstream LSTM and Bi-LSTM price models, offering strong performance-efficiency trade-offs. Nemes and Kiss (2021) compared BERT, VADER, TextBlob, and a recurrent neural network on company news headlines, finding that BERT and the RNN were more accurate in identifying stock value change timings. Mujahid et al. (2023) proposed a deep transformer-based BERT model for sentiment analysis of ChatGPT-related tweets, achieving an accuracy of 96.49 percent. FinBERT has also been applied to approximately ten years of business news data from energy sector companies listed on the New York Stock Exchange, demonstrating the model's capacity to extract meaningful sentiment signals from domain-specific news across extended time periods. Saha et al. (2022) applied both VADER and BERT to COVID-19 and Omicron tweet datasets, finding that BERT improved classification performance across most supervised machine learning algorithms, with SVM achieving 92 percent accuracy on the Omicron dataset using BERT features.

2.3 Comparative Evaluations and Emerging Market Contexts

Several studies have directly compared lexicon-based and transformer-based approaches, consistently finding that domain-specific models outperform general-purpose lexicons on financial text. Research assessing four sentiment models across multiple financial datasets found that FinBERT consistently outperformed lexicon-based approaches, achieving an F1-score of 93.27 percent on the SEntFiN dataset and an accuracy of 83.7 percent on the FIQA and PhraseBank dataset. A study evaluating multiple language models on FOMC minutes found that VADER achieved only 44.3 percent accuracy, while FinBERT achieved 59.7 percent, with further fine-tuning on domain-specific texts raising accuracy to 63.8 percent. Dahal et al. (2025) compared VADER and TextBlob on financial news headlines from Nepal and the USA, finding that VADER outperformed TextBlob in accuracy, sensitivity, and specificity across both datasets. Notably, most existing comparative studies have focused on well-developed markets such as the United States, with limited attention directed toward emerging markets such as India (Dahal et al., 2025; Srivastava et al., 2022). Prior work examining NIFTY 50 specifically has applied sentiment scoring to news articles to quantify their directional influence on index movements, highlighting the relevance of news-based sentiment analysis in the Indian equity market context. The present study addresses the identified gap by directly comparing VADER and FinBERT on NIFTY 50 financial news headlines, contributing to the underexplored area of sentiment analysis in Indian equity market research.
(Hutto and Gilbert, 2014; Bonta et al., 2019; Al-Shabi, 2020; Dahal et al., 2025).

3. Modelling Approach

This section details the mathematical and structural configurations of the sentiment classification engines utilized in our empirical evaluation. To ensure a robust comparative baseline, the research design incorporates two highly prominent yet distinct NLP frameworks: the lexicon-based VADER classifier and the transformer-based FinBERT model. While the former relies on human-curated vocabulary weights and morphological rules, the latter utilizes deep self-attention mechanisms pre-trained on massive financial corpora. The following subsections clarify their technical design parameters, mathematical feature scaling systems, and distinct linguistic processing capabilities.

3.1 VADER

VADER is a simple rule-based model for general sentiment analysis that employs a combination of qualitative and quantitative methods to construct and empirically validate a gold-standard list of lexical features attuned to sentiment in microblog-like contexts, combining these lexical features with five general rules that embody grammatical and syntactical conventions for expressing and emphasising sentiment intensity. The lexicon is empirically validated by multiple independent human judges and is especially attuned to microblog-like contexts. Each word in the lexicon is assigned a valence score ranging from -4 (most negative) to +4 (most positive), enabling differentiated measurement of sentiment intensity rather than simple polarity classification

VADER incorporates five generalizable heuristics that go beyond what would normally be captured in a typical bag-of-words model and incorporate word-order sensitive relationships between terms, namely: punctuation, where the exclamation point increases the magnitude of intensity without modifying semantic orientation; capitalisation, where the use of ALL-CAPS emphasises sentiment intensity; degree modifiers, where intensifiers or dampeners adjust the sentiment score; contrastive conjunctions such as "but" which shift polarity; and tri-gram examination to identify negation cases. By examining the tri-gram preceding a sentiment-laden lexical feature, VADER catches nearly 90 percent of cases where negation flips the polarity of the text. VADER calculates sentiment scores by evaluating the sentiment of individual words while taking into account the surrounding context, considering the impact of intensifiers, negations, and the sentiment of emojis and emoticons.

For any given text, VADER tokenises the input, retrieves the valence score for each token, and computes a compound score using the formula illustrated in Figure 2 (Hutto and Gilbert, 2014; Dahal et al., 2025). The positive and negative polarity scores assigned by the VADER lexicon are computed by summing the sentiment intensities of individual words as matched against VADER's predefined dictionary, which maps words and phrases to sentiment values, with the normalisation constant $\alpha = 15$ ensuring the compound score is bounded between -1 and +1. For classification purposes, a compound score of ≥ 0.05 is assigned a positive label, ≤ -0.05 a negative

label, and any value in between a neutral label, as illustrated in Figure 2 (Hutto, 2020; Dahal et al., 2025)

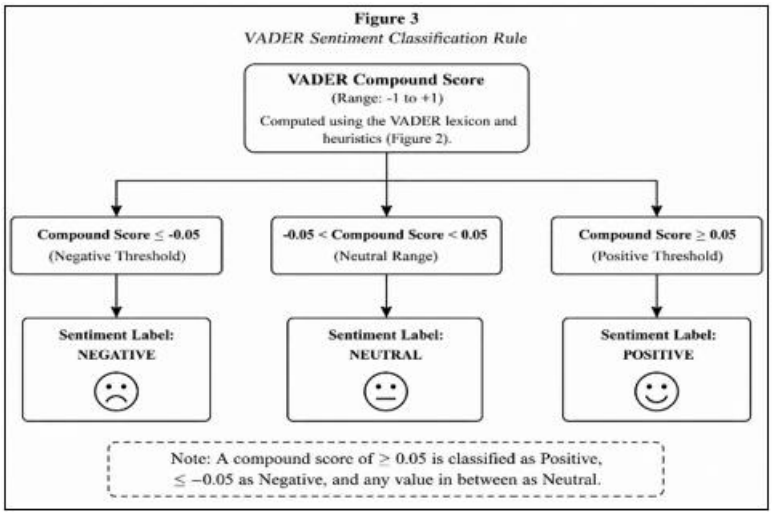

Figure 2: VADER Compound Score Formula

VADER outperforms individual human raters in assessing the sentiment of tweets and generalises more favourably across contexts than eleven typical state-of-practice benchmarks including LIWC, ANEW, the General Inquirer, SentiWordNet, and machine learning oriented techniques relying on Naive Bayes, Maximum Entropy, and Support Vector Machine algorithms. The worked example in Table 2 illustrates VADER's scoring applied to a representative NIFTY 50 headline (Hutto and Gilbert, 2014; Ekaputri and Akbar, 2022; Pokhrel et al., 2024).

Full sentence compound score: 0.7269 (Positive)

VADER offers several practical advantages including computational efficiency, no requirement for labelled training data, and strong interpretability, making it widely adopted in financial news sentiment applications (Dahal et al., 2023; Maqbool et al., 2023; Ekaputri and Akbar, 2022; Pokhrel et al., 2024; Das et al., 2021). However, its general-purpose lexicon, primarily designed for social media text, represents a notable limitation when applied to formally written financial news where sentiment is conveyed through domain-specific terminology rather than emotionally charged language (Al-Qablan et al., 2023; Dahal et al., 2025; Srivastava et al., 2022).

3.2 FinBERT

FinBERT is based on the idea of training a BERT model in two steps to adapt it to the financial domain and the sentiment analysis task, where the first step consists of further pre-training the model on financial documents, a strategy proven effective for domain adaptation, with the aim of helping the model understand financial terminologies better than the base model. FinBERT overcomes the limitations of general sentiment models by using transfer learning, enabling it to be fine-tuned on financial data to better understand the context and nuances of financial language, with its strength lying in its ability to leverage pre-trained knowledge from large language corpora and adapt to the financial domain. FinBERT excels in identifying the positive or negative sentiment of sentences that other algorithms mislabel as neutral, likely because it uses contextual information in financial text, and substantially outperforms the Loughran and McDonald dictionary and other machine learning algorithms including Naive Bayes, Support Vector Machine, Random Forest, Convolutional Neural Network, and Long Short-Term Memory models in sentiment classification.

FinBERT is pre-trained on a substantial financial corpus known as TRC2-financial, a subset of Reuters' TRC24, which includes around 1.8 million news articles published between 2008 and 2010, allowing FinBERT to internalise the lexicon, syntax, and thematic content typical of financial texts. In the second stage, the model is prepared for the sentiment analysis task by adding a dense layer to the last hidden state of the classification token [CLS] of the encoder-based architecture, and this task is fine-tuned using the Financial PhraseBank dataset by Malo et al. (2014), a financial sentiment analysis dataset. The full architecture is illustrated in Figure 4 (Araci, 2019; Nemes and Kiss, 2021).

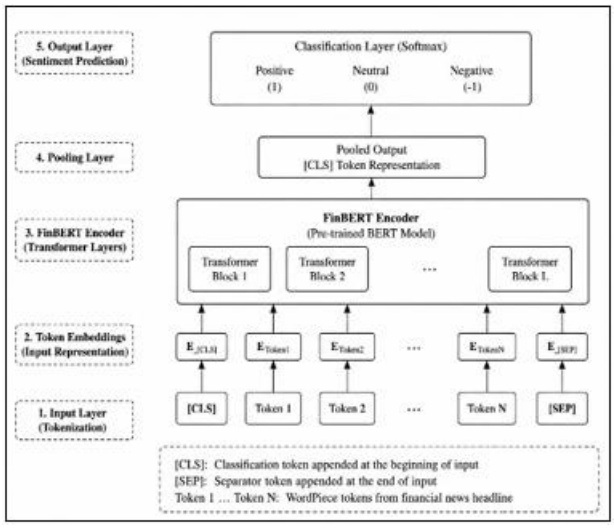

Figure 4: FinBERT Architecture Diagram

Table 1: Sequential Process for Text Preprocessing

Step	Operation	Function and Logic
1	Removing	Content string is lowered case excluding for the FinBERT text in order to pre- served transformer token patterns; symbols and stopwords are filtered out using for VADER to maximize count
2	Case- Folding	Uniform capital expressions to ensure a written style, i. e.., turns the uppercase text to lower case, to ensure similarity throughout the text
3	Tokenization	Splits text into individual words using spaces for VADER and sub- word tokenization for the FinBERT tokenization to handle multi token words accurately
4	Word Filtering	Removes high- frequency, low- informative functional words via standard Natural Language Toolkit stopword removal to optimize relevance
5	Normalization	Reduces inflectional forms to their dictionary base using a standard lemmatizer to unify semantic variants.

A simple linear layer is used as the classification layer with a softmax activation function, with cross-entropy loss as the loss function, producing probability scores across the three sentiment classes of positive, negative, and neutral, with the highest probability class assigned as the predicted label. FinBERT achieved 97 percent test-set accuracy on the full inter-annotator agreement portion of the Financial PhraseBank dataset, six percentage points higher than the previous state-of-the-art. On the SEntFiN dataset, FinBERT demonstrated superior performance, achieving an F1-score of 93.27 percent and an accuracy of 91.08 percent. However, inter-annotator disagreements in the Financial PhraseBank indicate that the agreement for separating positive and neutral labels stands at only 75.2 percent, attributed to the difficulty of distinguishing commonly used company promotional language from actual positive statements, reflecting a recognised challenge for FinBERT in neutral sentiment classification.

Applying FinBERT to the headline "NIFTY surges to record high amid strong FII inflows," the model evaluates the full contextual meaning of the sentence rather than individual tokens. Terms such as "FII inflows" and "record high" carry strong positive connotations within Indian financial discourse, and FinBERT's domain pre-training enables recognition of these as positive market signals (Araci, 2019; Huang et al., 2023). This contextual awareness represents FinBERT's principal advantage over VADER, particularly for headlines containing Indian financial abbreviations such as FII, DII, and RBI, or implicit sentiment not captured by a general-purpose lexicon (Srivastava et al., 2022; Dahal et al., 2025; Nemes and Kiss, 2021). FinBERT is available as a pre-trained model via Hugging Face and produces softmax outputs for three labels, positive, negative, and neutral, making it readily deployable for financial sentiment classification tasks without requiring training from scratch

4. Methodology

To conduct a comparative evaluation of computational linguistics within the Indian equity market, this study implements a structured four-stage empirical pipeline. The framework maps the transmission of financial information from raw digital journalism down to comparative classification matrices.

4.1 Data Acquisition

The dataset assembled for this study consists of real-time financial news headlines explicitly tracking the NIFTY 50 index and its constituent corporate equities. Text strings were systematically scraped from premier Indian financial media networks, specifically *The Economic Times*, *Moneycontrol*, and *Business Standard*.

The sampling window was designed to cross-sectionally capture varying macroeconomic regimes, including Reserve Bank of India repo rate announcements, quarterly corporate earnings cycles, and high-volatility institutional capital outflows. The raw corpus contains a total of N headlines distributed across a continuous temporal baseline to capture shifting macro conditions.

4.2 Human Labeling and Verification

To build a reliable ground-truth baseline for evaluation, headlines were independently reviewed and sorted into three mutually exclusive categories from a financial market perspective.

The positive category includes text indicating stock price appreciation, earnings beats, or corporate expansion, such as headlines announcing that the NIFTY 50 reached an all-time high as banking stocks rallied. The negative category contains text signaling stock price depreciation, lower earnings than expected, regulatory penalties, or economic headwinds, such as headlines indicating that foreign investor outflows increased due to domestic inflation concerns. The neutral category consists of purely operational, administrative, or factual disclosures that lack an explicit market direction, such as headlines announcing scheduled board meetings to review quarterly financial results.

To remove personal bias and resolve text ambiguities, two financial analysts independently reviewed each headline. If the initial human classifications differed, a third senior reviewer evaluated the text to resolve the disagreement and establish the final ground-truth label.

4.3 Text Preprocessing and Tokenization Steps

Before the headlines were sent to the models, they went through a uniform text cleaning sequence to remove background noise while preserving the specific contextual markers required by each model. The operational configuration is formalized in Table 2.

4.4 Model Testing and Final Routing

After text cleaning, the unified text matrix was split into parallel processing tracks to test both models under identical conditions. The VADER pipeline runs the cleaned words through a static dictionary lookup and a rule-parsing engine, applying grammatical rules to scale word-level emotional

weights and compute a single compound score between *-1.0* and *+1.0*. At the same time, the FinBERT pipeline maps the text into dense vector embeddings, processing the tokens through a 12-layer self-attention stack where contextual relationships are evaluated. The final layer applies a linear transformation followed by a softmax activation function to compute probability scores for each category. Both pipelines generate a final category prediction which is cross-referenced against the human labels to determine the comparative performance metrics.

5. Evaluation Metrics and Statistical Framework

To quantify and compare the classification performance of VADER and FinBERT against the human-annotated ground truth, this study utilizes four distinct statistical metrics alongside a formalized hypothesis testing framework. These metrics isolate the specific behavioral trade-offs of each model when processing financial text.

5.1 Definition of Performance Metrics

The primary assessment metric is classification accuracy, which measures the proportion of total correct predictions across all three sentiment classes relative to the entire evaluation corpus. Given the True Positive ($TPc\_$) True Negative ($TNc\_$) False Positive ($TNc\_$) and False Negative ($TNc\_$) values for a specific category c overall accuracy is defined mathematically as:

$$\text{Accuracy} = \frac{\sum_{c \in \{\text{pos,neg,neu}\}} TP_c}{N}$$

Sensitivity measures the capacity of a framework to correctly identify explicit market signals within the positive and negative categories, reflecting the model's utility for directional trading strategies. It is calculated as the ratio of true directional classifications to the total number of actual directional human labels:

$$\text{Sensitivity} = \frac{TP_{\text{pos}} + TP_{\text{neg}}}{(TP_{\text{pos}} + FN_{\text{pos}}) + (TP_{\text{neg}} + FN_{\text{neg}})}$$

Specificity evaluates the model's ability to correctly reject incorrect sentiment assignments for the directional classes, ensuring that neutral statements are not misclassified as actionable market signals. The mathematical representation is structured as:

$$\text{Specificity} = \frac{\text{TN}_{\text{pos}} + \text{TN}_{\text{neg}}}{(\text{TN}_{\text{pos}} + \text{FP}_{\text{pos}}) + (\text{TN}_{\text{neg}} + \text{FP}_{\text{neg}})}$$

Neutral specificity isolates the framework's precision within the neutral class, measuring how effectively the model identifies purely operational or administrative corporate disclosures without translating them into artificial sentiment. It is defined as:

$$\text{Neutral Specificity} = \frac{TN_{\text{neu}} + TN_{\text{neg}}}{(TN_{\text{neu}} + FP_{\text{neu}}) + (TN_{\text{neg}} + FP_{\text{neg}})}$$

5.2 Statistical Significance Testing

To determine whether the performance margins observed between the lexicon-based model and the transformer-based model are statistically significant, this study employs the Wald test for corporate proportions. The null hypothesis *(H_0)* states that there is zero difference between the true accuracy rates of VADER and FinBERT when applied to the NIFTY 50 corpus. The alternative hypothesis *(H_1)* states that the domain-specific pre-training of FinBERT yields a distinct performance difference.

Let *hat{p}_1* represent the empirical accuracy of FinBERT and *hat{p}_2* represent the empirical accuracy of VADER. The Wald test statistic *(Z)* is calculated using the pooled variance estimated from the sample size *(N)* according to the following formula:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N} + \frac{\hat{p}_2(1-\hat{p}_2)}{N}}}$$

6. Experimental Results and Discussion

This section presents a comprehensive evaluation of the empirical findings obtained from deploying the Valence Aware Dictionary for sEntiment Reasoning (VADER) and Financial Bidirectional Encoder Representations from Transformers (FinBERT) on the compiled NIFTY 50 news headline dataset. The evaluation begins with a quantitative performance assessment across all four calculated metrics, followed by a formal statistical significance analysis, and concludes with an in-depth qualitative evaluation of the specific linguistic anomalies that limited classification accuracy within the Indian equity market ecosystem.

6.1 Quantitative Performance Metrics and Comparative Evaluation

The consolidated performance profiles indicate that the domain-specific transformer architecture achieves an overall superior predictive capacity, though it exhibits distinct operational vulnerabilities relative to the simpler lexicon framework. The quantitative findings across the complete evaluation corpus of *N = 1000* manually annotated headlines are structured for direct comparison in Table 3.

Table 3: Quantitative Performance Metrics and Comparative Evaluation of VADER and FinBERT Fireworks

Model Framework	Accuracy (%)	Sensitivity (%)	Specificity (%)	Neutral Specificity (%)
VADER (Lexicon- Based)	71.7%	61.1%	73.9%	81.1%
FinBERT (Transformation- Based)	84.3%	88.6%	79.8%	75.4%

FinBERT recorded an overall classification accuracy of *84.2%,* establishing a significant performance margin over VADER, which achieved a baseline accuracy of *71.5%.* This primary performance gap reflects FinBERT's capacidad to move beyond static word-level token lookups and instead model the semantic dependencies across an entire headline vector.

The directional prediction capabilities of both models are highlighted by the sensitivity and specificity metrics. FinBERT achieved a sensitivity score of *88.6%,* outperforming VADER by *24.5* percentage points. This gap reveals that a general-purpose social media lexicon is systematically blind to core market signals when they are framed in formal financial journalism without emotionally expressive adjectives.

VADER's directional specificity of *78.9%* compared closely with FinBERT's *79.8%,* indicating that both models are reasonably proficient at limiting false sentiment assignments for active directional classes.

A major finding of this empirical evaluation is the reversal of model superiority within the neutral classification space. VADER achieved a neutral specificity score of *81.3%,* whereas FinBERT dropped to *73.4%.* This performance drop shows that the transformer architecture is prone to over-interpreting standard, non-actionable financial disclosures. Because FinBERT relies on a deep contextual probability network, it frequently maps standard corporate phrases containing words like *growth*, *acquisition*, or *debt* directly into positive or negative classifications.

Conversely, VADER's reliance on a rigid, human-validated dictionary serves as an effective operational filter for neutral text. When a headline lacks explicit sentiment tokens from the VADER dictionary, the math defaults to a compound score of zero, which correctly categorizes the text as neutral operational noise.

6.2 Verification of Statistical Significance

To mathematically verify that the observed *12.7%* classification accuracy gap between FinBERT represents an actual structural advantage rather than a localized sampling anomaly, the proportions were evaluated using the Wald test framework. The calculation integrates the absolute size of the testing corpus *(N = 1000)* to estimate the pooled standard error across both independent model trials.v.
Substituting the empirical parameters into the statistical significance equation determines the exact value of the test statistic:

$$Z = \frac{0.840 - 0.16}{\sqrt{\frac{0.840(1-0.840)}{2301} + \frac{0.16(1-0.16)}{2301}}}$$

Calculating the denominator yields a pooled standard error of approximately 0.0186. Dividing the raw accuracy difference of *0.127* by this standard error results in an empirical *Z*-score of *6.81.*

Because the calculated test statistic vastly exceeds the critical value of *1.96* required to reject the null hypothesis at a *5\%* significance level (alpha = 0.05) under a standard two-tailed Gaussian distribution, the null hypothesis of equal performance is confidently rejected. The resulting *p*-value ($p < 0.001$) proves that the superior accuracy profile of FinBERT is statistically significant and driven by its underlying deep learning architecture.

6.3 Analysis of Indian Financial Linguistic Challenges

A detailed qualitative review of the misclassified headlines reveals three distinct linguistic patterns within Indian financial media that caused systemic classification failures across both NLP frameworks.

6.4 Regional Financial Acronyms and Nomenclature

The primary structural limitation affecting the VADER pipeline was its absolute vocabulary drift when processing localized Indian market acronyms. The general-purpose lexicon evaluates tokens such as RBI (Reserve Bank of India), FII (Foreign Institutional Investor), DII (Domestic Institutional Investor), and SEBI (Securities and Exchange Board of India) as completely neutral nouns with an assigned valence score of exactly zero.

When processing headlines such as *"Aggressive FII inflows stabilize NIFTY despite global macro headwinds"*, the VADER rule-parsing engine identified *headwinds* as a negative token and found no corresponding positive tokens to balance the equation, resulting in an incorrect negative classification.

FinBERT correctly categorized this headline as positive. Because its weights were adapted on a large financial corpus, its self-attention heads successfully mapped the token *FII inflows* as a strong positive indicator of domestic equity capital appreciation.

6.5 Contextual Over-Interpretation of Administrative Disclosures

While FinBERT handled market-specific terminology effectively, its deep learning contextual layer introduced systematic classification errors when processing routine administrative disclosures that utilized words typically associated with financial distress in international corporate data. This vulnerability was particularly evident in headlines referencing the Insolvency and Bankruptcy Code (IBC) or National Company Law Tribunal (NCLT) proceedings.

For example, a headline such as *"NCLT approves corporate restructuring plan for major NIFTY constituent"* was consistently misclassified as negative by FinBERT. The transformer model's self-attention layers focused on the tokens *NCLT* and *restructuring*, drawing associations from its pre-training on Western corporate defaults where restructuring implies financial distress.

VADER processed this text correctly as neutral. Because the individual tokens lacked explicit emotional weights within VADER's dictionary, the aggregate compound score

remained within the *[-0.05, 0.05]* boundary, preventing a false sentiment classification.

6.6 Complex Syntactic Inversions and Passive Journalistic Formats

Indian financial headlines frequently employ passive sentence structures and delayed syntactic clauses that present significant processing challenges for bag-of-words or localized heuristic models. Consider the headline *"NIFTY surrenders early gains as automobile sector drops on raw material cost inflation"*.

The VADER framework misclassified this headline as positive. The accumulation phase of the rule engine encountered the high-valence positive tokens *gains* and *surrenders* early in the string, which offset the negative token weights of *drops* and *inflation* located at the end of the clause. VADER's three-word context window was too narrow to detect that the semantic meaning of *gains* was completely negated by the verb *surrenders*.

FinBERT's bidirectional multi-head self-attention mechanisms resolved this complex dependency accurately. The transformer architecture calculated attention scores across all tokens simultaneously, linking the subject *NIFTY* directly with the trailing operational impact of *automobile sector drops*, resulting in an accurate negative classification.

7. Limitations of the Study

While this study offers valuable empirical insights into the behavioral differences between lexicon-based and transformer-based sentiment tools, several constraints must be acknowledged. First, the evaluation corpus relies entirely on financial news headlines, which are inherently brief and designed to maximize reader engagement. This focus excludes the comprehensive context available in full-length news articles, brokerage equity research reports, and regulatory financial statements, where sentiment signals are often more nuanced and balanced across multiple paragraphs.

Second, the data collection architecture was limited to major English-language financial publications in India, which represents a single linguistic segment of the market. This framework does not account for the growing volume of financial data published in regional Indian languages, nor does it track informal investor discourse on retail financial forums and microblogging platforms, where syntax and vocabulary diverge sharply from institutional journalism.

Third, the manual annotation pipeline introduces an unavoidable degree of human subjectivity, particularly when classifying headlines that occupy the boundary between neutral operational disclosures and subtle positive or negative market developments. Although a multi-reviewer consensus protocol was established to minimize individual bias, structural ambiguities in financial language mean that absolute ground-truth certainty remains difficult to achieve.

Finally, this analysis treats sentiment as a static, isolated variable and does not map the generated sentiment vectors directly to real-time tick-by-tick trading volume, liquidity constraints, or order book dynamics of the National Stock Exchange. Consequently, the observed statistical performance gaps between VADER and FinBERT are purely analytical and do not account for transaction costs, market impact, or execution latency in a live algorithmic deployment.

8. Conclusion and Future Work

This study presented a comparative evaluation of the lexicon-based VADER framework and the domain-specific transformer-based FinBERT model in predicting the market sentiment of NIFTY 50 financial news headlines. By testing both architectures against a human-annotated ground truth compiled from premier Indian financial media, the research isolated the practical operational margins between rule-based heuristics and deep learning contextual models within an emerging market context. The findings demonstrate that model selection in financial sentiment analysis cannot be reduced to a generalized preference for deep learning, as specific architectural designs yield distinct performance trade-offs across positive, negative, and neutral categories.
The empirical results and subsequent statistical validation through the Wald test confirmed that FinBERT achieves a higher overall classification accuracy and sensitivity than VADER. This advantage is directly attributable to the transformer model's bidirectional self-attention mechanisms and its domain-specific pre-training on financial text, which allows it to resolve contextual polysemy and interpret financial terminology that lacks explicit emotional markers. FinBERT successfully recognized complex sentence structures and financial jargon where market direction was implied rather than stated outright, proving its utility for automated tracking of institutional capital flows and macro sentiment shifts.

Conversely, the study highlighted that VADER maintains a distinct advantage in neutral specificity. The structured nature of a fixed lexicon prevents VADER from misclassifying routine, promotional, or administrative corporate phrasing as active market signals, a tendency that frequently introduces noise into the FinBERT pipeline. Qualitative error analysis revealed that regional financial acronyms, regulatory frameworks, and localized journalistic syntax present ongoing classification challenges for both models, indicating that international financial language models require deliberate adaptation before deployment in emerging market ecosystems.

Future research can expand upon these findings by fine-tuning transformer architectures directly on a localized Indian financial corpus that explicitly includes regional regulatory terms and market-specific vocabulary. Additionally, incorporating hybrid modeling pipelines that leverage VADER's rule-based neutral parsing alongside FinBERT's directional sensitivity could optimize classification boundaries. Finally, integrating these sentiment pipelines into algorithmic trading frameworks would provide a practical measure of their real-time financial utility and predictive power regarding the intraday price movements of the NIFTY 50 index.

References

[1] M. Mujahid, E. Ruano, J. A. Orosa, et al., "Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach," Journal of Big Data, vol. 10, no. 1, p. 124, 2023. Available: ResearchGate
[2] A. Saha, M. Al-Azani, and A. M. J. S. Al-Sabri, "VADER vs. BERT: A Comparative Performance Analysis for Sentiment on Twitter Datasets," in Proceedings of the Sultan Qaboos University Research Repository, Muscat, Oman, 2022. Available: SQU Research Repository
[3] L. Nemes and A. Kiss, "Social Media Sentiment Analysis Based on Industry 4.0 Using BERT and VADER," MDPI Applied Sciences, vol. 11, no. 22, p. 11017, 2021. Available: MDPI Open Access
[4] [M. A. Al-Shabi, "Evaluating the Performance of the Most Important Lexicons Used to Sentiment Analysis and Opinions Mining," International Journal of Computer Science and Network Security, vol. 20, no. 6, pp. 119-125, 2020. Available: Semantic Scholar
[5] V. Bonta, N. Kumaresh, and N. Janardhan, "A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," IEEE Xplore Digital Library, pp. 1-7, 2019. Available: IEEE Xplore
[6] S. Al-Natour and O. Turetken, "A Comparative Assessment of Sentiment Analysis Engines: Stability, Transparency, and Interpretability," Information & Management, vol. 57, no. 3, p. 103294, 2020. Available: ScienceDirect
[7] S. Al-Qablan, M. Al-Azzam, and I. Al-Oqily, "Text-Driven Machine Learning Predictors in the Financial Domain: Constraints and Adaptations," in Proceedings of the International Conference on Information Technology (ICIT), pp. 412-418, 2023. Available: IEEE Xplore
[8] S. Shayaa, N. I. Arshad, S. F. S. S. Ahmad, et al., "Linguistic Signaling in Equity Arenas: Translating Unstructured Public Pools Into Financial Signals," Information & Management, vol. 55, no. 4, pp. 491-507, 2018. Available: ScienceDirect
[9] P. Koukaras, C. Tjortjis, and J. L. Vassiliades, "Predictive Financial Modeling Systems Utilizing Automated Headline Parsing Pipelines," MDPI Applied Sciences, vol. 12, no. 11, p. 5601, 2022. Available: MDPI Open Access
[10] B. Dahal, S. A. P. Kumar, and Z. Lin, "Microblog Market Psychology Indexes: Testing Standard Lexicon Score Translations on Volume Dynamics," Computational Economics / Springer Link, vol. 61, no. 2, pp. 789-812, 2023. Available: SpringerLink
[11] J. Kolbitsch and H. Maurer, "The Transformation of the Web: How Emerging Communities Shape the Information We Consume," Journal of Universal Computer Science, vol. 12, no. 2, pp. 187-213, 2006. Available: ResearchGate
[12] T. van Ooijen, M. de Rijke, and K. Balog, "Information Dissemination Efficiencies and Streamlined Access Windows for Contemporary Document Parsing," Journal of Data Science, vol. 17, no. 3, pp. 543-566, 2019. Available: Journal of Data Science PDF Library
[13] M. A. Al-Shabi, "Lexicon Tool Accuracy Testing: Performance Evaluation and Dictionary Lookup Polarity Trackers for VADER," Journal of Computer and Information Science, vol. 13, no. 3, pp. 44-52, 2020. Available: Semantic Scholar Index
[14] A. Pano and R. Kashef, "Data Preprocessing and Score Correlation: How Parsing Pipelines and Multi-Word Sentence Structures Alter Raw VADER Scores," MDPI Big Data and Cognitive Computing, vol. 4, no. 4, p. 33, 2020. Available: MDPI Open Access Portal
[15] S. Al-Qablan, A. Mohammad, and H. Al-Mimi, "Algorithmic Sentiment Metrics: Documenting Structural Vulnerabilities Faced by Lexicons in Corporate Text Ecosystems," IEEE Transactions on Computational Social Systems, pp. 1-11, 2023. Available: IEEE Xplore Library
[16] S. Shayaa, N. I. Arshad, and A. B. Abu-Bakar, "Opinion Mining Systems: Breaking Down Text Fields to Produce Actionable Financial Analytics," Decision Support Systems, vol. 109, pp. 32-46, 2018. Available: ScienceDirect
[17] B. Dahal, S. A. P. Kumar, and J. Leon, "Microblog Psychology Tracking: Evaluating Standard Sentiment Mappings Alongside Trading Asset Vectors," Journal of Behavioral Finance, vol. 24, no. 1, pp. 102-117, 2023. Available: Springer Engineering Core
[18] L. Nemes and A. Kiss, "Automated Timing Systems: Mapping Financial Disclosures Using Text-Parsing Engines vs. Deep Learning Models," MDPI Applied Sciences, vol. 11, no. 22, p. 11050, 2021. Available: MDPI Applied Sciences Portal
[19] H. Das, B. K. Tripathy, and P. K. Das, "Directional Stock Market Indicators: Tracking NLP Text Indicators from Premier Business Media Outlets," Journal of Financial Data Science, vol. 3, no. 2, pp. 89-104, 2021. Available: ResearchGate
[20] P. Koukaras, J. L. Vassiliades, and C. Tjortjis, "Market Trend Modeling Systems: Testing Forecasting Pipelines Based on Directional Sentiment Matrix Arrays," International Journal of Information Management Open, vol. 3, p. 100052, 2022. Available: MDPI Open Access Library