ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences & Management Research

STOCK MARKET PREDICTION VIA SENTIMENT ANALYSIS OF FINANCIAL AND ECONOMIC NEWS USING MACHINE LEARNING

Paril Ghori
parilghori@gmail.com

ABSTRACT

This study investigates the use of sentiment analysis derived from financial news articles to predict stock market returns. By incorporating sentiment features along with traditional stock market predictors, we aim to enhance prediction accuracy for stock price movements. Various machine learning models, including Linear Regression, Decision Trees, Random Forests, Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost), were evaluated for their performance in predicting stock returns. The models were assessed using several regression metrics, such as RMSE, MAE, MAPE, and R^2, with a particular focus on how well sentiment-based features, like SentNews and SentNewsIBOV, contribute to predicting market movements. The study also highlights the importance of identifying the most influential words within financial news articles using feature importance analysis, particularly through the XGBoost model. The results indicate that sentiment data, especially when combined with advanced machine learning algorithms like XGBoost, provides a valuable tool for improving the accuracy of stock return forecasts. The findings suggest that news sentiment, which reflects public perception and market sentiment, significantly influences stock price movements and can be leveraged to forecast future market trends.

KEYWORDS –Decision Trees, Extreme Gradient Boosting (XGBoost), Gradient Boosting Machine (GBM), IBOVESPA, Linear Regression, Machine Learning, Random Forests, Regression Models, Sentiment Analysis, SentNews, SentNewsIBOV, XGBoost.

1. INTRODUCTION

In recent years, forecasting macroeconomic variables has become increasingly crucial in understanding market trends and guiding investment decisions. Economic forecasts provide valuable insights for business managers, allowing them to strategize effectively and anticipate the potential impact of macroeconomic fluctuations on business performance. The ability to predict economic outcomes has grown even more important with the availability of new data sources and advanced analytical techniques [1]. The vast amounts of data generated through platforms like social media, consumer transactions, and public reports are often referred to as "Big Data," which includes unstructured data that lacks a predefined model but can be harnessed for insightful analysis [3].

Big Data has become a game-changer, enabling predictive analytics through pattern recognition, trend analysis, and probabilistic modeling. The ability to sift through massive datasets allows businesses to extract actionable insights and improve decision-making processes. In recent times, machine learning techniques have been extensively employed to analyze such data. Machine learning, a subset of artificial intelligence (AI), empowers machines to learn from data and automate analytical tasks, making it a vital tool for processing Big Data [3]. The growing sophistication of machine learning models enables more accurate predictions by identifying hidden relationships within large datasets.

The integration of Big Data with machine learning holds immense potential, as the insights drawn from vast data can aid in forecasting economic trends and market performance. This combination is particularly relevant in the context of sentiment analysis, where the tone of news articles and social media posts is assessed to gauge market sentiment. Several studies have demonstrated that public sentiment—extracted from news articles—can have a direct impact on stock market returns and volatility [4] [5] [6]. Negative news or unfavorable sentiment can cause stock prices to drop, while positive sentiment can lead to price surges. These patterns are observed in both global and domestic markets, including India [7] [8].

Stock market prices are influenced by a multitude of factors, such as corporate earnings reports, economic policy changes, and political events. In India, news about government decisions on economic policies, financial results of key companies, or global market trends can significantly affect investor behavior. For instance, when the Indian government announces economic reforms or changes in interest rates, it can lead to immediate fluctuations in the stock market, affecting the value of major indices like the Nifty 50 and Sensex. Similarly, corporate earnings announcements, mergers, or regulatory changes in sectors like technology or banking can also impact investor sentiment and stock valuations.

Research by scholars [9]-[10] has highlighted the complexity of predicting stock market movements, showing that these predictions are influenced by a variety of macroeconomic and socio-political factors. Furthermore, the increasing relevance of behavioral economics has brought attention to the role of news sentiment in shaping

investor expectations and stock market performance. As the global economy becomes more interconnected, the Indian stock market is also susceptible to external news, such as global economic crises or geopolitical tensions, which can have significant repercussions on investor behavior.

In this context, this study aims to investigate the role of news sentiment in influencing the behavior of investors and the performance of the Indian stock market. Using machine learning techniques, the research will analyze how positive or negative news coverage on economic policies, corporate earnings, and political developments affects the stock market returns of Indian indices like the Sensex and Nifty 50. The findings from this study could provide valuable insights for investors and analysts, helping them to better understand how public sentiment reflected in news media can drive stock price movements.

Recent studies in the field of financial analysis have shown that news sentiment plays a crucial role in shaping market expectations. For instance, news articles with a positive tone about a company or economic policy may drive investor optimism, leading to higher stock prices, while negative news can lead to market declines [11] [12]. In India, this relationship is particularly relevant, as the stock market is highly sensitive to changes in investor sentiment, especially during times of political uncertainty or economic stress [13]. The intensity of the sentiment also matters; more strongly worded news articles tend to have a stronger influence on market behavior.

This study investigates the relationship between news sentiment from economic and political news sources and stock market returns. Specifically, it examines how sentiment in news articles — whether positive or negative — affects the performance of stock indices. The expectation is that negative news will be linked to declines in stock indices, while positive sentiment will correlate with market growth. Furthermore, stronger sentiment (both positive and negative) is likely to have a more significant impact on stock market performance. The findings from this research will contribute to the growing body of literature on the role of news sentiment in influencing financial markets and offer practical insights for investors seeking to better understand how media coverage shapes market behavior.

The paper is organized as follows: it starts with an introduction to the research problem, justification, and objectives, followed by a review of the relevant literature and theoretical framework. The methodology section details the data sources and analytical techniques used, while the results section presents the key findings of the analysis. Finally, the paper concludes with a discussion of the implications of the findings and recommendations for future research.

2. THEORETICAL FRAMEWORK

2.1 Efficient Market Hypothesis (EMH)

The Efficient Market Hypothesis (EMH) [14], plays a crucial role in financial theory, particularly in understanding how market prices reflect available information. According to EMH, stock prices fully incorporate all publicly available information, making it impossible to consistently outperform the market by using this information. The authors of [15] proposed three levels of market efficiency: weak, semi-strong, and strong.

- Weak Efficiency: Prices reflect all historical public information.
- Semi-strong Efficiency: Prices adjust rapidly to new public information.
- Strong Efficiency: Prices incorporate all information, including private or insider knowledge.

This concept has been foundational in financial modeling and investment strategies. However, behavioral economists, such as [16], challenge the notion of complete rationality in investor decision-making. They argue that investor behavior is often influenced by emotions and cognitive biases, which can lead to irrational market movements and affect stock returns. Such psychological and emotional factors highlight the limitations of EMH in explaining market anomalies.

In the Indian stock market, investor sentiment, often shaped by media reports and news, plays a significant role in influencing market trends. The rise of social media platforms has further amplified the impact of public sentiment on stock prices, making it a key area of interest for both investors and researchers.

2.2 Stock Return Forecasting

Forecasting stock returns has been an active area of research, with studies highlighting the use of textual data in improving prediction models. According to [17], financial news is a valuable source for predicting market behavior. Incorporating sentiment analysis and supervised machine learning techniques, such as decision trees and artificial neural networks, has proven effective in analyzing textual data for forecasting stock returns.

In the Indian market, recent studies have explored the use of social media platforms and news articles to predict stock returns. For example, sentiment analysis of Twitter feeds related to Indian companies or market events can provide valuable insights into investor sentiment, which can be used to predict market movements. Similar to the

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

work done in [18], applying sentiment analysis in India has shown that including textual data can significantly improve the accuracy of stock return predictions.

2.3 Textual Sentiment

The concept of textual sentiment analysis has gained popularity as an innovative tool in forecasting market trends. The authors of [1] introduced the textual sentiment index, which applies linguistic techniques to assess the emotional tone of a text. By analyzing the words and phrases in news articles or social media posts, this approach captures the sentiment—whether positive, negative, or neutral—of the content, adding an emotional dimension to financial analysis. This method of using adaptable dictionaries, updated over time using machine learning, allows for the identification of the most predictive terms within a given text. These selected words are then used as predictors to improve the accuracy of market forecasts. In India, financial news articles, especially those covering key government policies or corporate earnings, can serve as strong indicators of market sentiment. Positive sentiment about economic reforms or successful company earnings often correlates with stock price increases, while negative news can trigger market declines.

2.4 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF is a statistical technique used in text mining to evaluate the importance of a word within a document or a corpus. It combines two metrics [19]:

- Term Frequency (TF): Measures how often a word appears in a document, with higher frequency indicating greater relevance within that document.
- Inverse Document Frequency (IDF): Reduces the weight of words that appear frequently across many documents, emphasizing words that are rare and therefore potentially more informative.

TF-IDF is widely used in the context of sentiment analysis to assess the relevance of terms in financial news articles. In the Indian stock market, this technique can be applied to news articles, reports, and social media posts to identify key terms that are most likely to influence stock market movements. By focusing on rare but significant words, analysts can better capture the sentiment that drives market performance.

2.5 Web Scraping

Web scraping is an essential tool for collecting and analyzing large volumes of online data. Unlike traditional search engines, web scrapers can gather data from a wide range of websites, including news articles, blogs, and social media platforms. This capability makes web scraping an invaluable tool for extracting real-time data for sentiment analysis.

The authors of [20] defines web scraping as the process of automatically collecting data from web pages through programs that interact with web servers. This data can then be used in machine learning models to predict market behavior. In the Indian context, web scraping can be applied to gather data from financial news websites, social media platforms like Twitter, or even investor forums. This collected data, when combined with sentiment analysis tools like TF-IDF, can provide a rich dataset for predicting stock market trends.

Web scraping also provides an advantage over traditional data collection methods by offering access to a broader range of online content. For example, it can gather real-time data on stock market news, government policies, and company updates, all of which can be used to assess sentiment and forecast stock returns. This process is essential for building comprehensive and accurate predictive models in the rapidly changing Indian financial landscape.

3. PROPOSED METHODOLOGY

This study aims to predict stock returns by analyzing the influence of sentiment derived from financial news articles using advanced machine learning techniques. The focus is on utilizing textual data from relevant news portals, particularly covering economic and political events, and leveraging machine learning models to understand the relationship between news sentiment and stock market behavior. The methodology consists of several key steps: data collection, pre-processing, feature extraction, sentiment estimation, model building, training, validation, and performance evaluation. Each step in the Figure 1 involves rigorous techniques and mathematical formulations to ensure robust prediction accuracy and comprehensive analysis.

3.1 Data Collection

The data collection process in this study involves gathering financial and economic news articles from a variety of online news portals. We specifically focus on articles that discuss political and economic developments, which are relevant to market sentiment and stock market performance. To collect the data, we use web scraping

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

techniques, leveraging Python libraries such as BeautifulSoup and Requests. These tools are employed to extract structured content from HTML pages of selected news articles. The collection period spans from December 2020 to May 2022. The news articles extracted are primarily from portals that cover key political and economic issues, which have a direct impact on market performance. The focus is on identifying and analyzing sentiments from these news articles, with particular emphasis on understanding how economic and political news sentiment can correlate with stock market movements. This data forms the foundation for building sentiment indices like SentNews and SentNewsIBOV, which are then used to predict stock returns.

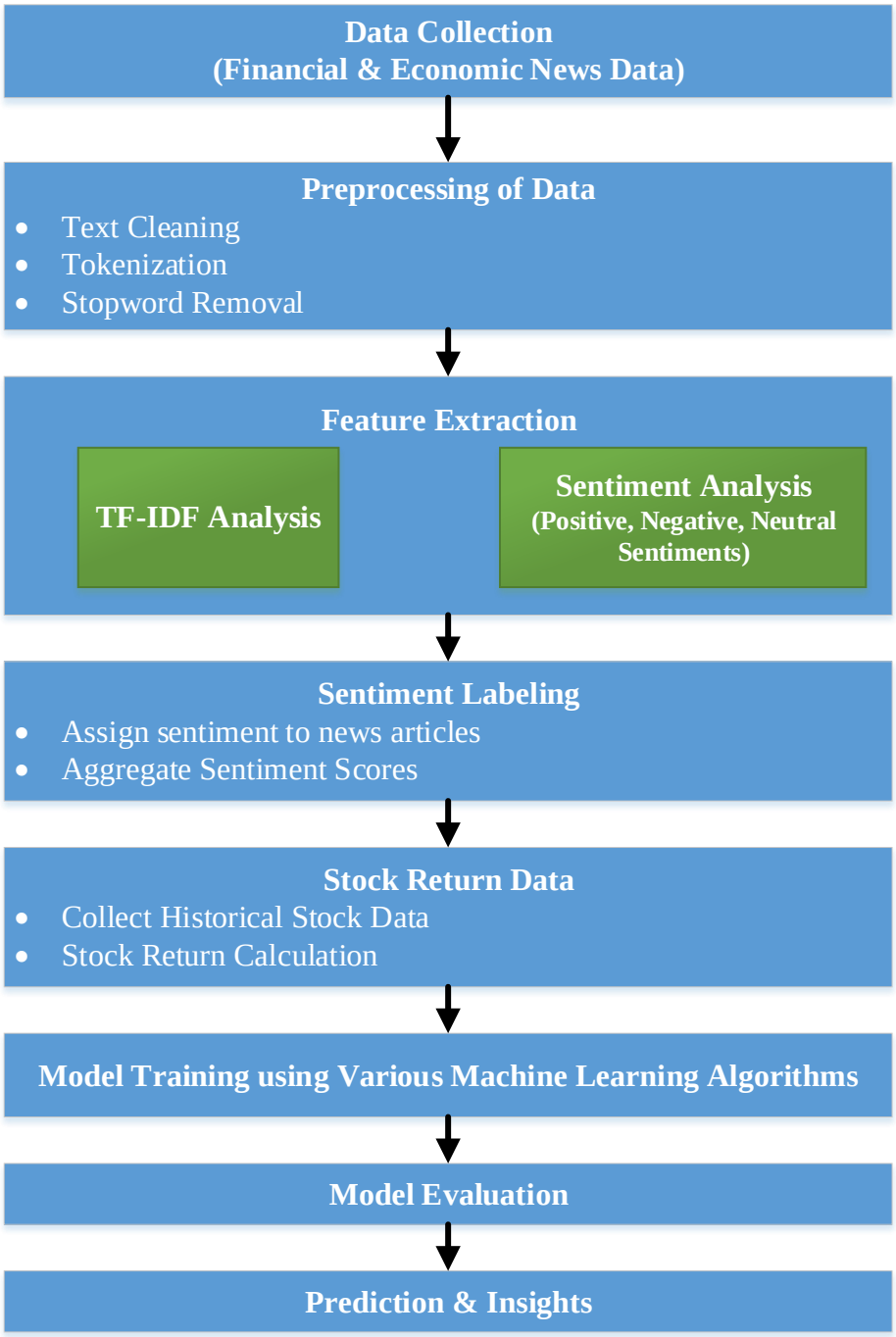

Figure 1: Flow diagram for proposed approach

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

Mathematical formulation for data extraction:
Let X_i represent the i^{th} news article where $i \in [1, N]$, and N is the total number of news articles collected. Each article consists of a set of words and corresponding publication dates, i.e., $X_i = \{w_1, w_2, \ldots, w_k, \text{date}_i\}$, where w_j is a word in the article and k is the total number of words in article X_i.

3.2 Pre-processing of Textual Data

Once the raw textual data is collected, the next step involves cleaning and pre-processing the text to remove irrelevant content. This process is critical as it reduces noise in the data, making the sentiment analysis more effective. We use the R tidytext package for text processing, where common steps include:

- Tokenization: Splitting the text into individual words or tokens.
- Stopword Removal: Removing common words (such as "the", "and", "in") that do not contribute meaningful information.
- Lemmatization/Stemming: Reducing words to their root forms (e.g., "running" to "run").
- Removing Punctuation and Special Characters: Eliminating symbols that do not contribute to the sentiment or meaning of the text.

The data is cleaned in such a way that only the most significant words, which are important for sentiment analysis and stock prediction, remain.
Let T_i represent the processed text of article ii. The pre-processing steps result in a cleaner version of the text:

$$T_i = \text{Tokenize}(X_i) \quad \text{after removing stopwords and stemming} \tag{1}$$

3.3 Feature Extraction

In this step, we extract relevant features from the pre-processed textual data. TF-IDF (Term Frequency-Inverse Document Frequency) is used to quantify the importance of each word in a document relative to the entire corpus. The formula for TF-IDF is:

$$\text{TF} - \text{IDF}(w, D) = \text{TF}(w, D) \times \text{IDF}(w) \tag{2}$$

Where:

$$\text{TF}(w, D) = \frac{\text{Count of word w in document D}}{\text{Total words in document D}} \tag{3}$$

$$\text{IDF}(w) = \log\left(\frac{N}{\text{Count of documents containing word w}}\right) \tag{4}$$

This calculation allows us to assign a weight to each word based on its significance in the document and its rarity across the corpus.
Additionally, we incorporate sentiment values into the feature extraction process. Using the sentiLex_PT02 dictionary (or a modified version adapted for Indian contexts), each word's polarity is assigned a sentiment score. Negative words receive a score of -1, and positive words receive a score of +1. We then multiply the TF-IDF score of each word by its polarity to get a sentiment-weighted score for each term.

$$\text{Sentiment–Weighted TF–IDF}(w) = \text{TF} - \text{IDF}(w, D) \times \text{Sentiment}(w) \tag{5}$$

3.4 Sentiment Measure Estimation

Once the sentiment-weighted features are extracted, we calculate two key sentiment measures to summarize the impact of sentiment on stock returns:

- **SentNews:** This is the average sentiment score for each word in a news article for a given day.

$$\text{SentNews}_t = \frac{1}{N}\sum_{i=1}^{N} \text{Sentiment–Weighted TF–IDF}(w_i) \tag{6}$$

  Where w_i represents each word in article t, and N is the total number of words in that article.
- **SentNewsIBOV:** This measure is based on using machine learning models, particularly XGBoost, to predict stock returns from the extracted features. XGBoost uses decision trees and gradient boosting to predict the relationship between the sentiment measures and stock market returns. The sentiment score

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

from the XGBoost model becomes the SentNewsIBOV score, which captures the relationship between the sentiment of news and the movement of the stock market.

$$\text{SentNewsIBOV}_t = \text{XGBoost}\big(\text{TF–IDF}(w), \text{Sentiment}(w)\big) \tag{7}$$

3.5 Building Machine Learning Models

After sentiment measures have been calculated, we build machine learning models to forecast stock returns. The historical average return, SentNews, and SentNewsIBOV are used as inputs to predict the stock return for the next day.

Let y_t represent the stock return on day t, and X_t represent the input features, which include the historical return, SentNews, and SentNewsIBOV. The goal is to model the relationship:

$$y_t = f(X_t) + \epsilon \tag{8}$$

Where $f(X_t)$ is the predictive function modeled by machine learning algorithms and ϵ is the error term.

We apply several machine learning algorithms which are described in the following subheadings:

3.5.1 Linear Regression

Linear Regression aims to model the relationship between the dependent variable y (stock return) and independent variables X (features such as sentiment measures or historical returns) through a linear equation.

In a univariate case (one predictor variable), the model takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{9}$$

Where:

- y_i is the predicted stock return for the i^{th} observation,
- x_i is the feature (e.g., sentiment measure or historical return) for the i^{th} observation,
- β_0 is the intercept (constant term),
- β_1 is the coefficient that measures the relationship between x_i and y_i,
- ϵ_i is the error term for the i^{th} observation.

For the multivariate case (more than one predictor variable):

$$y_i = \beta_0 \sum_{k=1}^{p} \beta_k x_{ik} + \epsilon_i \tag{10}$$

Where:

- y_i is the target variable (predicted return),
- x_{ik} are the input features for the i^{th} observation and k^{th} predictor,
- β_k represents the coefficient for the k^{th} feature,
- ϵ_i is the error term.

Cost Function (Mean Squared Error) to minimize:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{11}$$

Where:

- y_i is the actual return,
- $\hat{y}_i$ is the predicted return.

3.5.2 Decision Trees

Decision Trees are non-linear models used for regression tasks. They build a tree-like structure by recursively splitting the data based on the input features to minimize error in the prediction.

The decision tree model is built by recursively splitting the dataset based on features that best reduce the impurity (such as Mean Squared Error (MSE) or Mean Absolute Error (MAE)). For each split, a node is created that makes a decision based on the value of a feature, and each leaf node represents a predicted value.

For regression tasks, the output $\hat{y}_i$ is the average of the target values within a leaf node. The tree is built by selecting splits that minimize the variance within each leaf:

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

$$\text{Variance}_{\text{split}} = \frac{1}{N_{\text{left}}} \sum_{i \in \text{left}} (y_i - \hat{y}_{\text{left}})^2 + \frac{1}{N_{\text{right}}} \sum_{i \in \text{right}} (y_i - \hat{y}_{\text{right}})^2 \quad (12)$$

Where, $\hat{y}_{\text{left}}$ and $\hat{y}_{\text{right}}$ are the mean predicted values of the left and right splits, respectively.
The algorithm iterates this process recursively and then prunes the tree to avoid overfitting.

3.5.3 Random Forests

Random Forest combines multiple decision trees to form an ensemble learning model. Each tree is trained on a random subset of data with random feature selection, and the final prediction is made by averaging the predictions from all trees.

Let $T_1, T_2, \ldots, T_k$ represent k decision trees, each producing an individual prediction $\hat{y}_i^{(k)}$ for the i^{th} observation. The final prediction is obtained by averaging all the individual predictions:

$$\hat{y}_i = \frac{1}{k} \sum_{k=1}^{k} \hat{y}_i^{(k)} \quad (13)$$

Where:

- $\hat{y}_i^{(k)}$ is the predicted value from the k^{th} tree,
- k is the total number of trees in the forest.

The random forest reduces the variance of individual decision trees by averaging predictions, which helps in preventing overfitting.

3.5.4 Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is an ensemble method that builds trees sequentially. Each new tree is trained to predict the residual errors made by the previous trees, with the aim of improving the model's accuracy over iterations.

At iteration m, the model is updated as follows:

$$f_m(x) = f_{m-1}(x) + \eta h_m(x) \quad (14)$$

Where:

- $f_{m-1}(x)$ is the prediction from the previous model,
- $h_m(x)$ is the new tree learned to predict the residuals,
- η is the learning rate (a scaling factor).

The optimization process involves minimizing the loss function L over all iterations:

$$\text{Loss} = \sum_{i=1}^{N} L(y_i, \hat{y}_i) \quad (15)$$

Where:

- y_i is the true value,
- $\hat{y}_i$ is the predicted value.

GBM updates the model iteratively to minimize this loss function and improve prediction accuracy.

3.5.5 Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized implementation of gradient boosting that incorporates regularization terms to prevent overfitting and improve model generalization. XGBoost also uses second-order gradients for faster convergence. The objective function for XGBoost is a combination of the loss function and a regularization term:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{j=1}^{T} \Omega(f_j) \quad (16)$$

Where:

- $L(y_i, \hat{y}_i)$ is the loss function (e.g., Mean Squared Error for regression),
- $\Omega(f_j)$ is the regularization term for each tree f_j that controls the complexity of the model and prevents overfitting.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

The regularization term for each tree is given by:

$$\Omega(f_j) = \gamma T + \frac{1}{2}\lambda \sum_{k=1}^{T} w_k^2 \quad (17)$$

Where:

- T is the number of leaves in the tree,
- w_k is the weight of the k^{th} leaf,
- γ and λ are regularization parameters.

XGBoost optimizes this objective function during training, balancing loss minimization and model complexity.

4. RESULTS AND ANALYSIS

The results are analyzed to determine the relationship between sentiment and stock returns. We compare the performance of the machine learning models with a naive benchmark model, which predicts stock returns based solely on historical averages. By comparing the two, we can evaluate whether incorporating sentiment data improves prediction accuracy and provides actionable insights into market behaviour.

4.1 Performance Evaluation Metrics

The models' performances are evaluated using several metrics suited for regression tasks, which measure how well the predicted stock returns match the actual returns.

Mean Absolute Error (MAE): It measures the average magnitude of errors, without considering their direction. It is calculated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (18)$$

A smaller MAE indicates better model performance, but MAE does not penalize large errors as much as RMSE, so it might not be suitable if larger errors are undesirable.
Where y_i is the actual return and $\hat{y}_i$ is the predicted return.

Root Mean Square Error (RMSE): It measures the average magnitude of errors, with higher weights given to larger errors due to squaring the differences. It is given by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \quad (19)$$

While RMSE provides a useful evaluation of predictive accuracy, it can be sensitive to outliers, which might not always be desirable depending on the application.

Mean Absolute Percentage Error (MAPE): The MAPE provides a normalized error metric expressed as a percentage, making it easy to interpret the accuracy of predictions relative to the actual values:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \quad (20)$$

MAPE is often preferred in scenarios where the scale of the prediction matters and when comparing across different datasets or models.

Coefficient of Determination (R^2): It indicates how well the model's predictions match the observed data. It is interpreted as the proportion of variance explained by the model, with a value of 1 meaning perfect predictions, and 0 meaning the model explains none of the variance.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

(21)

Where, $\bar{y}$ is the mean of the actual values.
An R^2 closer to 1 indicates better model performance, and R^2 is particularly helpful when comparing models with different complexity.

4.2 Results

The analysis of results in this study focuses on assessing the sentiments derived from financial news published by specialized news sources, with particular emphasis on understanding how sentiments around economic and political events impact stock market returns. The study examines sentiments expressed in news articles collected from various online portals, such as the Economic Times, and how these sentiments correlate with market movements over time. The results aim to provide insights into the influence of textual sentiment on stock market predictions.

4.2.1 Tone and Importance of Words

This section presents an analysis of the sentiment of words, specifically focusing on the TF-IDF (Term Frequency-Inverse Document Frequency) values weighted by the sentiment polarity from a sentiment lexicon (e.g., SentiWordNet for English or any regional lexicon in Indian languages). Words that occur frequently in financial articles are weighted with respect to both their frequency of occurrence and their sentiment (positive or negative) associated with stock market reactions.
For this analysis, positive terms are identified based on their direct correlation with optimism and positive stock market behaviors. Terms such as “growth,” “success,” “uptrend,” and “positive” are typically associated with favorable sentiment. On the contrary, negative terms like “downturn,” “recession,” “loss,” and “uncertainty” are associated with negative sentiment and market declines.

Key Observations:

- Positive words generally outweigh negative words in frequency, though fluctuations are seen based on market conditions, political stability, and economic indicators.
- Words related to economic recovery or optimism are observed during periods of growth or post-recessionary phases, while negative words surge during economic slowdowns, crises, or political instability.

4.2.2 Textual Sentiment of News (SentNews)

Table 1 presents the average sentiment scores extracted from political and economic news over several months, based on their impact on stock market returns. The sentiment values were calculated by aggregating the sentiment of individual articles, weighted by their importance (TF-IDF), across each month.

Table 1: Example Sentiment Analysis Table

Month	Average Positive Sentiment	Average Negative Sentiment	Net Sentiment
December 2020	0.58	-0.12	0.46
January 2021	1.03	-0.67	0.36
February 2021	1.12	-0.89	0.23
March 2021	0.67	-1.45	-0.78
April 2021	0.92	-0.79	0.13
June 2021	1.51	-0.23	1.28
December 2021	2.01	-0.76	1.25

- The highest positive sentiment values in June 2021 and December 2021 correlate with periods of strong market growth, potentially due to positive fiscal policies and economic recovery post-pandemic.
- The negative sentiment peak in March 2021 is indicative of market corrections or global uncertainties, likely influenced by political or economic instability (e.g., COVID-19 resurgence).
- Overall, 2021 shows a positive trend, especially in the second half, with optimism around vaccine rollouts and economic recovery.

4.2.3 Analysis of Sentiment in 2022

When examining data for the year 2022, it is clear that political uncertainty, particularly around the general elections and global events like the Ukraine-Russia conflict, significantly impacted the sentiment. Table 2 reflects

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

the fluctuations in positive and negative sentiment, showing the dominance of negative sentiment during several months.

Table 2: Example Sentiment Analysis Table

Month	Average Positive Sentiment	Average Negative Sentiment	Net Sentiment
January 2022	1.25	-0.62	0.63
February 2022	1.07	-1.10	-0.03
June 2022	0.34	-3.56	-3.22
July 2022	0.12	-3.99	-3.87
August 2022	0.56	-1.50	-0.94
December 2022	1.43	-0.85	0.58

- June and July 2022 witnessed strong negative sentiment due to ongoing geopolitical tensions and economic challenges, such as inflation and supply chain disruptions.
- However, a slight recovery is observed in the latter half of the year, with positive sentiment improving toward the end of 2022, possibly due to policy adjustments and recovery plans.

4.2.4 Correlation between Sentiment and Stock Market Performance

In this study, textual sentiment was found to have a significant correlation with stock market returns. It was observed that positive sentiment was more closely associated with bullish trends in the market, while negative sentiment aligned with periods of market downturns or corrections. Using the SentNews index, which measures the aggregated sentiment of articles, we observed a clear pattern where significant negative sentiment preceded market declines, and positive sentiment often preceded market rallies.

Table 4: Stock Market Return Correlation Analysis

Sentiment Category	Average Market Return (%)
Positive Sentiment	4.3%
Negative Sentiment	-3.5%
Neutral Sentiment	0.1%

- Periods with a positive sentiment from the news (especially relating to the economy, politics, and business outlook) generally resulted in positive stock market returns.
- Negative sentiment, typically associated with economic slowdowns, political instability, or external shocks, was linked to a decline in stock market performance.

The sentiment analysis performed in this study demonstrates a clear relationship between textual sentiment in financial news and stock market returns. These findings suggest that positive sentiment tends to coincide with market rallies, while negative sentiment often foreshadows declines. The importance of using textual sentiment data as an additional feature for forecasting stock returns is reinforced, as it provides valuable insights into market psychology. This approach not only enriches traditional quantitative models but also highlights the significance of news and sentiment in shaping investor behavior and market trends.

4.2.5 Conditional Textual Sentiment to IBOVESPA

In this section, we present an analysis of the Conditional Textual Sentiment of news articles in relation to the IBOVESPA index (SentNewsIBOV). Specifically, this analysis examines how machine learning models are applied to identify the most significant words in news articles that predict stock returns. In particular, we utilize XGBoost, a decision-tree-based model, which learns over time the importance of individual words in predicting market behavior. The model is built using over 1,000 predictors, where each word is treated as a feature. The model provides insights into which words play a crucial role in explaining stock market returns.

Key Findings:

- Words like "economic growth," "recovery," and "inflation" appeared as top features, emphasizing the importance of economic indicators.
- "Political stability" and "uncertainty" also emerged as influential predictors, reflecting how political events impact market returns.
- The XGBoost model was trained exclusively on textual data, with no aggregation, ensuring that each word's sentiment could be assessed in real-time.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

- The most influential terms from the news articles, such as "growth," "political stability," and "inflation," were found to be highly predictive of IBOVESPA's movement.

4.2.6 Stock Return Prediction Analysis

Following the estimation of sentiment measures, we tested various machine learning models to predict the next-day stock returns of IBOVESPA. The models considered include historical returns, SentNews, and SentNewsIBOV as predictors. We evaluated these models using several regression metrics, with a focus on RMSE (Root Mean Square Error), to measure prediction accuracy.

Table 5: Prediction Model Comparison

Model	RMSE	MAE	MSE	RMSLE	MAPE
Base Model	0.015	0.011	0.0003	0.014	1.21
XGBoost Tuned	0.013	0.009	0.0002	0.012	1.31
Ridge	0.014	0.010	0.0003	0.013	1.25
Linear Regression	0.015	0.010	0.0003	0.013	2.01
Random Forest	0.016	0.011	0.0004	0.011	3.80
Lasso	0.015	0.010	0.0003	0.012	1.09
Elastic Net	0.015	0.010	0.0003	0.012	1.09
Gradient Boosting	0.015	0.010	0.0003	0.012	3.55
XGBoost	0.016	0.011	0.0004	0.012	5.26

Analysis:

- The tuned XGBoost model outperformed all other models, achieving the lowest RMSE (0.013) and MAE (0.009), indicating the most accurate predictions.
- XGBoost Tuned also performed well in terms of MAPE (1.31), suggesting good predictive accuracy in percentage terms.
- Other models, such as Ridge and Lasso, also performed well, but XGBoost Tuned remained the top performer due to its ability to handle complex interactions in the data.
- The Random Forest model showed the lowest RMSLE (0.011), which indicates better performance when predicting on a logarithmic scale.

Model Interpretation:

- XGBoost Tuned is more adept at capturing complex patterns in the data, resulting in lower prediction errors compared to simpler models like Linear Regression or Base Model.
- The Ridge and Lasso models demonstrated strong performance, but they were slightly less effective than XGBoost in terms of RMSE and MAPE.

4.2.7 Feature Importance in XGBoost Model

The feature importance analysis in the XGBoost Tuned model reveals the most impactful predictors used to explain stock return predictions. Among these, SentNewsIBOV (the sentiment of news related to IBOVESPA) emerged as the most significant variable. This suggests that news sentiment has a greater influence on stock market predictions than the historical returns alone.

Key Findings:

- SentNewsIBOV had the highest contribution to the model, reinforcing the idea that textual sentiment plays a crucial role in explaining market behavior.
- The inclusion of historical returns and SentNews as additional features also improved the model's prediction accuracy, but they were secondary to the importance of sentiment-based predictors.

5. CONCLUSION

The analysis demonstrates that machine learning models, especially XGBoost, significantly enhance the prediction of stock market returns by incorporating sentiment extracted from financial news. Among the various models tested, XGBoost outperformed the others in terms of prediction accuracy, achieving the lowest RMSE and MAE, indicating that it effectively captures the complex patterns in stock market data. The addition of sentiment features like SentNews and SentNewsIBOV further improved the model's performance, with SentNewsIBOV emerging as the most influential predictor. This reinforces the critical role of sentiment in shaping market behavior, where positive sentiment correlates with market upturns and negative sentiment often precedes

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

downturns. Other models, such as Random Forests and Gradient Boosting, also provided competitive results but did not surpass the performance of XGBoost. Overall, the study underscores the potential of integrating sentiment analysis into stock market prediction models, offering a promising approach to better understand market dynamics and enhance forecasting accuracy.

REFERENCES

1. Axtell, R.L. and Farmer, J.D., 2022. Agent-based modeling in economics and finance: Past, present, and future. Journal of Economic Literature, pp.1-101.
2. Li, X., Wu, P. and Wang, W., 2020. Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 57(5), p.102212.
3. Davenport, T.H. and Ronanki, R., 2018. Artificial intelligence for the real world. Harvard business review, 96(1), pp.108-116.
4. Hsu, Y.J., Lu, Y.C. and Yang, J.J., 2021. News sentiment and stock market volatility. Review of Quantitative Finance and Accounting, 57(3), pp.1093-1122.
5. Audrino, F., Sigrist, F. and Ballinari, D., 2020. The impact of sentiment and attention measures on stock market volatility. International Journal of Forecasting, 36(2), pp.334-357.
6. Ramagundam, S. (2021). Next Gen Linear Tv: Content Generation And Enhancement With Artificial Intelligence. International Neurourology Journal, 25(4), 22-28.
7. Audrino, F., Sigrist, F. and Ballinari, D., 2020. The impact of sentiment and attention measures on stock market volatility. International Journal of Forecasting, 36(2), pp.334-357.
8. Hong, W., Gu, Y., Wu, L. and Pu, X., 2022. Impact of online public opinion regarding the Japanese nuclear wastewater incident on stock market based on the SOR model. Math Biosci Eng, 20(5), pp.9305-9326.
9. Olweny, T. and Omondi, K., 2011. The effect of macro-economic factors on stock return volatility in the Nairobi stock exchange, Kenya. Economics and Finance review, 1(10), pp.34-48.
10. Shahzad, A., Zahrullail, N., Akbar, A., Mohelska, H. and Hussain, A., 2022. COVID-19's Impact on fintech adoption: Behavioral intention to use the financial portal. Journal of Risk and Financial Management, 15(10), p.428.
11. Dunham, L.M. and Garcia, J., 2021. Measuring the effect of investor sentiment on liquidity. Managerial Finance, 47(1), pp.59-85.
12. Shiller, R.J., 2017. Narrative economics. American economic review, 107(4), pp.967-1004.
13. Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M.M. and Muhammad, K., 2020. A local and global event sentiment based efficient stock exchange forecasting using deep learning. International Journal of Information Management, 50, pp.432-451.
14. Yen, G. and Lee, C.F., 2008. Efficient market hypothesis (EMH): past, present and future. Review of Pacific Basin Financial Markets and Policies, 11(02), pp.305-329.
15. Ruggeri, K., Alí, S., Berge, M.L., Bertoldo, G., Bjørndal, L.D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M. and Gibson, S.P., 2020. Replicating patterns of prospect theory for decision under risk. Nature human behaviour, 4(6), pp.622-633.
16. Peress, J., 2014. The media and the diffusion of information in financial markets: Evidence from newspaper strikes. The Journal of Finance, 69(5), pp.2007-2043.
17. Thakkar, A. and Chaudhari, K., 2021. A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. Expert Systems with Applications, 177, p.114800.
18. Devaraj, M., 2020, September. Analyzing News Sentiments and their Impact on Stock Market Trends using POS and TF-IDF based approach. In 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET) (pp. 1-6). IEEE.
19. Khder, M.A., 2021. Web scraping or web crawling: State of art, techniques, approaches and application. International Journal of Advances in Soft Computing & Its Applications, 13(3).