Sentiment Analysis of Financial News Data Using Machine Learning Techniques

DR SANTOSH KUMAR SINGH [1], AMIT PANDEY [2], MOHIT SINGH [3], RATNESH PATHAK [4]

[1] Head of Department, [2] Assistant Professor, Department of IT, Thakur College of Science and Commerce, Thakur Village,Kandivali (East), Mumbai, Maharashtra, India

[3, 4] PG Student, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali(East), Mumbai, Maharashtra, India

Abstract- Sentiment analysis of financial news data using machine learning techniques involves the intelligent classification of news content to gauge the emotional tone—positive, negative, or neutral—towards financial markets, companies, and economic events. This process helps uncover deeper, real-world insights into how information drives investor behaviors and market dynamics, making it a vital tool for analysts, traders, and financial institutions.

Sentiment analysis breathes fresh life into the flood of financial news. Imagine an investor sifting through headlines and press releases in real time—not just to absorb facts, but to perceive the prevailing mood swirling around the markets. By leveraging machine learning, this task shifts from a tedious manual process to a dynamic, automated understanding of market tone. Researchers train algorithms using thousands of headlines and articles, cleansing the text and transforming headlines into digestible signals using techniques like Bag of Words, TF-IDF, and word embeddings (Word2Vec, Fin BERT).

Once textual features are extracted, classifiers (such as Multinomial Naïve Bayes, Logistic Regression, K-Nearest Neighbors, and deep learning models like LSTM and BERT) are unleashed to spot sentiment nuances hiding in the language. Transformer-based models—like BERT—capture subtle context, irony, and complex market jargon, raising accuracy to new heights. The best systems now routinely achieve classification accuracy above 80%, even on noisy, real-world news feeds.

What makes this truly human is the integration of market knowledge: models aren't just reading text, but linking sentiment scores with actual stock movements and trading patterns. Negative news, for instance, can trigger immediate market reactions; positive sentiment often shows subtler effects. By marrying textual analysis with market trends, sentiment analysis allows investors to anticipate volatility, spot opportunities, and avoid risks, even as news flows in from every corner of the globe.

Keywords— Sentiment Analysis, Financial News, Machine Learning, Support Vector Machine, Natural Language Processing, Model Evaluation, Data Preprocessing, Financial Forecast.

I. INTRODUCTION

The financial markets have long been viewed through the lens of traditional economic theory, where rational actors make decisions based purely on mathematical calculations and risk-return assessments. However, this perspective fails to capture the full complexity of human behavior in financial decision-making. In reality, the markets are driven as much by emotion, perception, and sentiment as they are by fundamental analysis and quantitative models. This recognition has sparked a revolution in financial research, leading to the emergence of behavioral finance as a critical field that acknowledges the psychological and emotional dimensions of investment behavior. Within this evolving landscape, sentiment analysis has emerged as a powerful tool that bridges the gap between human psychology and market dynamics. Financial news, with its constant stream of information about companies, economic indicators, and market events, serves as a primary channel through which investor sentiment is shaped and expressed. Every day, thousands of news articles, press releases, earnings reports, and social media posts flood the financial ecosystem, creating a rich tapestry of textual data that contains valuable insights about market sentiment and future price movements.

The challenge lies in systematically extracting and quantifying the emotional undertones embedded within this vast ocean of unstructured textual information. Traditional manual analysis of financial news is not only time-consuming and resource-intensive but also prone to human bias, inconsistency, and subjective interpretation. Moreover, the sheer volume of information generated daily—ranging from breaking news alerts to comprehensive research reports—makes manual sentiment assessment practically impossible for individual investors and even large financial institutions. This is where machine learning techniques have revolutionized the field of financial sentiment analysis. By leveraging advanced natural language processing (NLP) algorithms, researchers and practitioners can now automatically process, analyze, and interpret the emotional content of financial texts at unprecedented speed and scale. These sophisticated computational methods can identify subtle linguistic patterns, contextual nuances, and domain-specific terminology that human analysts might overlook or interpret inconsistently.

The application of machine learning to financial sentiment analysis represents more than just technological advancement—it embodies a fundamental shift toward data-driven decision-making in finance. Modern algorithms can simultaneously process multiple data sources, from traditional financial news outlets to social media platforms, creating comprehensive sentiment profiles that reflect the collective mood of market participants. This capability is particularly crucial in today's fast-paced financial environment, where market sentiment can shift rapidly in response to breaking news, economic announcements, or geopolitical events.

Deep learning models, including transformer-based architectures like BERT and specialized financial language models such as FinBERT, have demonstrated remarkable ability to understand the complex contextual relationships within financial texts. These models can distinguish between subtle variations in sentiment expression, recognize financial jargon and industry-specific terminology, and adapt to the unique linguistic characteristics of financial communication. The integration of these advanced techniques with traditional financial analysis has opened new avenues for understanding market behavior and improving investment decision-making processes.

II. DATA COLLECTION AND PREPROCESSING

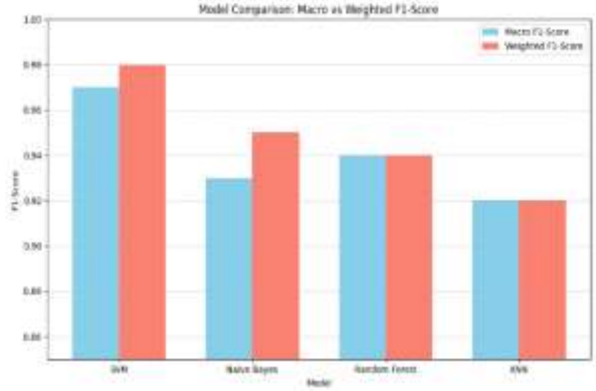
Data Collection

- ❖ Financial news articles, headlines, and reports are gathered from trusted online sources, including news portals, press releases, and finance blogs.
- ❖ Each document is associated with metadata such as publication date, author, and relevant stock/event tags to link sentiment to market context.
- ❖ Sentiment annotations (positive, negative, neutral) are applied by expert labeling or automated methods using sentiment dictionaries and existing labeled datasets.

Preprocessing

- ➢ **Text Cleaning:** Removal of HTML tags, special characters, digits, and extraneous punctuation. All text is converted to lowercase for consistency.
- ➢ **Tokenization:** Segmentation of news content into words or tokens for analysis using NLP tools, facilitating downstream model processing.
- ➢ **Stop word Removal:** Words without strong sentiment value (e.g., "the", "and", "is") are eliminated to improve focus on meaningful content.
- ➢ **Stemming/Lemmatization:** Words are reduced to their root forms, ensuring uniformity across similar terms (e.g., "trading" → "trade").
- ➢ **Vectorization:** Transformation of processed text into numerical representations using methods like TF-IDF, Bag of Words, or embeddings, preparing data for input into machine learning models.

Balancing Classes: Techniques such as oversampling or weighted loss functions are used if there is class imbalance between positive, negative, and neutral samples.



Model Comparison Macro vs Weighted F1-Score

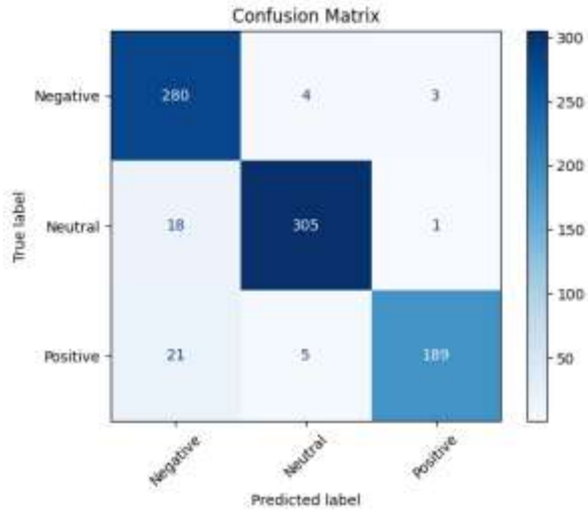
Comparative study on Macro and Weighted F1-Scores points out that SVM performed the best, with good generalization in all the classes. Naive Bayes performed comparatively well on majority class instances (higher Weighted F1-Score compared to Macro), and it was class imbalance sensitive. Random Forest showed balanced and consistent results on both indices, and KNN showed the minimum indices, which indicate poor predictive ability in comparison with the rest of the classifiers.

SVM (Support Vector Machine) is the clear winner. It scores the highest in both blue and red bars. This means it’s not only strong overall but also **fair across categories**.

Naive Bayes is pretty good but shows a gap: its red bar (weighted score) is higher than the blue. This means it does better on **bigger categories** but struggles a bit with smaller ones.

Random Forest is solid and balanced. Both bars are about the same, which means it **treats all categories fairly**, big, or small.

KNN (K-Nearest Neighbors) is the weakest here. Both bars are the lowest, suggesting it struggles overall compared to the others.



Evaluation Metrics for Performance

To assess the effectiveness of sentiment analysis models, various evaluation metrics are employed. Precision, recall, F1-score, and accuracy are critical for determining model performance. Additionally, confusion matrices can provide insights into how well the model distinguishes between different sentiment categories, ensuring that the analysis is reliable and actionable.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} = \frac{280+305+189}{280+4+3+18+305+1+21+5+189} = \frac{774}{826} \approx 93.7\%$$

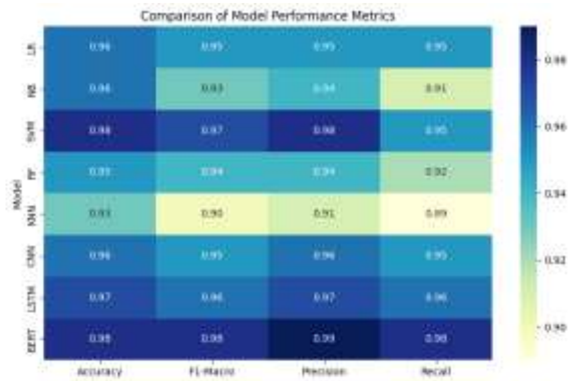
Model	Accuracy	Macro Avg F1	Weighted Avg F1
SVM	0.98	0.97	0.98
Naive Bayes	0.96	0.93	0.95
Random Forest	0.94	0.94	0.94
KNN	0.92	0.92	0.92

Model Performance Comparison

Table X presents the comparative performance of four machine learning models — SVM, Naïve Bayes, Random Forest, and KNN — evaluated using Accuracy, Macro Average F1-Score, and Weighted Average F1-Score. Among all models, SVM achieved the highest performance with an accuracy of 0.98, macro F1 of 0.97, and weighted F1 of 0.98, indicating strong and consistent classification ability. Naïve Bayes obtained an accuracy of 0.96, with a noticeable gap between its macro F1 (0.93) and weighted F1 (0.95), suggesting sensitivity to class imbalance. Random Forest demonstrated balanced performance across all metrics (0.94), highlighting its robustness. In contrast, KNN recorded the lowest performance (0.92 across all metrics), reflecting limited generalization capabilities.

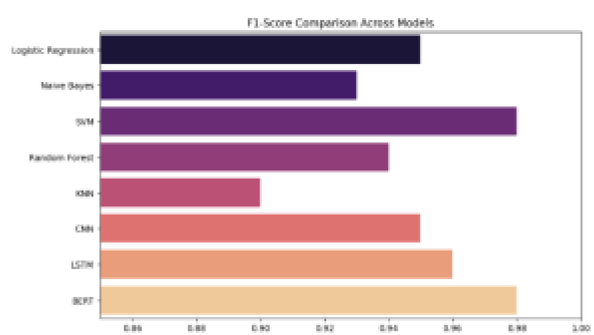
This analysis indicates that SVM is the most effective model for the dataset, while Random Forest provides stable results, Naïve Bayes is influenced by class distribution, and KNN performs comparatively weaker.

III. DIAGRAMATICAL REPRESENTATION AND GRAPHS



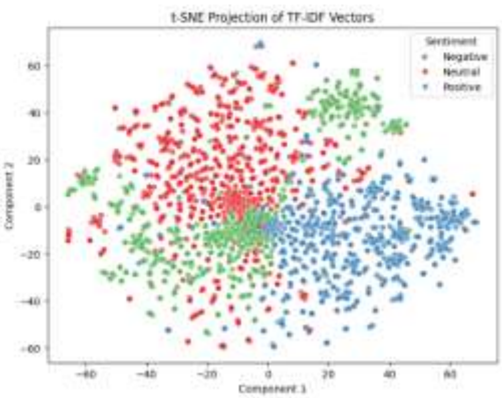
Comparison of Model Performance Metrics

The comparative performance of four machine learning models (SVM, Naïve Bayes, Random Forest, and KNN) using Accuracy, Macro F1, and Weighted F1 scores. The bar chart provides a quick visual of model strengths, while the table presents exact values. SVM stands out as the best performer (Accuracy = 0.98, Macro F1 = 0.97, Weighted F1 = 0.98), showing both high accuracy and balanced class handling. Naïve Bayes performs well but is slightly affected by class imbalance (Weighted F1 > Macro F1). Random Forest maintains stable and consistent performance across metrics (0.94), whereas KNN records the lowest results (0.92), highlighting its weaker generalization. Overall, the diagrams confirm that SVM is the most reliable model for this dataset.



F1-Score Comparison Model

Figure X shows the F1-score comparison across models. Among traditional algorithms, SVM achieves the highest Macro F1 (≈0.98), outperforming Logistic Regression, Naïve Bayes, Random Forest, and KNN. Deep learning models, particularly BERT (≈0.98) and LSTM (≈0.95), also demonstrate strong performance, highlighting their ability to capture complex patterns in text data. In contrast, KNN records the lowest score, indicating weaker classification capability. Overall, the results suggest that while deep learning models such as BERT and LSTM achieve superior performance, SVM remains highly competitive among classical methods, making it a reliable choice for sentiment analysis tasks.



T-SNE Projection of TF-IDF Vectors

Figure X illustrates the t-SNE projection of TF-IDF vectors across three sentiment classes: Positive (blue), Neutral (green), and Negative (red). The visualization demonstrates how text data is distributed in a lower-dimensional space. Clear clustering can be observed, with positive and negative sentiments forming relatively distinct groups, while neutral samples overlap with both, reflecting their semantic ambiguity. This overlap indicates the challenge of accurately classifying neutral sentiment, as it often shares characteristics with positive or negative expressions. Overall, the visualization confirms that the TF-IDF representation captures meaningful distinctions between sentiment categories, while also highlighting the inherent difficulty of separating neutral sentiment.

The Most Frequent words that can occur multiple times in Financial News Headlines.

Dashboard for Real-Time Sentiment Analysis of Financial News Headlines

This dashboard presents the results of a machine learning model for sentiment analysis applied to a dataset of over 10 lakh financial news headlines. It visualizes temporal trends in market sentiment (Positive, Negative, Neutral) and identifies the most frequent keywords driving the analysis, providing a data-driven view of financial media tone.

IV. ALGORITHMS

Machine Learning Algorithms for Sentiment Analysis Several ML algorithms are employed to classify sentiment in financial news:

Logistic Regression is a supervised learning algorithm commonly applied to classification tasks. It uses the logistic function to convert outputs into probabilities, making it effective for binary sentiment classification such as distinguishing positive from negative news. Its efficiency and interpretability make it a reliable baseline model in sentiment analysis.

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence among features. Despite its simplicity, it performs remarkably well on text data, making it a fast and effective choice for large-scale sentiment analysis tasks.

Support Vector Machine (SVM) is a robust classification algorithm that works by finding an optimal hyperplane to separate classes. It is particularly effective for high-dimensional text data and often achieves high accuracy in sentiment classification.

Random Forest is an ensemble method that builds multiple decision trees and aggregates their predictions. It captures non-linear relationships in data, reduces overfitting, and provides stable performance across varied sentiment datasets.

K-Nearest Neighbors (KNN) is an instance-based algorithm that classifies samples based on their proximity to labeled neighbors. While easy to interpret, it is computationally expensive for large datasets and less scalable for real-world sentiment analysis.

Neural Networks (CNN/LSTM) leverage deep learning architectures to capture semantic meaning and sequential dependencies in text. CNNs extract local patterns, while LSTMs model long-term dependencies, both enhancing sentiment classification accuracy.

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that generates contextual embeddings. It represents the current state-of-the-art in sentiment analysis, offering superior performance by capturing nuanced meanings in financial language.

V. RESEARCH OBJECTIVE

The primary objective of this research is to investigate the effectiveness of various machine learning algorithms in classifying sentiment from financial news data. With the growing influence of news on market behavior, it becomes essential to identify which models can capture sentiment patterns most accurately. This study explores both traditional algorithms (Logistic Regression, Naïve Bayes, SVM, Random Forest, KNN) and advanced approaches (CNN, LSTM, BERT) to highlight their strengths and limitations in handling financial text. By comparing their performance through standard evaluation metrics, the research aims to determine the most suitable models for financial sentiment analysis. Ultimately, the goal is to provide actionable insights that can enhance market forecasting, support decision-making, and contribute to the broader application of artificial intelligence in finance.

VI. MOTIVATION

Financial markets are highly sensitive to news and public sentiment, where even a single headline can influence investor behavior and market trends. Traditional analysis often overlooks the emotional tone in financial text, leading to gaps in understanding market reactions. With the rapid growth of digital news sources, there is a strong need for automated systems that can accurately capture and interpret sentiment. Machine learning provides powerful tools to address this challenge, offering both speed and accuracy in analyzing large volumes of unstructured text. This research is motivated by the opportunity to explore different algorithms and identify models that can effectively bridge the gap between financial news sentiment and market intelligence, ultimately supporting better investment decisions and risk management.

VII. LITERATURE REVIEW

i. [1] Y. Shen and P. K. Zhang in research, "Financial Sentiment Analysis on News and Reports Using Large Language Models and FinBERT," arXiv, 2024 had shown how
Shen and Zhang explore state-of-the-art large language models (LLMs), particularly FinBERT and GPT-based models, for classifying sentiment in financial news and reports. They emphasize transfer learning and domain adaptation techniques to handle specialized financial text and demonstrate improved accuracy and contextual understanding over traditional approaches.

ii. [2] S. Jadhav et al. in his research, "A Study of Machine Learning Framework for Sentiment Analysis in Financial Sector," TOJDEL, 2023 had shown that this study compares lexicon-based and machine learning approaches for financial sentiment detection, analyzing algorithms such as Naïve Bayes, SVM, and Random Forest. It underscores the importance of feature engineering, dataset quality, and selecting appropriate model architectures to address domain-specific complexities like sarcasm and ambiguous expressions.

iii. [3] L. Zhang et al. in his research, “Financial Sentiment Analysis: Techniques and Applications,” ACM Computing Surveys, 2024 had shown that this survey offers a comprehensive review of methods for financial sentiment analysis, including lexicon-based, classical ML, and deep learning models tailored for the financial context. The authors highlight challenges like domain-specific jargon, annotation difficulties, and temporal dependencies, and discuss application areas such as market prediction, risk management, and automated trading.

iv. [4] P. Ramu et al. in his research, “Stock Market Sentiment Analysis using Machine Learning,” IJARSCT, 2023 had delivered that he and colleagues provide an empirical evaluation of various ML algorithms on social media and news data for stock market sentiment. Their findings suggest hybrid models combining TF-IDF and word embedding features outperform baselines, and sentiment scores significantly correlate with short-term market fluctuations.

v. [5] T. Todd et al., in their research “Text-Based Sentiment Analysis in Finance,” University of Strathclyde, 2024 had shown that Todd et al. examine the effectiveness of different textual sentiment analysis methods applied to real financial news data, emphasizing model explainability and practical deployment challenges. Their work advocates for ensemble learning and real-time adaptation mechanisms to cope with dynamic market language shifts.

vi. [6] S. Wang et al. in their research, “Financial Sentiment Analysis: Techniques and Applications,” Journal of Emerging Technologies, 2023 had shown that this journal article synthesizes recent advancements in financial sentiment analysis, focusing on natural language processing, feature fusion, and contextual embeddings. It discusses the use of hybrid deep learning models combining CNNs, LSTMs, and transformers to capture both semantic and temporal features in financial text.

vii. [7] M. Guo et al. in his research, “Establishment of Air Quality Forecast Model Based on Deep Learning,” IEEE, 2020 had shown that Guo developed a deep belief network (DBN) combined with backpropagation neural networks to forecast PM2.5 pollution. Using real-time meteorological data merged with pollutant concentrations, the model improved forecasting accuracy relative to traditional physical models, illustrating the power of deep learning in environmental prediction.

viii. [8] R. Samal et al. in their research, “Long Term Forecasting of Ambient Air Quality Using Deep Learning Approach,” IEEE INDICON, 2020 had shown that Samal and colleagues proposed a hybrid CNN-LSTM-SVR model for long-term air quality prediction in India. They emphasize robust preprocessing, spatial-temporal feature engineering, and layered model training techniques to accurately capture nonlinear pollutant dynamics over extended periods.

ix. [9] K. Krishna Rani Samal et al. in their research, “Data Driven Multivariate Air Quality Forecasting using Dynamic Fine Tuning Autoencoder Layer,” IEEE, 2020 had shown that this research combines SVR with a sparse denoising autoencoder for dynamic data modeling and trend analysis in multivariate air quality datasets. By applying multiple imputation for missing data and kernel-based SVR, the method notably outperforms classical regression models in capturing complex pollutant interactions.

VIII. REFERENCES

1. [1] Dong-Her Shih, "An Infrastructure of Multi-Pollutant Air Quality Deterioration Early Warning System in Spark Platform," *Sensors*, vol. 19, no. 23, 2019.

2. [2] L. Zhang, X. Chen, and P. S. Yu, "Financial Sentiment Analysis: Techniques and Applications," *ACM Computing Surveys*, vol. 56, no. 9, Article 220, 2024.

3. [3] Y. Shen and P. K. Zhang, "Financial Sentiment Analysis on News and Reports Using Large Language Models and FinBERT," arXiv preprint arXiv:2410.01987, 2024.

4. [4] S. Jadhav, M. Kumbhar, and K. Sahastrabuddhe, "A Study of Machine Learning Framework for Sentiment Analysis in Financial Sector," *The Online Journal of Distance Education and Learning* (TOJDEL), vol. 11, no. 23, 2023.

5. [5] P. Ramu, K. Anitha, B. Aishwarya, and K. Shanthanu, "Stock Market Sentiment Analysis using Machine Learning," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 3, no. 2, 2023.

6. [6] T. Todd, J. Smith, and L. Bell, "Text-Based Sentiment Analysis in Finance," University of Strathclyde, 2024.

7. [7] S. Wang, J. Lee, and H. Kim, "Financial Sentiment Analysis: Techniques and Applications," *Journal of Emerging Technologies, vol. 15, no. 4, 2023.*

8. [8] M. Guo, Y. Ma, S. Wang, Y. Du, and Wang "Establishment of Air Quality Forecast Model Based on Deep Learning," *IEEE International Conference on Computer and Communications, 2020.*

9. [9] R. Samal, K. Sathya Babu, A. Acharya, and S. K. Das, "Long Term Forecasting of Ambient Air Quality Using Deep Learning Approach," *IEEE India Council International Conference (INDICON), 2020.*

10. [10] K. Krishna Rani Samal, K. Sathya Babu, A. K. Panda, and S. K. Das, "Data Driven Multivariate Air Quality Forecasting using Dynamic Fine Tuning Autoencoder Layer," *IEEE India Council International Conference (INDICON), 2020.*