Borsa Istanbul Review xxx (xxxx) xxx

Contents lists available at ScienceDirect

Borsa Istanbul Review

journal homepage: www.elsevier.com/journals/borsa-istanbul-review/2214-8450

From headlines to stock trends: Natural language processing and explainable artificial intelligence approach to predicting Turkey's financial pulse

Mahat Maalim Ibrahim *, Asad Ul Islam Khan, Muhittin Kaplan

Ibn Haldun University, Turkey

ARTICLE INFO

Keywords:
Transformers
Natural language processing
Stock market
Explainable AI
Machine learning
Sentiment analysis

ABSTRACT

The dynamic field of financial markets is constantly in search of new ways to understand complex market dynamics. The increasing availability of vast amounts of text data offers new avenues for investigation (Botchway et al., 2020). This study aims to shed light on the dynamics between stock market movements and news narratives in Turkey. To address this issue, the study will include the analysis of business, financial, and economic news from four major news journals (The Economist, The New York Times, The Guardian, and Yeni Şafak) along with local tweets. Yeni Şafak and local tweets serve as proxies for local news sentiment. The analysis rests on daily Turkish stock market data from January 1, 2015, to February 27, 2024, obtained from Yahoo Finance. The issue was addressed using state-of-the-art Natural Language Processing (NLP), machine learning, and explainable AI techniques. The findings reveal that international news significantly predicts the Turkish Stock market, with the majority of machine learning models yielding approximately 80 percent predictive accuracy. The Explainable AI methods demonstrate that traditional international news media have a significant impact on the Turkish stock market in comparison to local news sources such as Yeni Şafak and Twitter which serve as less effective predictors. Notably, the ensemble algorithms, comprising Random Forest, Gradient Boosting, and XGBoost, demonstrate robust performance across all datasets.

1. Introduction

Financial markets have been the subject of intense scrutiny and analysis for decades, with researchers and investors alike seeking to unravel the complexities of financial and economic systems and predict market movements (Zhang & Ulku, 2019). Traditional approaches have focused mainly on time series analysis, fundamental economic indicators, and quantitative models to predict and decipher market trends, resulting in a lack of research regarding the impact of wider political, macroeconomic uncertainties on the stock market (Pal et al., 2020). These methods, while valuable, often struggle to capture the nuanced, rapidly changing landscape of global finance, particularly in an era of instantaneous information flow and complex market interrelations. Predictive models have become increasingly sophisticated as technology has progressed, including larger datasets and complex algorithms (Quinlan, 1986; Sako, Mpinda, & Rodrigues, 2022). The prevalence of Big Data, driven by advancements in information technology (Shen et al., 2022), has triggered a transformative shift revolutionizing data usage from traditional time series to encompassing diverse forms such as text data, thereby enhancing decision-making processes in financial markets (Rizkiana, Sari, Hardjomidjojo, Prihartono, & Yang, 2019; Tian, Zhou, & Zhang, 2020; Xing, Cambria, & Welsch, 2018; Yuanyuan, Kumari, Ilyas, Bhayo, & Marwat, 2023).

Investor decision-making is influenced by both current events and expectations about future developments. In this context, negative news[1] sentiment plays a significant role in shaping investor behavior, especially among international investors,[2] who tend to be more sensitive to uncertainty and market volatility.

As a result, the financial sector is increasingly leveraging text data to gain deeper insights into market behavior, sentiment, and emerging trends (Balaneji et al., 2024). This shift toward integrating unstructured, noisy text data such as news articles (Chari, Desai, Borde, & George,

* Corresponding author.
E-mail addresses: mahatibrahim@ihu.edu.tr (M.M. Ibrahim), asad.khan@ihu.edu.tr (A.U. Islam Khan), muhittin.kaplan@ihu.edu.tr (M. Kaplan).

[1] Plunge in Lira, Turkey's Currency, Fuels Fears of Financial Contagion - The New York Times.

[2] Plunge in Lira, Turkey's Currency, Fuels Fears of Financial Contagion - The New York Times.

https://doi.org/10.1016/j.bir.2025.06.013
Received 28 October 2024; Received in revised form 30 June 2025; Accepted 30 June 2025
Available online 1 July 2025

2023; Chatchawan, 2021; Jin, Chen, & Yang, 2024; Krinitz, Alfano, & Neumann, 2017; Pröllochs, Feuerriegel, & Neumann, 2015, 2016), social media posts (Bollen, Mao, & Zeng, 2011; Botchway, Jibril, Oplatková, Chovancová, & McMillan, 2020), and financial reports, into predictive models represents a frontier in financial analysis, enabling more dynamic and context-rich assessments of market conditions. However, the complexity of unstructured data deters many researchers, resulting in a nascent body of literature employing this methodology.

Similarly, sentiment analysis has gained significant momentum in various social science disciplines, particularly in the field of economics. In one such study, González and Cruz Tadle (2020) demonstrate how central banks can influence market expectations through the sentiment expressed in their press releases, underlining the critical role of clear communication in financial policy. The widespread use of central bank statements for sentiment analysis is further substantiated by a plethora of ongoing research efforts throughout the world. Studies by Kahveci and Odabaş (2016), Omotosho (2019), Kumari and Pandey (2025), Lee, Wang, Hsieh, and Chen (2023), and Picault, Pinter, and Renault (2022), all contribute to this increasing body of work. These studies emphasize the role of central bank communication and its importance on sentiment analysis in deciphering complex economic dynamics and predicting financial market behavior. In a similar, researchers have used data from various news and social media sources. Djourelov et al. (2017) highlighted the evolving media landscape, and the impact of online competition on local newspapers. Their research underlines the potential consequences of media consolidation by indicating that decreased coverage of political news, a result of shrinking newsrooms, can have a substantial impact on investor information and behavior. Additionally, there is evidence from previous research regarding the connection between news sentiment and asset prices. Pyo & Kim (2012) demonstrated a positive correlation in Korea, with investor optimism leading to higher stock prices and lower volatility. Johnman et al. (2014) suggested similar findings for the FTSE 100 index, with positive sentiment reducing volatility even without impacting returns. Öztürk (2022) extended this concept to the cryptocurrency market, highlighting the predictive power of Twitter sentiment on Bitcoin returns and trade volume. Notably, the influence of overall sentiment seemed stronger than that of individual accounts.

Despite these advancements, a critical gap exists in research combining local and international news sentiment analysis to predict the Turkish stock market. Previous studies are based on either financial reports (Atak, 2023), or Twitter only (Gumus & Sakar, 2021). This study aims to bridge this gap by leveraging Natural Language Processing (NLP) techniques and sentiment analysis on international news sources, local news outlets, and Twitter. Sentiment from both global news media (The Economist, the New York Times, The Guardian) and domestic ones (Yeni Şafak and Local Tweets) were taken into account to analyze the Turkish stock market. Moreover, to enhance the accuracy and interpretability of the predictions, the study employs advanced explainable AI (XAI) techniques alongside state-of-the-art natural language processing using Financial Bidirectional Encoder Representations from Transformers (FinBERT) transformer models. FinBERT, a variation of BERT specifically fine-tuned for financial text, offers superior performance in understanding and analyzing nuances of financial language. Its high capacity to comprehend entire texts contextually and compute sentiment as continuous values, rather than binary classifications, supports our research objectives.

Transformers, including FinBERT, offer significant advantages over classical models by combining a self-attention mechanism, parallelization, and positional encoding. These features are particularly beneficial when working on projects that involve semantic feature extraction from large datasets (Vaswani et al., 2023). Unlike traditional neural networks such as Long-Short Term Memory (LSTM) and Convolutional Neural Network (CNN) components, which rely on sequential analysis and can be limited in their effectiveness, transformer models excel at capturing complex relationships within text data (Heaton et al., 2017).

The remainder of the paper is structured as follows: Section 2 presents a review of the related literature, providing context and background for the study. Section 3 describes the data sources and the methodology employed, detailing the analytical framework used in the research. Section 4, presents the results and discussion of classical supervised machine learning models, analyzing their performance and implications. Section 5 focuses on the application of explainable AI techniques, offering insights into the interpretability of the models. Finally, the paper concludes with Section 6, which summarizes the findings and suggests directions for future research.

2. Related literature

Sentiment analysis has become a crucial tool in understanding market dynamics (Nabeen et al., 2022; Nyakurukwa & Seetharam, 2023). Case and Clements (2021) emphasize the importance of sentiment analysis in extracting subjective, speculative information from media, which can influence investment decisions and asset prices. They utilized contextual word representations from pre-trained BERT language models to analyze Wall Street Journal articles and found that financial news sentiment significantly predicts daily returns, while longer-term economic news sentiment is more relevant for monthly returns. On the other hand, Gu et al. (2021) constructed a financial news sentiment lexicon and an auxiliary lexicon specific to the financial domain. They introduced a Comprehensive Sentiment Index (CSI) that incorporates both textual sentiment and structured stock market data. Their comparative analysis of GARCH and LSTM models revealed that LSTM models, augmented with these sentiment indices, outperform traditional approaches in predicting stock returns.

The scope of sentiment analysis in financial markets extends beyond general news to specific topics of global concern. Fedorova and Iasakova (2024) investigate the impact of climate change news on US stock indices, employing a pre-trained bidirectional FinBERT neural network for sentiment analysis. They examined the influence of five main news topics (finance and politics, natural disasters and consequences, industrial sector and innovations, activism and culture, and the coronavirus pandemic) on stock index dynamics. The study reveals that these topics significantly influence the financial market. Other studies focus on the predictive power of social media sentiment. For instance, Jiang et al. (2021) explore the relationship between Twitter sentiment and stock returns using a deep learning framework that integrates both textual and visual information from tweets. The findings suggest that incorporating multimodal social media data can enhance stock return prediction accuracy. Similarly, Zhang and Ulku (2019) propose a hybrid model combining LSTM networks with attention mechanisms to capture both long-term dependencies and key information in financial news. Their approach demonstrates improved accuracy in stock price prediction compared to traditional methods.

Beyond text-based sentiment analysis, Chen et al. (2023) investigates the impact of audio sentiment in earnings conference calls on stock price movements. The study highlights the importance of considering multiple modalities of communication in assessing market sentiment. On the other hand, the geographical scope of sentiment analysis in financial markets is also diverse. While many studies focus on US markets, some studies explore other regions. For example, Li et al. (2020) examines the Chinese stock market, developing a sentiment-aware stock trend prediction model that incorporates both market data and news sentiment. Similarly, methodological innovations are evident across the literature. Sezer and Ozbayoglu (2021) employed capsule networks for financial sentiment analysis, demonstrating improved performance over traditional convolutional neural networks. Similarly, Wang et al. (2023) propose a multi-task learning framework that predicts stock movement and classifies sentiment simultaneously.

The impact of major global events on market sentiment is also a significant area of study. Several researchers examine the effects of the COVID-19 pandemic on stock market sentiment and returns,

highlighting the importance of sentiment analysis during periods of high uncertainty (e.g., Liu, Pan, & Yu, 2019). Meanwhile, other studies explore approaches that are more expansive to sentiment analysis. For instance, Xu et al. (2023) proposes a multi-modal framework combining textual and visual information from social media. Additionally, Zhang, Li, and Shen (2021) suggests an attention-based neural network for sentiment-aware stock prediction. Furthermore, other studies focused on specific sectors or events. Yang, Ma, and Wang (2021) analyzed the impact of sentiment on renewable energy stocks, while Johnson and Smith (2022) examined how geopolitical events influence market sentiment and returns.

The highlighted literature reveals a rapidly evolving field at the intersection of sentiment analysis, news media, and stock market prediction (Azimi and Agrawal, 2021). The integration of advanced machine learning techniques, particularly transformer-based models represent a significant advancement.

3. Data and methodology

3.1. Data and sources

A multifaceted approach was utilized, incorporating news articles, both from international and local sources alongside local social media data. Three prominent international publications were selected for news data: The Economist, The New York Times, and The Guardian. These newspapers were chosen due to their extensive global reach and reputation for in-depth coverage. Additionally, a local newspaper, Yeni Şafak, was included to offer a domestic perspective. To capture real-time sentiment surrounding Turkish markets, we further enriched the dataset with local Twitter posts. Due to limitations in accessing data from other local news outlets, Yeni Şafak and local Twitter posts were selected as proxies for the broader local media landscape. To ensure data relevance and minimize noise from extraneous content, we focused on the economic, finance, and business sections for all news sources.

The obtained data provides a dataset for analysis, spanning from January 1, 2015, to February 27, 2024. This timeframe aligns with our financial market data, which includes Turkish daily stock returns, opening prices, lows, highs, and trading volumes. All financial data is sourced from Yahoo Finance.

3.1.1. Data acquisition

Extracting large datasets is often a cumbersome task. Fortunately, modern software infrastructure simplifies this process. Most of our chosen news sources possess sophisticated Application Programming Interfaces (APIs) that facilitate smooth and systematic data retrieval. The New York Times, The Guardian, and Yeni Şafak offer user-friendly and free APIs that streamline data extraction. For The Economist, we utilized the university's subscribed database as well as manual data collection to access the latest relevant articles. Similarly, Twitter's academic API was utilized to extract relevant tweets. In conjunction with APIs, we employed Python libraries such as Requests and BeautifulSoup, as well as regular expressions, to further refine the extracted data. These tools assisted in removing unnecessary characters and extra spaces, ensuring a clean and organized dataset. The extracted data encompassed the publication date, article title, summary, full body, and URLs for each news article and tweet.

Finally, we consolidated the data from all sources and saved it in a single CSV file, which was readily prepared for the subsequent preprocessing stage. Fig. 1 illustrates the data extraction processes.

3.1.2. Data preprocessing

Raw text data obtained from various sources often presents challenges in the form of noise and irrelevant information. To address these issues, we implemented a series of data preprocessing and cleaning tasks using Pandas, SpaCy, and Neattext libraries. SpaCy and Neattext are powerful Natural Language Processing (NLP) tools specifically designed for text-cleaning and preprocessing tasks (Agbe, 2020; Honnibal & Montani, 2017).

The cleaning stage focused on removing extraneous elements such as URLs, HTML tags, excessive spaces, special characters, numbers, email addresses, hashtags, and improper quotations. The subsequent preprocessing stage involved several steps.

1. Lemmatization: reducing words to their base dictionary form.
2. Case conversion: converting text to lowercase to minimize textual variability and enhance consistency.
3. Punctuation removal: eliminating unnecessary symbols to streamline the data.
4. White space removal: ensuring textual coherence.
5. Normalization: standardizing the text format.
6. Spelling correction: improving the overall quality and readability of the text, minimizing potential confusion or misunderstanding.
7. After analyzing the removal and retention of stop-words, the decision was made to retain them as the analysis improved when the stop words were maintained.

Through this systematic preprocessing pipeline, we transformed the raw text data into a clean and structured format, ready for further analysis. Fig. 2 depicts the data-preprocessing flowchart and techniques employed.

After preprocessing the data, we carried out the following tasks: utilization the FinBERT transformer model for advanced feature extraction, topic modeling to identify key themes and narratives in the news articles, sentiment analysis to gauge positive, negative, and neutral

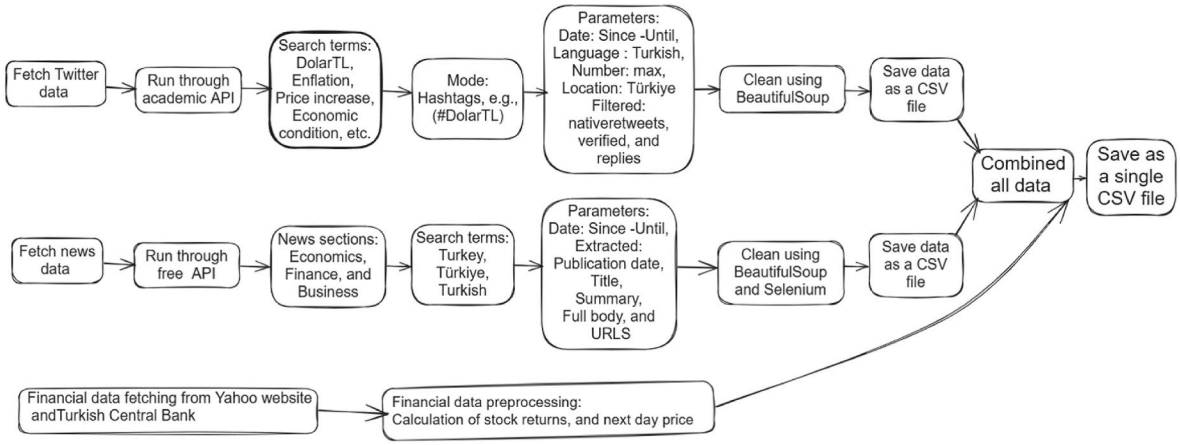

Fig. 1. Data extraction workflow.

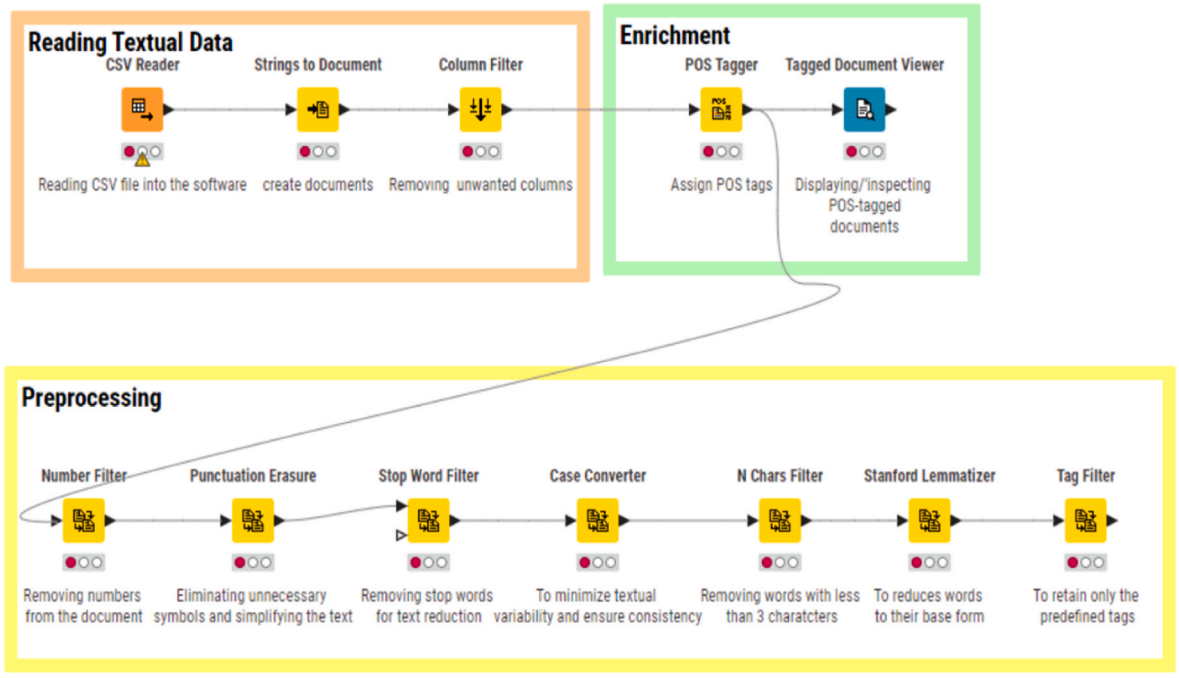

Fig. 2. Data preprocessing flowchart.

sentiment towards financial events, and named entity recognition to identify relevant entities such as companies, government policies, and economic indicators. Volatility and event indicators as well as market trends capture potential signals from news content. N-grams (unigrams, bigrams, and trigrams) capture the frequency and co-occurrence of words or phrases that are related to the stock market. Parts of speech classify words based on their syntactic and semantic roles in a sentence, such as verbs, adverbs, and adjectives, and their sentiment levels.

Our selection of FinBERT was driven by five compelling methodological advantages. First, FinBERT's specialized training on financial corpora allows for precise interpretation of domain-specific terminology that general language models often misinterpret. Second, its transformer-based architecture captures contextual relationships between words when analyzing complex financial statements where meaning depends heavily on surrounding text. Third, empirical comparisons demonstrate FinBERT's superior classification accuracy compared to traditional lexicon-based approaches like (Loughran & McDonald, 2016) when distinguishing sentiment variations. Fourth, FinBERT shows remarkable performance even with the limited Turkish financial dataset available to us, which can be seen as an advantage on account of the relatively smaller size of the Turkish market compared to major global exchanges. Finally, its proven capability to identify informational content in financial texts that alternative models miss ensures more comprehensive coverage of market-moving news events in our analysis framework (Huang et al., 2022).

To attest the robustness of our analysis, we also employed the Loughran-McDonald dictionary, which also offers five methodological strengths for financial text analysis. First, this lexicon was specifically developed to address misclassification issues found when general-purpose dictionaries are applied to financial contexts. Second, the dictionary categorizes financial terminology into distinct sentiment categories that effectively capture the tone of corporate documents and financial news. Third, empirical validation demonstrates its effectiveness in measuring sentiment when examining market reactions, trading volumes, and volatility patterns. Fourth, the dictionary shows consistent performance across various financial contexts including fraud detection, material weakness identification, and unexpected earnings analysis. Finally, its application requires minimal computational resources compared to transformer-based models, which makes it a practical complementary approach for verifying sentiment analysis results in financial market studies (Bodnaruk et al., 2015; Huang et al., 2022; Loughran & McDonald, 2011, 2016, 2024).

Sentiment Calculation Formula.

Given a text input T, we calculated the sentiment score S(T) as follows:

$S(T) = f\,(M\,(E\,(T\,[:512])))$Where:

T [:512] represents the first 512 characters of the input text T.

E is the encoding function that converts text to tokens.

M is the BERT model function that processes the encoded tokens.

f is a function that maps the model output to a sentiment score (1–5).

In our implementation:

$\mathbf{f(x)} = \text{argmax}(x) + 1$.

The sentiment scoring function processes text in chunks of 512 characters due to the FinBERT model's input size limitation. For texts longer than 512 characters, we iteratively applied the sentiment analysis to successive 512-character segments to ensure that the entire text was analyzed, regardless of its length. Thus, each text was split into consecutive 512-token chunks due to model input constraints. The system processed the first 512-token chunk completely before continuing to the next 512-token chunk, until the whole text is processed. Sentiment was then computed for each chunk, and the final sentiment score was obtained by combining or aggregating the scores across all chunks of each row.

Using a 1–5 scale for sentiment scores offers greater granularity compared to a 0–1 scale. This allows capturing subtle differences between slightly positive, neutral, and slightly negative sentiments better (Vaswani et al., 2023). The integer scores are also more intuitive for human interpretation, between 1 (very negative) and 5 (very positive), facilitating easier analysis and reporting of results (Goldberg Y. 2017).

3.2. *Methodological framework*

Predicting stock market trends has been a formidable task in financial analysis, traditionally approached using statistical methods like ARIMA (Autoregressive Integrated Moving Average) and econometric models such as GARCH (Generalized Autoregressive Conditional

Heteroskedasticity) (Tsay, 2005). While these methods have been instrumental in understanding and forecasting time series data, they are often constrained by assumptions of linearity and stationarity, making them less effective in capturing the complex and non-linear nature of financial markets (Atsalakis & Valavanis, 2009).

In response to the limitations of traditional methods, the development and application of Machine Learning (ML) models have emerged as powerful tools for stock market prediction (Hu et al., 2021). ML models have demonstrated superior performance by leveraging large datasets and identifying intricate patterns in data without the stringent assumptions required by classical methods (Cortes & Vapnik, 1995; Breiman, 2001). Given the potential of ML in financial forecasting, the selection of appropriate algorithms is crucial to ensure robust and reliable predictions. In this study, a diverse array of machine learning techniques, ranging from individual models to ensemble-based approaches have been employed, each offering distinct advantages in capturing the dynamics of the Turkish stock market (T. Chen & Guestrin, 2016; W. Chen, Song, & Hong, 2020). Alongside machine learning models, GARCH model has been employed as the foundational framework for our analysis. The GARCH model did not reveal statistically significant relationships between stock market returns and sentiment indicators derived from various news sources. The result of the GARCH model mean equation is presented in the Appendix (Table A.1).

The machine learning models utilized in this research include Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, and Artificial Neural Network. These models are widely recognized for their ability to learn complex patterns from data and provide interpretable insights into the underlying relationships (Hastie et al., 2009; Goodfellow et al., 2016). Logistic Regression, for instance, excels at modeling the probability of binary outcomes, making it well-suited for stock market classification tasks (Hosmer et al., 2013). Similarly, Decision Trees and their extensions have demonstrated remarkable performance in capturing non-linear interactions within the data, with the added benefit of easy interpretation (Breiman, 2001).

To further enhance predictive accuracy, ensemble-based algorithms have also been incorporated into the modeling framework. These methods, including Random Forest, XGBoost, Gradient Boosting, and AdaBoost, leverage the strengths of multiple individual models to produce more accurate, robust, and consistent predictions (Abbott, 2014; Rokach, 2010; Dietterich, 2000). Ensemble techniques have gained widespread popularity in finance and economics due to their ability to handle complex, non-linear relationships and their inherent robustness to outliers and noise in the data (Breiman, 2001; Chen & Guestrin, 2016). The utilization of this diverse array of algorithms serves for robustness checks and to enhance the predictive accuracy of stock market forecasts. The hyperparameters of these models were fine-tuned using GridSearchCV, a method that systematically evaluates model parameters to optimize performance. This approach ensured that each model was calibrated to effectively handle the complexities of the financial data, to achieve more reliable, and strong predictions (Chen & Guestrin, 2016.

Following the initial supervise machine learning prediction phase, explainable AI (XAI) techniques were employed to shed light on the rationale behind the model's predictions. The Shapley Additive exPlanations (SHAP) library was used specifically for its user-friendliness and ability to provide human-interpretable explanations of the model's features (Lundberg & Lee, 2017). Understanding the local explanations and importance of these features allows for a deeper comprehension of how media discourse influences market movements.

3.2.1. Evaluation metrics

The evaluation metrics include Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Accuracy, Precision, Recall, F1 Score. The Receiver Operating Characteristic (ROC) has also been reported, as well as specificity and sensitivity. These are detailed as follows:

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

$$RMSE = \sqrt{\Sigma_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{n}\right)^2}$$

$$MAE = \frac{1}{N}\sum\nolimits_{i=1}^{n}|y_i - \hat{y}_i|$$

Where y_i is the actual value and $\hat{y}_i$ is the predicted value of the i^{th} observation and N is the total number of observations.

When evaluating the accuracy of regression models, two key metrics emerge: Mean Absolute Error (MAE) and Mean Squared Error (MSE).[3] Both metrics measure the discrepancies between predicted and actual values but do so in fundamentally different ways. MAE focuses on the average magnitude of errors, disregarding their direction. It calculates the absolute difference between each predicted and actual value, sums all these differences, and then divides by the total number of predictions. This approach gives equal weight to all errors, regardless of whether they are over- or under-predictions. MAE is robust to outliers, as a single large error cannot significantly skew the final value. However, it can be difficult to interpret because it lacks a natural scale comparable to the data itself. MSE, on the other hand, amplifies the influence of larger errors. It squares the difference between each predicted and actual value, essentially penalizing deviations more severely. This leads to a higher overall error measure when large errors are present, highlighting their detrimental impact on model performance. While MSE allows for easier interpretation in the same units as the data, it is sensitive to outliers, potentially inflating the overall error due to a few extreme cases.

To address this sensitivity, Root Mean Squared Error (RMSE) emerges as a compromise. RMSE is the square root of MSE, bringing the metric back to the original scale of the data while retaining some emphasis on larger errors. This offers a more balanced interpretation compared to both MAE and MSE, making it a popular choice for evaluating regression models in several scenarios. RMSE exaggerates the importance of greater error values but still maintains a scale comparable to MAE, facilitating easier comparison between these two metrics and enhancing our understanding of model performance.

On the contrary, when evaluating the accuracy of classification models, the key metrics that emerge are Accuracy, Precision, Recall, F1 score, Sensitivity, Specificity, and Receiver Operating Characteristic (ROC) curve.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{True Positives (TP)} + \text{True Negatives (TN)} + \text{False Positives (FP)} + \text{False Negatives (FN)}}$$

[3] https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives(FP)}}$$

$$\text{Sensitivity (Recall)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Specificity (True Negative Rate)} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive(FP)}}$$

Precision and recall are essential classification metrics, primarily defined using the following terms.

- TP (True Positive): Instances that are true and correctly classified as true.
- TN (True Negative): Instances that are false and correctly classified as false.
- FP (False Positive): Instances that are false but incorrectly classified as true.
- FN (False Negative): Instances that are true but incorrectly classified as false.

Accuracy, however, is one of the most widely employed metrics in machine learning evaluations (Rainio et al., 2024). Its popularity stems from its simplicity and ease of understanding, facilitating straightforward model comparisons. However, despite its user-friendly nature, it may not always be the most appropriate evaluation metric, as it oversimplifies complex scenarios. Therefore, in this study, accuracy has been complemented with additional metrics, such as precision, recall, and F1 score. These metrics are predominantly utilized in binary classifications but can also be applied to other classification tasks.

Precision represents the percentage of correctly labeled instances out of all positively labeled instances. In other words, it indicates the accuracy of positive class predictions among all instances labeled as positive or belonging to a specific class. On the other hand, Recall, also known as Sensitivity or True Positive Rate, denotes the percentage of correctly labeled positive instances out of all actual positive instances. It reveals how effectively the model captures instances belonging to a specific class, whether positive or negative. Precision and recall are often used together because they offer different perspectives on the same model, providing insights into its performance and highlighting areas for improvement.

The F1 score, a combination of precision and recall, is often preferred by researchers. It represents the harmonic mean of recall and precision, offering a balanced metric that considers both false positives and false negatives. In binary classification problems, the Receiver Operating Characteristic (ROC) curve serves as a graphical assessment tool. It plots the true positive rate (sensitivity) against the false positive rate (1 - specificity). The Aea Under this Curve (AUC) offers a valuable metric for comparing the performance of different prediction models. As it is independent of scaling, AUC allows for objective ranking of various methods. One major advantage of AUC is its classification-threshold invariance.[4] It measures the inherent predictive power of a model irrespective of the chosen classification threshold, unlike metrics like F1 score or overall accuracy, which can be sensitive to threshold selection. AUC determines the overall unbiased predictive power of a classifier by accounting for potential biases. A perfect classifier achieves an AUC of 1, while a model with no predictive power scores 0.5 (equivalent to random chance). Real-world scenarios typically present AUC values between these extremes.

4. Results and decision

Tables 1 and 2 provide the performance of various classification models in predicting stock returns, evaluating them across several key metrics: Accuracy, Precision, F1 Score, Specificity, Sensitivity (Recall), and ROC AUC (Receiver Operating Characteristic Area Under the Curve). The analysis of the various classification models in Table 1 reveals important insights into the relationship between model performance and investment behaviors of risk aversion and risk-taking. In addition, Fig. 5 is a graphical illustration of the ROC AUC scores of the various algorithms used.

4.1. Ensemble methods and risk management

The ensemble methods, including Random Forest, Gradient Boosting, and XGBoost, consistently demonstrate high performance across metrics such as accuracy, precision, and specificity. This is crucial for risk-averse investors who prioritize minimizing losses over potential gains. High specificity indicates the models' ability to correctly identify negative instances (i.e., when a stock will not perform well), which is essential in avoiding bad investments. This aligns with the risk management strategies of conservative investors who prefer to avoid losses even at the cost of missing out on some profitable opportunities.

The ensemble methods' robustness, particularly their strong ROC AUC scores, further underscores their suitability for investors who seek to minimize false positives (unwarranted investments) and false negatives (missed profitable investments). These models' ability to aggregate multiple decision trees helps mitigate overfitting, ensuring stronger and increasingly reliable predictions in diverse market conditions-a key factor for long-term, risk-averse strategies. These algorithms have been the most widely used algorithms since 2000 for price completion. The Netflix[5] Prize, one of the most well-known ensemble competitions, offered a $1 million reward for improving the accuracy of its movie recommendation algorithm. Participants were tasked with reducing the Root Mean Square Error (RMSE) by at least 10 % compared to the existing model. The top-ranking teams, including the winner and runner-up, predominantly relied on model ensembles, combining hundreds of predictive models to achieve superior results. This competition underscored the power of ensemble methods in achieving significant predictive accuracy gains in complex tasks. Vorhies (2016) aptly highlighted the value of model ensembles in predictive analytics, particularly in competitive settings like Kaggle. To secure a top spot on the leaderboard or achieve a commendable ranking, it's essential to not only understand but also strategically apply ensemble methods.

4.2. Sensitivity, risk-taking, and speculative investments

Sensitivity (recall), on the other hand, is a critical metric for risk-taking investors who tend to identify and capitalize on all possible opportunities, even at the risk of incurring a few losses. Models like Neural Networks and Support Vector Machines, while not outperforming ensemble methods, show higher sensitivity, making them further suitable for aggressive investment strategies. These models' ability to detect true positives (successful investments) can be valuable in speculative markets, where missing out on a high-yield opportunity is more detrimental than making a poor investment.

However, the relatively lower specificity of these models indicates a higher likelihood of false positives, which corresponds to a higher risk of poor investment decisions. This trade-off between sensitivity and specificity reflects the inherent risk-return trade-off in financial investments, where higher potential returns often come with higher risks.

[4] https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce.

[5] https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data.

Table 1
Evaluation Metrics of Machine Learning Models with Control Variables (FinBERT model).

Model	Accuracy	Precision	F1 Score	Specificity	Sensitivity (Recall)	ROC
Decision tree	0.90[A]	0.88[B]	0.88[B]	0.91[A]	0.88[B]	0.89 [B]
Random forest	0.88[B]	0.87[B]	0.86[B]	0.90[A]	0.85[B]	0.97[A]
Logistic regression	0.81[B]	0.76[C]	0.79[C]	0.80[B]	0.83[B]	0.89[B]
Neural network	0.68[D]	0.62[D]	0.65[D]	0.68[D]	0.69[D]	0.77[C]
XGBoost	0.89[B]	0.86[B]	0.87[B]	0.90[B]	0.87[B]	0.95[A]
Support vector machine	0.84[B]	0.80[B]	0.82[B]	0.85[B]	0.83[B]	0.87[B]
kNearest neighbor	0.86[B]	0.83[B]	0.84[B]	0.87[B]	0.84[B]	0.94[A]
AdaBoost	0.76[B]	0.70[B]	0.74[B]	0.76[B]	0.77[B]	0.86[B]
Gradient boosting	0.91[A]	0.90[A]	0.89[B]	0.93[A]	0.88[B]	0.96[A]

Note: [A:] Greater than 90, [B:] Greater than 80, [C:] Greater than 70, [D:] Greater than 60.

Table 2
Evaluation Metrics of Machine Learning Models without Control Variables (FinBERT model).

Model	Accuracy	Precision	F1 Score	Specificity	Sensitivity (Recall)	ROC
Decision tree	0.86[B]	0.83[B]	0.83[B]	0.88[B]	0.83[B]	0.86[B]
Random forest	0.86[B]	0.84[B]	0.84[B]	0.88[B]	0.84[B]	0.95[A]
Logistic regression	0.57[E]	0.49[F]	0.57[E]	0.49[F]	0.67[D]	0.55[E]
Neural network	0.66[D]	0.58[E]	0.66[D]	0.60[D]	0.75[C]	0.73[C]
XGBoost	0.81[B]	0.76[C]	0.78[C]	0.82[B]	0.79[C]	0.90 [A]
Support vector machine	0.78[C]	0.72[C]	0.76[C]	0.76[C]	0.81[B]	0.83[B]
kNearest neighbor	0.85[B]	0.83[B]	0.83[B]	0.88[B]	0.82[B]	0.90 [A]
AdaBoost	0.67[D]	0.60[D]	0.64[D]	0.66[D]	0.69[D]	0.73[C]
Gradient boosting	0.86[B]	0.84[B]	0.83[B]	0.88[B]	0.83[B]	0.95 [A]

Note: [A:] Greater than 90, [B:] Greater than 80, [C:] Greater than 70, [D:] Greater than 60, [E:] Greater than 50.

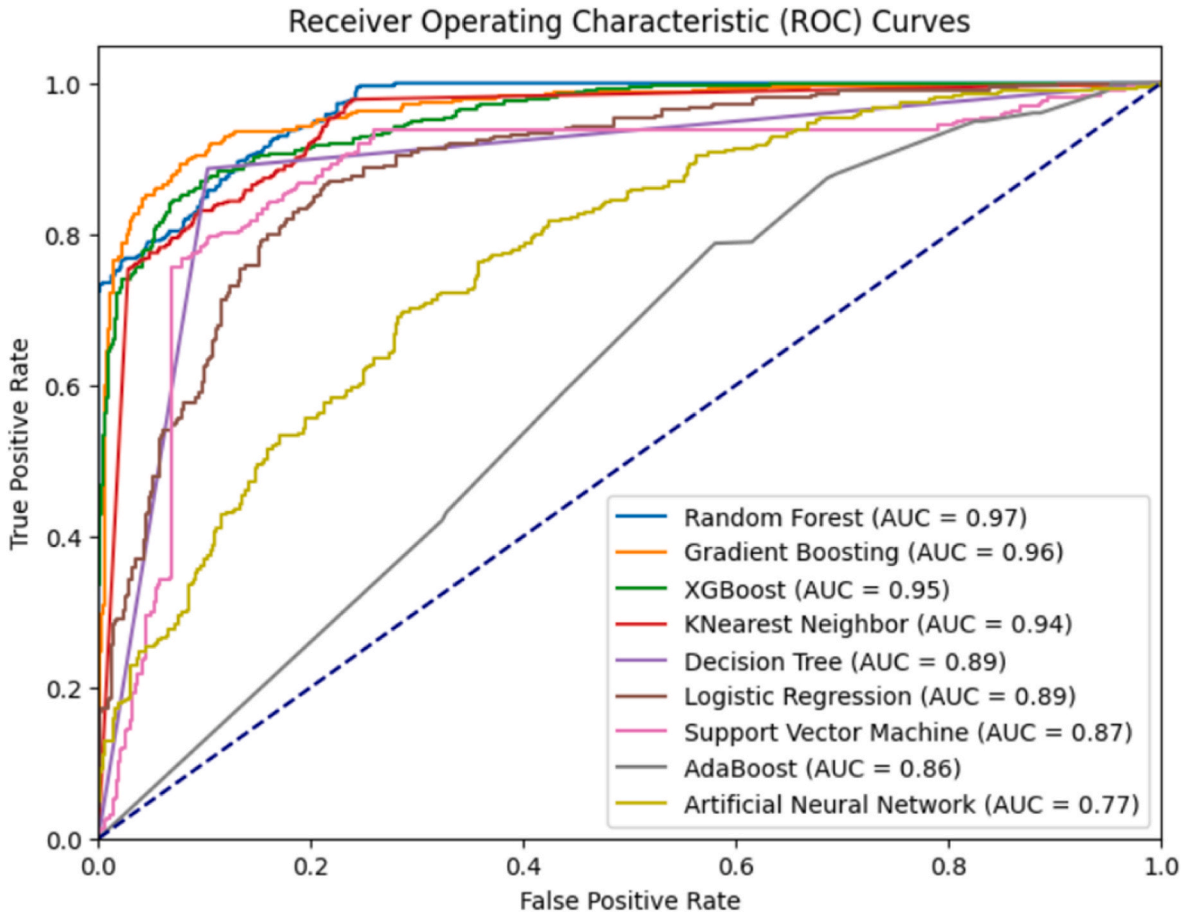

Fig. 3. Combined ROC curve scores.

4.3. F1 score and balanced investment strategies

The F1 score, which balances precision and recall, emerges as a crucial metric for investors seeking a balanced approach. The strong F1 scores of the ensemble methods and k-Nearest Neighbors suggest their ability to provide a balanced investment strategy that mitigates risks while still capturing opportunities. This suggest that investors with a moderate risk appetite might prefer these models as they offer a more reliable basis for decision-making without leaning heavily towards risk-averse or risk-taking behaviors.

4.4. Financial, economic implications and model selection

Economically, the superiority of ensemble methods in terms of F1 score and ROC AUC indicates their potential for application in algorithmic trading, where both risk management and opportunity capture are critical. The consistent performance across metrics suggests that these models are well-suited for dynamic financial environments, providing a reliable tool for forecasting in both bullish and bearish markets.

In contrast, the weaker performance of linear models like Logistic Regression highlights the challenges of employing simple linear approximations to the complex, non-linear nature of financial markets. The analysis emphasizes the importance of model selection in financial forecasting, considering both statistical performance and economic implications. Selecting models that align with investment strategies whether focused on risk aversion, risk-taking, or a balanced approach allows investors to navigate the complexities of financial markets more effectively, leading to further informed decision-making.

Fig. 3 presents the combined ROC curves of all the models, arranged in descending order of performance, as indicated by the legend. Models with higher performance have curves that are positioned closer to the top-left corner of the plot, indicating better predictive accuracy. The higher the curve, the more effective the model is at distinguishing between classes.

To ensure further robustness of the evaluation, we applied 10-fold cross-validation, a widely used technique that balances bias and variance by dividing the data into 10 equal parts, training on 9 and testing on 1 in rotation. This method provides a more reliable estimate of model generalizability. The cross-validation results closely align with the main test outcomes, confirming that the better-performing models are not only accurate but also consistent across different data subsets, while models with lower test performance also exhibited lower mean F1-scores and higher variability during validation. The result of the 10-fold cross-validation is presented in the Appendix (Figure A.2).

In addition to 10-fold cross-validation, we implemented walk-forward validation to evaluate how the models perform over time. The results show that models such as Gradient Boosting, XGBoost, and Random Forest perform well in both validation setups. Models such as Support Vector Machine and Artificial Neural Network perform worse in the walk-forward test. This pattern shows that some models can fit static data well but struggle with time-dependent data. Walk-forward validation gives a more realistic test of model stability in sequential prediction tasks. The comparative results of the 10-fold cross-validation and the walk-forward validation are shown in the Appendix (Figure A.3).

As evidenced in Table 3, the ensemble-based algorithms (the Random Forest and XGBoost) along with Decision Tree demonstrate superior performance metrics in the implementation of Loughran-McDonald sentiment analysis. Conversely, AdaBoost exhibited suboptimal efficacy in this analysis. Gradient Boosting and K-Nearest Neighbor algorithms achieved moderate performance outcomes. The models demonstrating the least favorable performance metrics include Artificial Neural Networks, Support Vector Machines, and Logistic Regression, which yielded relatively lower accuracy and predictive validity. The overall performance patterns of the Loughran-McDonald-based sentiment analysis are similar to those of the main FinBERT transformer-based model. However, it performs relatively worse than the FinBERT model in this analysis.

Along with the classification models, linear machine learning models were applied to the log of the stock market price to evaluate their performances using metrics such as Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error as shown in Table 4. Among the models, Linear Regression was observed to perform the best with the least errors, indicating that it makes relatively smaller errors and delivers more accurate predictions compared to the other models. On the other hand, the Decision Tree produced the highest errors, suggesting that it has larger discrepancies in its predictions. The Random Forest model performed slightly better than the Decision Tree but still falls short of Linear Regression in terms of accuracy. The Neural Network model falls in between, showing moderate accuracy.

5. Explainable artificial intelligence (XAI)

The rapid advancement of AI has given rise to increasingly complex and opaque models employed to address challenging real-world problems (Lundberg et al., 2020). Unlike their predecessors, such as time series ARIMA models, the complicated structures of these models often make them difficult to understand and operate. Decoding the source of their outputs, especially unexpected ones, becomes a monumental task, limiting our ability to effectively diagnose and remedy issues. This is where Explainable AI (XAI) steps in as a critical ally in the evolving landscape of AI models. XAI shines a light on the inner workings of these intricate models, demystifying their decision-making processes. It empowers researchers to understand the performance of the model as well

Table 4
Linear machine learning models.

Model	MAE	MSE	RMSE
Linear regression	0.0112	0.0002	0.0155
Decision tree	0.0168	0.0005	0.0226
Random forest	0.0121	0.0003	0.0166
Neural network	0.0151	0.0004	0.0198

Table 3
Classification performance metrics (Loughran-McDonald sentiment analyzer).

Model	Accuracy	Precision	Recall	F1 Score	Sensitivity (TPR)	Specificity (TNR)	ROC
Logistic regression	0.64 [D]	0.56 [E]	0.78 [C]	0.65 [C]	0.78 [C]	0.54 [E]	0.76 [C]
Random forest	0.88 [B]	0.86 [B]	0.88 [B]	0.87 [B]	0.88 [B]	0.89 [B]	0.93 [A]
Decision tree	0.88 [B]	0.85 [B]	0.87 [B]	0.86 [B]	0.87 [B]	0.89 [B]	0.88 [B]
XGBoost	0.89 [B]	0.87 [B]	0.89 [B]	0.88 [B]	0.89 [B]	0.91 [A]	0.93 [A]
KNearest neighbor	0.81 [B]	0.77 [C]	0.78 [C]	0.78 [C]	0.78 [C]	0.82 [B]	0.91 [A]
Support vector machine	0.66 [C]	0.59 [E]	0.72 [C]	0.65 [D]	0.72 [C]	0.62 [D]	0.72 [C]
Artificial neural network	0.66 [C]	0.58 [E]	0.80 [C]	0.67 [C]	0.80 [C]	0.56 [D]	0.75 [C]
AdaBoost	0.70 [D]	0.63 [D]	0.70 [D]	0.66 [D]	0.70 [D]	0.70 [D]	0.7 [C]
Gradient boosting	0.77 [C]	0.70 [D]	0.80 [C]	0.75 [C]	0.80 [C]	0.75 [C]	0.87 [B]

Note: [A:] Greater than 90, [B:] Greater than 80, [C:] Greater than 70, [D:] Greater than 60, [E:] Greater than 50.

as the impact of individual features or variables. Shapley values, in particular, provide a clear understanding of the contribution of each feature to the model's predictions within the dataset.

Introduced by Lundberg and Lee (2017), Shapley Additive Explanations (SHAP) has become a widely adopted framework for model interpretability across various fields. Building upon the classic Shapley values from game theory and their extensions (Lundberg et al., 2020), SHAP effectively links optimal credit allocation with local explanations, fostering model transparency and trust (Ali et al., 2023; Barredo Arrieta et al., 2020).

This study employs SHAP values to interpret predictions and understand the key features in news articles that drive model predictions, thereby enhancing transparency and trust. SHAP is more appropriate for research production compared to its predecessors, such as LIME (Local Interpretable Model-agnostic Explanations) introduced by Ribeiro et al. (2016). While LIME was one of the earliest well-known methods in the explainability area and is more suitable for large data processing and industrial production, this project aims to employ SHAP methods due to their advantages in research contexts. Fig. 4 indicates the obscurity of the conventional machine learning algorithms, while Fig. 5 shows the explainable AI's end process.

We only ran the explainable models of the two best-performing supervised classification models, the Gradient Boost and XGBoost based on the ROC curve result portrayed in Fig. 3. The Random Forest model provided results almost identical to the Gradient Boost leading us to therefore, only report the Gradient Boost to restrict the paper's length and avoid repetitive tasks. Positive SHAP values indicate contributions pushing predictions higher, while negative values suggest a decreasing effect. This approach ensures an accurate and interpretable explanation of the outcome (Moustakidis et al., 2023). The SHAP values, derived from cooperative game theory, represent the average marginal contribution of a feature across all possible feature combinations. This approach provides a unified method for interpreting model predictions, assigning each feature an importance value for a particular prediction. The integration of SHAP into our analytical framework represents a step towards increasingly transparent and accountable machine learning applications in financial prediction tasks, bridging the gap between model complexity and interpretability.

Fig. 6 illustrates feature importance analyses for high-performing machine learning models, XGBoost (on the left), and Gradient Boosting (on the right). Across all models, historical price data and control features, such as "Low," "Open," and "High" consistently emerge as the most significant predictors. However, the relative importance of other features exhibits notable variation among the different models.

The XGBoost algorithm, attributes greater importance to The Economist, Yeni Şafak, and the Guardian. On the other hand, Twitter data is seen as the least or the weakest predictor of the stock returns. The gradient boost algorithm switches the relative importance of Yeni Şafak and The Economist while other features largely remain the same as in that of the XGBoost algorithm.

The observed variations in feature importance across different machine learning models suggest that the choice of algorithm can influence the perceived relevance of input variables in predictive modeling. However, this comparative analysis between the two models indicates the interplay between algorithm selection and feature evaluation in the context of financial prediction tasks, which is crucial for interpretable outcomes in financial predictions.

Nevertheless, feature importance only indicates the overall significance of each feature in the model's predictions but does not provide insights into how individual features influence the probability of a positive prediction. This limitation is addressed by using Beeswarm plots and SHAP mean values, which help visualize and quantify the impact of each feature on the model's output, highlighting how features drive predictions in specific directions.

5.1. Gradient Boosting and XGBoost: mean SHAP analysis

Fig. 7 presents the mean SHAP values across all observations for the Gradient Boosting and XGBoost models, respectively. These visualizations elucidate the most salient features for each model, based on the mean of the absolute SHAP values to prevent the offsetting effect of positive and negative contributions. The analysis employs a bar chart representation, with each bar corresponding to a distinct feature.

As depicted in Fig. 7, the Gradient Boosting model reveals that The New York Times, The Economist, and Yeni Şafak exhibit the highest mean SHAP values of 0.27, 0.22, and 0.20, respectively. This indicates that these news sources, on average, contribute positively to the prediction of stock market returns. The Guardian and Twitter data demonstrate comparatively lower influence, with mean SHAP values of approximately 0.18 and 0.01, respectively.

In contrast, the XGBoost model presents a different hierarchy of feature importance. In this model, The Guardian emerges as the most influential predictor, with a substantial mean SHAP value of 0.63. This is followed by The New York Times and The Economist, mirroring the pattern observed in the Gradient Boosting model. Yeni Şafak and Twitter maintain relatively lower positions in terms of predictive power, analogous to their status in the Gradient Boosting analysis in Fig. 7.

The divergence in feature importance rankings between the two models underscores the complexity of the relationship between media sentiment and stock market dynamics. However, a consistent pattern emerges, with traditional news journals, particularly those with international reputations, demonstrating greater predictive power for stock returns compared to social media platforms like Twitter. This suggests that these established media outlets play a more significant role in shaping investor sentiment, emotions, and decision-making processes than social media discourse.

These findings contribute to our understanding of the interplay between media sentiment and financial markets, highlighting the differential impact of various sources of information on stock market behavior. Future research could explore the temporal stability of these relationships and investigate potential causal mechanisms underlying the observed correlations.

Next, Figs. 8 and 9 depict the Beeswarm visualization of all the SHAP values. On the y-axis, the values are grouped by features. For each group, the color of the points is determined by the feature value (higher feature values are redder while the lower values are bluer). Just like mean SHAP, the Beeswarm highlights important relationships. In fact, the features in Figs. 8 and 9 are ordered by mean SHAP.

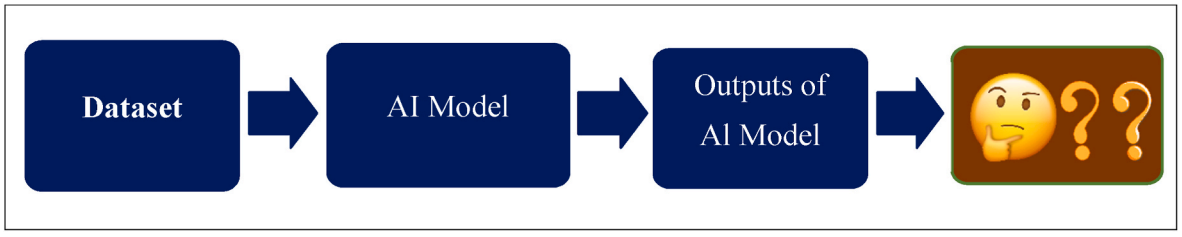

Fig. 4. The process of Artificial Intelligence Models without Explanation.

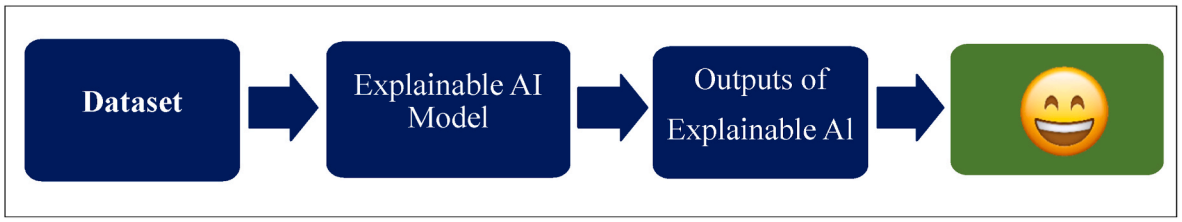

Fig. 5. The process of artificial intelligence models with explanation.

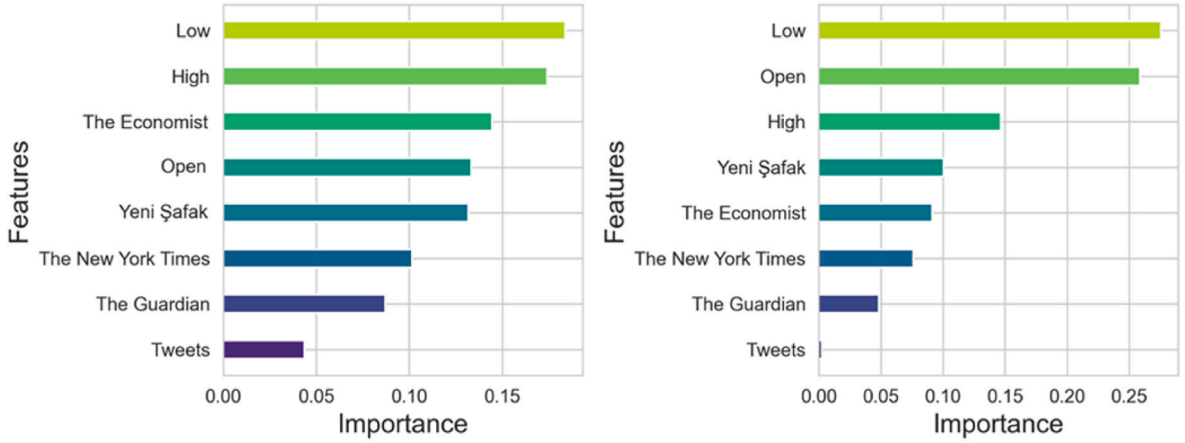

Fig. 6. Feature importance in XGBoost and gradient boosting models.

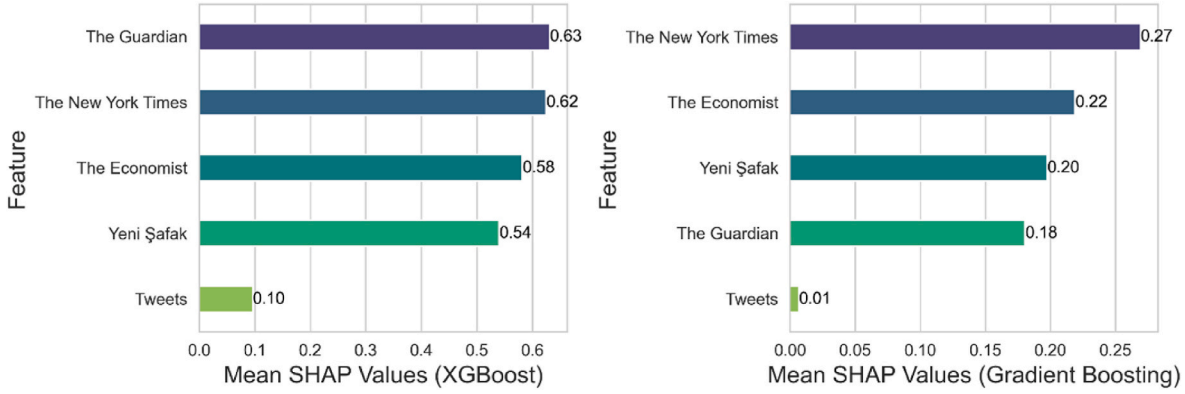

Fig. 7. Mean SHAP values of XGBoast, and gradient boosting models.

The Beeswarm plot of the Gradient Boosting model in Fig. 8 reveals the influence of the features on stock market predictions and investor sentiment. Traditional news outlets, The New York Times and The Economist, emerge as pivotal predictors with wide-ranging impacts (SHAP values ranging from approximately −1.0 to 1.2 for The New York Times). Their broader distribution of SHAP values indicates that these sources can sway predictions towards both positive and negative stock returns, reflecting the complex nature of financial news interpretation. Yeni Şafak and The Guardian show a more moderate range of impact (mostly between −1.5 and 1.0). The minimal influence of Tweets (SHAP values tightly clustered around 0) underscores that established journalism outweighs social media sentiment in predicting market movements in this model. The correlation between higher feature values and positive SHAP values for major news outlets suggests that positive sentiment tends to predict upward market movements, reflecting the intricate relationship between media sentiment and investor psychology. However, the presence of negative SHAP values, even for high-value features, highlights the sensitivity of the market to nuanced interpretations of news.

The XGBoost Beeswarm plot in Fig. 9 reveals a subtle shift in the relative importance and impact of news sources compared to the Gradient Boosting model. The Guardian now exhibits the most substantial influence (SHAP values ranging from approximately −3.0 to 3.0), surpassing The New York Times (range of about −3.0 to 2.5). This contrasts with the previous model where The New York Times had the strongest impact. The Economist maintains a significant but more concentrated effect (mostly between −2.0 and 2.0), while Yeni Şafak shows a broader distribution than before (from about −4.0 to 2.0). Tweets continue to have the least variation, though with a slightly larger impact than in the Gradient Boosting model. All major news sources in the XGBoost model demonstrate further extreme SHAP values, suggesting a higher sensitivity to media sentiment. The color distribution indicates a stronger correlation between high feature values (red points) and positive SHAP values across sources, implying that positive sentiment consistently predicts upward market movements in this model.

The Beeswarm visualizations for both the Gradient Boosting and

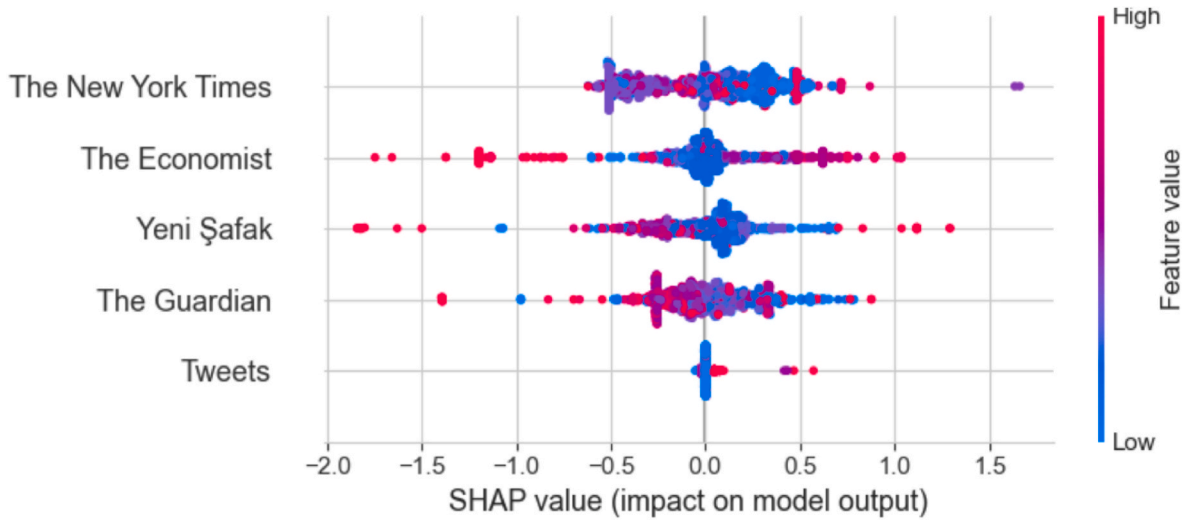

Fig. 8. Beeswarm Plot of Gradient Boosting model.

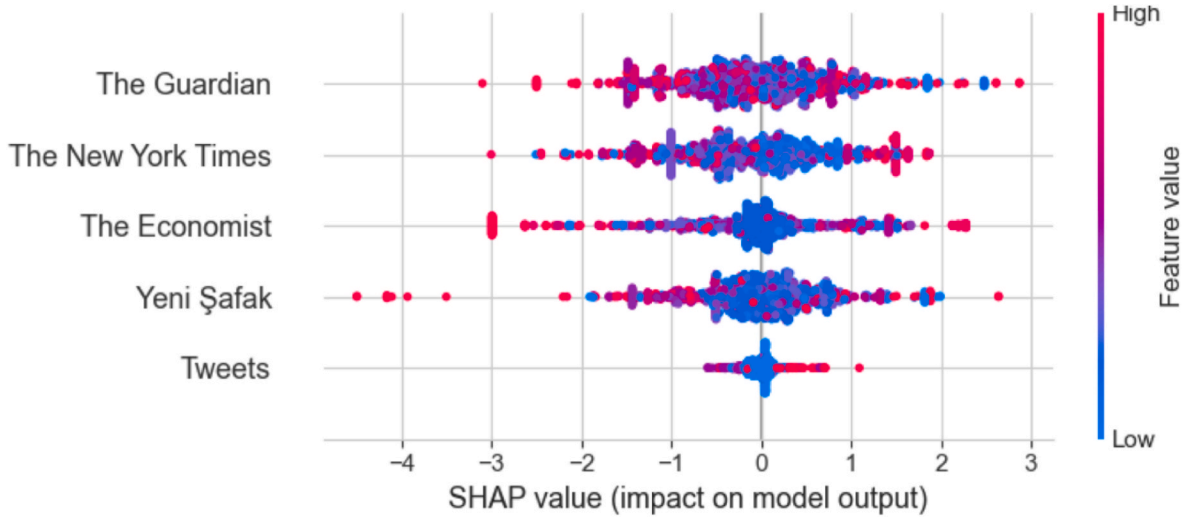

Fig. 9. Beeswarm Plot of Gradient Boosting model.

XGBoost models illuminate the crucial role of news sentiment in stock market prediction and investor decision-making processes. These plots vividly demonstrate how different information sources are weighted in the collective judgment of market participants, with established news media consistently emerging as dominant predictors. While both models emphasize the pronounced influence of reputable news outlets, the XGBoost model reveals a more nuanced picture. Its heightened impact and wider ranges of SHAP values suggests an enhanced sensitivity to subtle variations in media sentiment, potentially capturing more complex relationships between news coverage and market behavior. The comparison between these models not only reinforces the significance of media sentiment analysis in financial forecasting but also highlights the potential for machine learning algorithms to uncover increasingly sophisticated patterns in the interplay between information dissemination and market reactions.

6. Conclusion

Reflecting on the findings, it is evident that the complex interplay of media discourse and investor sentiment on the Turkish stock market is best captured by sophisticated ensemble models. These models, such as Gradient Boosting, XGBoost, and Random Forest, excel in their ability to decode the emotional undercurrents that drive market movements. The narratives woven through media reports often sway investor behavior. Stories of economic turmoil, financial risks, speculations, and optimism can stir emotions that ripple through trading floors. The highest predictive accuracy of these models suggests they are particularly attuned to these nuances, effectively translating the collective mood into market forecasts.

Linear models, like Linear Regression, also show a remarkable ability to predict market outcomes with precision, reflecting a more measured, perhaps rational, aspect of investor behavior. Yet, the superior performance of ensemble methods hints at the emotional complexity of the market, where sentiments, fear, and greed interplay, making simplistic approaches less effective. The Explainable AI methods reveal that the traditional international news media affects the Turkish stock market while local news from (Yeni Şafak and the Twitter), has lesser predictive power in comparison. These findings highlight the importance of the role of news investors' sentiment in understanding market dynamics, reminding us that beneath the cold numbers lies a very human market, swayed by emotions as much as by economic fundamentals.

6.1. Limitations of the study

The scope of this study was constrained by the use of a single local newspaper, Yeni Şafak, and Twitter, primarily due to the inaccessibility of historical data from other local news sources. This limitation restricts the generalizability of the findings, as the selected media outlets may not fully capture the broad spectrum of public sentiment and information flows that influence stock market dynamics. Future research should consider expanding the range of data sources by incorporating additional local newspapers and diverse social media platforms to provide a more comprehensive analysis of how various media channels impact different stock indices.

Author contributions

Conceptualization, Mahat Maalim Ibrahim; **methodology**, Mahat Maalim Ibrahim and Asad Ul Islam Khan; **software**, Mahat Maalim Ibrahim; **validation**, Mahat Maalim Ibrahim and Asad Ul Islam Khan; **formal analysis and investigation**, Mahat Maalim Ibrahim; **resources**, Asad Ul Islam Khan and Muhittin Kaplan; **data curation;** Mahat Maalim Ibrahim; **writing—original draft preparation**, Mahat Maalim Ibrahim; **writing—review and editing**, Asad Ul Islam Khan and Muhittin Kaplan; **visualization**, Mahat Maalim Ibrahim; **supervision**, Asad Ul Islam Khan; funding **acquisition**, Asad Ul Islam Khan and Mahat Maalim Ibrahim; **project administration**, Asad Ul Islam Khan.

Acknowledgements

This work was supported by the Research Fund of Ibn Haldun University. Project Number: 2236.

Appendix

Table A.1
GARCH (2,1) Model Mean Equation

Variable	Coefficient	Std. Error	t-Statistic	p-Value	Interpretation
Const (μ)	0.001352	0.000879	1.537	0.124	Intercept; not significant
Lagged return	0.007746	0.02218	0.349	0.727	AR (1); not significant
The economist	−0.000999	0.001514	−0.660	0.509	Neg. Impact; not significant
Yeni şafak	0.000505	0.000441	1.146	0.252	Pos. Impact; not significant
The guardian	0.001335	0.001315	1.015	0.310	Pos. Impact; not significant
The New York Times	0.000392	0.000525	0.748	0.454	Pos. Impact; not significant
Tweets	0.000090	0.000204	0.443	0.658	Pos. Impact; not significant

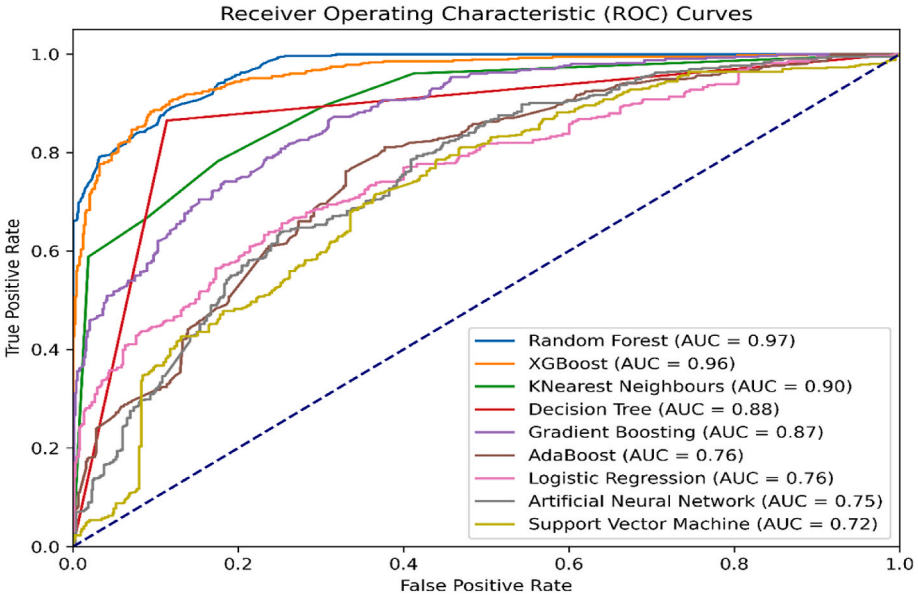

Fig. A.1. ROC curve from the Loughran-McDonald sentiment-based result.

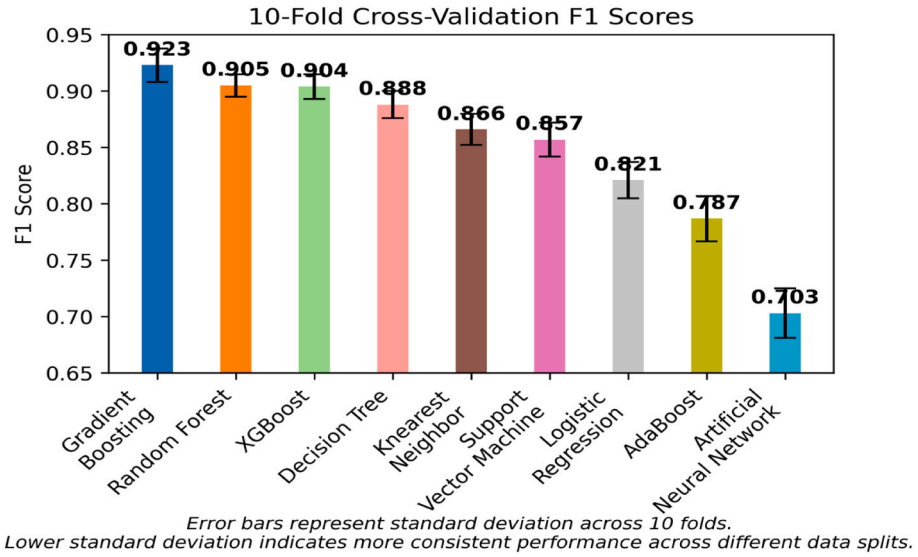

Fig. A.2. Cross validation of the FinBERT model-based in Table 1.

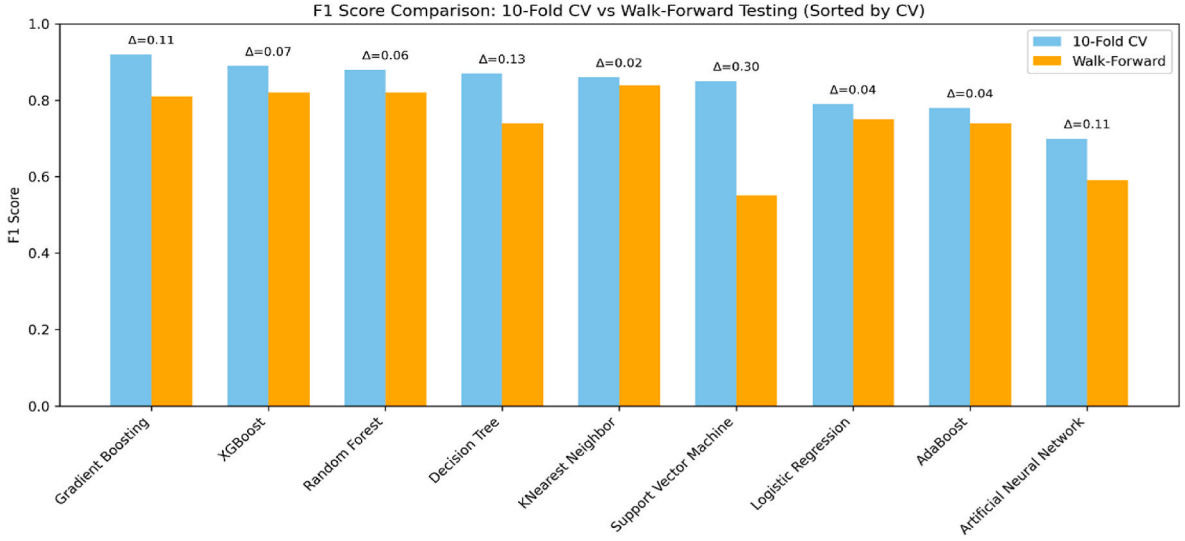

Fig. A.3. 10-Fold versus Walk-Forward Testing.

References

Agbe, J. E. (2020). Neattext. *GitHub* Version 0.1.3. https://github.com/Jcharis/neattext.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion, 99*, Article 101805. https://doi.org/10.1016/j.inffus.2023.101805

Atak, A. (2023). Exploring the sentiment in Borsa Istanbul with deep learning. *Borsa Istanbul Rev., 23*(Supplement 2), S84–S95. https://doi.org/10.1016/j.bir.2023.12.010

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques – Part II: Soft computing methods. *Expert Systems with Applications, 36*(3), 5932–5941.

Azimi, J., & Agrawal, M. (2021). Predicting stock market trends using machine learning and deep learning techniques. *J. Financ. Data Sci., 3*(4), 1–19.

Balaneji, F., Maringer, D., & Spasić. (2024). The power of words: Predicting stock market returns with fine-grained sentiment analysis and XGBoost. In K. Arai (Ed.), *Intelligent systems and applications. IntelliSys 2023. Lecture notes in networks and systems, 822.* Cham: Springer. https://doi.org/10.1007/978-3-031-47721-8_39.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Garcia, S. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis, 50*(4), 623–646. https://doi.org/10.1017/S0022109015000448

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *J. Comput. Sci., 2*(1), 1–8.

Botchway, R. K., Jibril, A. B., Oplatková, Z. K., Chovancová, M., & McMillan, D. (2020). Deductions from a sub-saharan african bank's tweets: A sentiment analysis approach. *Cogent Eco. Finance, 8*(1). https://doi.org/10.1080/23322039.2020.1776006

Case, J., & Clements, A. (2021). The impact of sentiment in the news media on daily and monthly stock market returns. Data Mining. AusDM 2021. In *Communications in computer and information science, 1504*Singapore: Springer. https://doi.org/10.1007/978-981-16-8531-6_13.

Chari, S., Desai, P. H., Borde, N., & George, B. (2023). Aggregate news sentiment and stock market returns in India. *JRFM, 16*(8), 1–18.

Chatchawan, S. (2021). Can sentiment from news headlines explain stock market returns? Evidence from Thailand. *Thammasat Rev., 24*(1), 317–333.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, W., Song, W., & Hong, Z. (2020). Financial market prediction using news and social media sentiment. *Journal of Business Research, 120*, 175–188.

Fedorova, E., & Iasakova, P. (2024). The impact of climate change news on the US stock market. *The Journal of Risk Finance, 25*(2), 293–320. https://doi.org/10.1108/JRF-06-2023-0133

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00762ED1V01Y201703HLT037

González, M., & Cruz Tadle, R. (2020). Signaling and financial market impact of Chile's Central Bank communication: A content analysis approach. *Economía, 20*(2), 127–178. https://www.jstor.org/stable/27007015.

Gu, W., Zhang, L., Xi, H., & Zheng, S. (2021). Stock prediction based on news text analysis. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 25*(5), 581–591.

Heaton, J., Polson, N., & Witte, J. H. (2017). Deep learning in finance. *Annual Rev.Financ. Eco., 9*, 145–181.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Unpublished manuscript.

Hu, Z., Zhao, Y., & Huang, Z. (2021). Neural networks for financial time series forecasting: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems, 32*(5), 1788–1804.

Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research, 40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Jin, X., Chen, C., & Yang, X. (2024). The effect of international media news on the global stock market. *International Review of Economics & Finance, 89*(Part A), 50–69.

Kahveci, E., & Odabaş, A. (2016). Central banks' communication strategy and content analysis of monetary policy statements: The case of Fed, ECB and CBRT. *Procedia - Social and Behavioral Sciences, 235*, 618–629. https://doi.org/10.1016/j.sbspro.2016.11.039.

Krinitz, J., Alfano, S., & Neumann, D. (2017). How the market can detect its own mispricing: A news sentiment index to detect irrational exuberance. In *50th Hawaii international conference on system sciences*. https://doi.org/10.24251/HICSS.2017.170

Kumari, V., & Pandey, D. K. (2025). Market reactions to the central bank's mandate on climate-related financial risk disclosures: Evidence from the Indian banking sector. *Finance Research Letters, 84*, 107774. https://doi.org/10.1016/j.frl.2025.107774.

Lee, C.-C., Wang, C.-W., Hsieh, H.-Y., & Chen, W.-L. (2023). The impact of central bank digital currency variation on firm's implied volatility. *Research in International Business and Finance, 64*, 101878. https://doi.org/10.1016/j.ribaf.2023.101878.

Li, X., Xie, H., & Wang, S. (2020). News impact on stock price predictability based on deep learning models. *Complexity*.

Liu, Y., Pan, Z., & Yu, W. (2019). Ensemble learning based on sparse representation and its application in financial prediction. *Applied Intelligence, 49*(11), 3795–3807.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research, 54*(4), 1187–1230.

Loughran, T., & McDonald, B. (2024). Measuring firm complexity. *Journal of Financial and Quantitative Analysis, 59*(6), 2487–2514. https://doi.org/10.1017/S0022109023000612

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Moustakidis, S., Plakias, S., Kokkotis, C., Tsatalas, T., & Tsaopoulos, D. (2023). Predicting football team performance with explainable AI: Leveraging SHAP to identify key team-level performance metrics. *Future Internet, 15*(5), 174. https://doi.org/10.3390/fi15050174

Nabeen, A., Yasir, M., Ansari, Y., Yasmin, S., Moon, J., & Rho, S. (2022). An empirical study of macroeconomic factors and stock returns in the context of economic uncertainty news sentiment using machine learning. *Complexity, 2022*, 1–18. https://doi.org/10.1155/2022/4646733

Nyakurukwa, K., & Seetharam, Y. (2023). Investor reaction to ESG news sentiment: Evidence from South Africa. *Economia, 24*(1), 68–85. https://doi.org/10.1108/ECON-09-2022-0126

Omotosho, B. S.. Central bank communication in Ghana: Insights from a text mining analysis [Working paper]. African Development Bank; Central Bank of Nigeria. https://doi.org/10.2139/ssrn.3526451.

Pal, S., Garg, A. K., & McMillan, D. (2020). Macroeconomic surprises and stock market responses in view of global linkage – a study of Indian stock market. *Cogent Eco. Finance, 8*(1). https://doi.org/10.1080/23322039.2020.1839171

Picault, M., Pinter, J., & Renault, T. (2022). Media sentiment on monetary policy: Determinants and relevance for inflation expectations. *Journal of International Money and Finance, 124*, 102626. https://doi.org/10.1016/j.jimonfin.2022.102626.

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2015). Enhancing sentiment analysis of financial news by detecting negation scopes. In *Hawaii international conference on system sciences, kauai, HI, USA* (pp. 959–968). https://doi.org/10.1109/HICSS.2015.119

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2016). Negation scope detection in sentiment analysis: Decision support for news-driven trading. *Decision Support Systems, 88*, 67–75. https://doi.org/10.1016/j.dss.2016.05.009

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports, 14*, 6086. https://doi.org/10.1038/s41598-024-56706-x

Rizkiana, A., Sari, H., Hardjomidjojo, P., Prihartono, B., & Yang, Z. (2019). The development of composite sentiment index in Indonesia based on internet-available data. *Cogent Eco. Finance, 7*(1). https://doi.org/10.1080/23322039.2019.1669399

Sako, K., Mpinda, B. N., & Rodrigues, P. C. (2022). Neural networks for financial time series forecasting. *Entropy, 24*(5), 657. https://doi.org/10.3390/e24050657.

Shen, S., Xia, L., Shuai, Y., & Gao, D. (2022). Measuring news media sentiment using big data for Chinese stock markets. *Pacific-Basin Finance Journal, 74*, Article 101810.

Tian, Y., Zhou, C., & Zhang, J. (2020). Application of machine learning methods in financial market prediction: A literature review. *Financ. Innov., 6*(1), 1–19.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I.. https://arxiv.org/abs/1706.03762.

Xing, F., Cambria, E., & Welsch, R. E. (2018). Intelligent asset allocation via market sentiment views. *IEEE Computational Intelligence Magazine, 13*(4), 16–28.

Yang, X., Ma, J., & Wang, C. (2021). Sentiment analysis for financial markets using news and social media: A survey. *Computational Economics, 57*(1), 1–26.

Yuanyuan, Z., Kumari, S., Ilyas, M., Bhayo, M.-U.-R., & Marwat, J. (2023). Media coverage and stock market returns: Evidence from China Pakistan economic corridor (CPEC). *Heliyon, 9*(3), Article e14204. https://doi.org/10.1016/j.heliyon.2023.e14204

Zhang, W., Li, Y., & Shen, D. (2021). Integrating news sentiment into stock market forecasting: A deep learning method. *Journal of Forecasting, 40*(1), 121–137.

Zhang, Z., & Ulku, N. (2019). Attitudes towards ambiguous information and stock returns. *Cogent Eco. Finance, 7*(1). https://doi.org/10.1080/23322039.2019.1693678