International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 16 (2024) pp. 378–394

Article

Smart Finance: Utilizing AI to Predict Stock Prices from News and Market Data

Veena Madhuri Sangala *, Sirisha Alamanda and Prathima Tirumalareddy

Chaitanya Bharathi Institute of Technology, Hyderabad 500075, India; asirishanaidu@gmail.com (S.A.); prathimareddyt@gmail.com (P.T.)
* Correspondence author:sangala.veenamadhuri@gmail.com

Received date: 1 March 2024; Accepted date: 21 March 2024; Published online: 10 July 2024

Abstract: This study contributes to Stock Price Prediction by examining how market sentiment and news impact stock movements. We've developed a new model that integrates market data and sentiment analysis of news to predict these movements accurately. Our model is crucial for investors and companies aiming to safeguard their interests. We meticulously clean and prepare data, engineer features, and employ various training methods to optimize model performance. Through rigorous testing, we achieve impressive accuracy, predicting next-day stock prices with about 77% accuracy. While our model extends predictions up to 10 days, accuracy slightly decreases over longer periods. Notably, our approach combines market data and news sentiment analysis, enhancing prediction precision. This research isn't just theoretical; it's a practical tool for decision-makers in finance. Our model represents a significant advancement in accuracy, offering valuable insights for strategic stock investments.

Keywords: stock price; market data; news data; classification models; data preprocessing; feature engineering

1. Introduction

Forecasting stock prices is vital for investors aiming to optimize financial gains by predicting future asset values. The accuracy of these predictions directly impacts profit margins, driving the pursuit of improved prediction methods in current research. Market data, encompassing metrics like opening and closing prices, trading volume, and stock returns, is pivotal. Additionally, news data, including word count, sentence count, relevance, and sentiment scores, significantly influences daily stock prices, driving fluctuations.

Investors meticulously analyze a company's historical performance, share values, and profit trends before deciding on investments. This often involves sifting through numerous news articles to stay abreast of the latest developments. This accumulated knowledge aids in determining whether to buy, sell, or hold shares, highlighting the significance of news data in decision-making. Time series analysis forms the bedrock for examining stock price data, involving the extraction of characteristics and statistics and employing forecasting models to predict future values based on historical trends.

Market data encompasses a wide array of metrics, including opening and closing prices, trading volume, and historical returns, providing essential insights into the performance of stocks and financial assets. This data is fundamental for understanding market trends, identifying patterns, and making informed investment decisions. However, market data alone may not capture the full spectrum of factors influencing stock prices.

On the other hand, news data plays a crucial role in shaping market sentiment and investor perceptions. News articles, press releases, and social media posts can contain valuable information about company performance, industry trends, regulatory changes, and geopolitical events that impact stock prices. Sentiment analysis techniques are often employed to gauge the mood and attitude expressed in news articles, providing additional context for understanding market dynamics.

By integrating market data and news data, analysts can gain a more comprehensive understanding of the factors driving stock price movements. Market data provides quantitative measures of past performance and market trends, while news data offers qualitative insights into the underlying factors influencing investor sentiment and market behavior. By combining these two sources of information, analysts can develop more robust predictive models that account for both quantitative and qualitative factors.

Using market data and news data together in stock price prediction allows for a more holistic approach that captures the complex interplay between market fundamentals and investor sentiment. This integration enables analysts to identify emerging trends, anticipate market reactions to news events, and make more accurate predictions about future stock price movements. Overall, the significance of using market and news data together lies in its ability to enhance the accuracy and reliability of stock price predictions, ultimately enabling investors to make more informed and profitable investment decisions.

To enhance stock price prediction, researchers deploy various classification models such as Multi-layer Perceptron (MLP), Recurrent Neural Network (RNN), Light Gradient Boosting Machine (LGBM), and Extreme Gradient Boosting Machine (XG-Boost). These models utilize different techniques, including hidden layers, activation functions, and pre-tuned parameters, to ensure accurate predictions.

Inspired by a Kaggle challenge from Two-Sigma [1], this research focuses on a subset of dynamically changing US-listed companies, utilizing market data with calculated returns and news data with pre-determined sentiment scores. The methodology encompasses data collection, preprocessing, feature engineering, model application, rigorous evaluation, and comprehensive results analysis. This research contributes to the evolving field of Stock Price Prediction by addressing the critical influence of market sentiment and news on stock movements.

Recognizing the profound implications of news in shaping stock prices, we propose a novel model designed to anticipate these movements by leveraging both market data and sentiment analysis of news. The envisioned framework holds significant potential for investors and corporate entities seeking to safeguard shareholder interests. In the pursuit of accurate predictions, our methodology involves meticulous data preprocessing, thoughtful feature engineering, and training on diverse classification models. The selection of the most accurate model, based on rigorous evaluation, is a key facet of our approach. Notably, our model demonstrates a commendable accuracy of approximately 77% in predicting stock prices for the next day. Furthermore, we extend our analysis to forecast stock price movements up to 10 days in advance, revealing a gradual decline in predictive accuracy over an extended timeframe.

This research distinguishes itself by providing a comprehensive solution that incorporates market data and sentiment analysis of news to enhance the precision of stock price predictions. Our findings not only contribute to the academic discourse on stock price prediction but also offer a practical tool for investors and firms aiming to make informed decisions in the dynamic financial landscape. Our model showcases a notable advancement in accuracy, underscoring its potential significance for strategic decision-making in the realm of stock investments.

This research explores the intricate landscape of stock price prediction, emphasizing the interplay between market and news data. By leveraging advanced classification models and a robust methodology, the study not only contributes to academic discussions but also addresses the practical needs of investors navigating the complexities of the financial market. The insights gained from this research have the potential to empower investors with a more informed and strategic approach to stock trading.

2. Literature Review

In recent years, the field of stock market prediction has witnessed remarkable efforts aimed at developing models capable of forecasting market trends and future stock prices. The interconnectedness between news articles about a company and its stock price movements has been a focal point of research, revealing the profound impact of information dissemination on financial markets.

Stock market forecasting, vital for business planning, has attracted interdisciplinary attention from computer science, finance, statistics, operations research and economics. While traditional methods like Long Short-Term Memory (LSTM) are adept at pattern recognition, they often overlook the crucial relationship between past and future values. Addressing this, a novel LSTM technique is proposed, incorporating a Three-Point Moving Gradient approach. By integrating reversed Linear Regression (LR) and mean price calculations, this method achieves comparable precision to conventional LSTM models, offering enhanced predictive accuracy for financial analysts [2].

With the rise of market participants, algorithmic trading has gained prominence, necessitating careful strategy development for profitability. Emerging markets like Bangladesh's stock market have yet to fully utilize ML-based algorithmic trading. This study fills this gap by developing an ML-based trading strategy for the Dhaka Stock Exchange (DSEX) and validating it through back testing. Using five years

of DSEX index and 49 company data, the study shows promising results with a sharp ratio of 0.50, offering practical implications for trading applications and enhancing understanding of ML-based back testing and investment strategy [3].

This study reviews recent advancements in using artificial intelligence (AI) to predict stock prices. Researchers analyzed papers published from 2020 to 2023, identifying 14 influential ones. Various AI techniques, including technical, fundamental, and sentiment analysis, were employed, with hybrid models showing superior performance. These models integrate deep learning to handle complex data patterns. The findings suggest AI holds significant potential in stock market prediction, guiding future research and practical applications in finance, shaping investment strategies, and decision-making processes [4].

This paper addresses the limited exploration of expanded feature engineering for complex and noisy stock data in financial forecasting. It introduces a solution involving Discrete Wavelet Transform (DWT) based feature engineering and hyperparameter-tuned ensemble models, termed as Wavelet-Particle Swarm Optimization (WPSO). The method, Multi-Stage Feature Engineering (MSFE), utilizes DWT decomposition for noise handling, followed by two-stage feature reduction. Tested on NIFTY, NASDAQ, and NYSE indices, WPSO demonstrates improved accuracy and outperforms state-of-the-art methods, emphasizing enhanced predictive outcomes through feature quality improvement and hyperparameter tuning [5].

This study introduces VGC-GAN, a framework for predicting stock prices. Unlike traditional approaches relying on predetermined graphs, VGC-GAN generates multiple correlation graphs from historical stock data, offering a comprehensive view of inter-stock relationships. It combines a Multi-Graph Convolutional Network and Gated Recurrent Unit within a Generative Adversarial Network framework to improve predictive performance. Evaluated on real datasets, VGC-GAN demonstrates effectiveness in stock price prediction by exploring hidden correlations and time dependence of stocks while mitigating noise impact through Variational Mode Decomposition [6].

Investors traditionally rely on personal experience and expert advice for decision-making. However, the emergence of Machine Learning in Quantitative Investment offers a more objective approach. Researchers developed a stock price prediction model using multi-layer training networks, integrating stock data and AI. This model, blending machine learning with traditional statistical methods, outperforms conventional techniques. Comparative experiments validate its effectiveness, marking a significant advancement in predictive analytics for investors [7].

Utilizing deep learning, particularly time series neural networks, for stock price prediction is crucial in quantitative finance. Although GAN-based models incorporating LSTM or GRU as generators have emerged, they often lack robust feature extraction, resulting in slightly lower prediction accuracies. Addressing these issues, a novel approach integrating multiple factors and GAN-TrellisNet is proposed. This method employs a multi-factor strategy and a GAN comprising TrellisNet and CNN for enhanced prediction accuracy. Comparative experiments across four markets demonstrate superior performance over existing GAN-based methods, highlighting the effectiveness of the proposed approach [8].

This study explores the profitability of employing machine learning algorithms to pick stocks from the S&P 500 using various factors as features. Tree-based algorithms like Decision Tree, Random Forest, and XG Boost are utilized due to their feature importance extraction capabilities. A back test is conducted to train the models and rebalance the portfolio using recent data. Despite higher risks, the selected assets outperform the index. The study demonstrates that the importance of factors determining asset performance evolves over time, offering profitability insights with explainability [9].

This research explores using Artificial Neural Network (ANN) techniques combined with Fuzzy Logic (FL) to enhance time-series data forecasting, focusing on DOW30 and NASDAQ100 US stock indices. The study introduces Adaptive Neuro-Fuzzy Inference System (ANFIS) and Wavelet Transform (WT) models, investigating different fuzzy Membership Functions (MFs) and WT filters. Results reveal that WT-ANFIS models surpass original ANFIS ones, with the trapezoidal MF yielding the most accurate forecasts. The study underscores the significant impact of MF types and numbers on forecasting accuracy [10].

This study assesses random forest models with AI for predicting stock trends, crucial for investors. The evaluation examines accuracy and efficiency on a test set of four stocks with optimal parameters [11].

The analysis delves into diverse methods for forecasting stock prices, encompassing statistical, machine learning, and deep learning models. It discusses their merits and drawbacks, emphasizing the efficacy of traditional statistical methods like autoregressive integrated moving average while noting their limitations in addressing non-linear problems. Machine learning algorithms, such as artificial neural networks, offer promise in capturing non-linear information, while deep learning models like convolutional and recurrent neural networks excel at decoding complex stock price patterns. The study

also explores hybrid models that blend different approaches to enhance predictive accuracy, shedding light on future directions in stock price forecasting [12].

Portfolio management involves deciding the best times to buy or sell stocks for maximum profit. Many individual investors find this challenging due to unclear goals and a lack of structured decision-making. With numerous stocks available, it's hard to choose. Effective decision support systems are needed to aid investors. Researchers explore methods like sentiment analysis of news and historical trends, but the impact of financial reports on stock prices is underexplored. This paper fills the gap by using machine learning to predict stock prices based on Nifty 50 companies' financial reports, including quarterly and annual reports, cash flow statements, and ratios [13].

This paper explores the resilience of a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) in predicting Bitcoin (BTC) daily closing prices amidst disturbances. Using Gaussian noise, the study evaluates the network's robustness, observing increased Root Mean Square Error (RMSE) with noise layers. Comparisons with Autoregressive Integrated Moving Average (ARIMA) models reveal LSTM's superior resilience to random disruptions. While highlighting DL techniques' robustness over linear methods, the study underscores the need for thorough testing, acknowledging varied network behavior based on circumstances [14].

This article discusses the use of machine learning, particularly LSTM structures, for predicting stock prices, focusing on technology companies listed on NASDAQ. It highlights the potential of machine learning in financial forecasting despite limitations in accuracy. While not sufficient for investment decisions, it offers insights and solutions for personal finance. The study suggests further enhancements in structure and logic to improve accuracy. Overall, it aims to expand research perspectives on machine learning in finance [15].

This study [16] centres on collecting Iranian news articles related to stocks and employs the HESNEGAR lexicon for polarity analysis post-data preprocessing. Pioneering the application of a semantic Persian lexicon, the research introduces three convolutional neural network-based models: one relying solely on stock prices, another incorporating news sentiments, and a hybrid model combining both. Experimental results reveal that the hybrid model, integrating technical indicators and news sentiments via the HESNEGAR lexicon, significantly improves prediction accuracy compared to models focusing solely on stock prices or news sentiments. Positioned as a reference for effective trading strategies in the Iranian stock market, the study contributes to market analysis.

Market prediction research in finance is crucial, given the growing data volume and diversity. A new model, HJPSO, combines jellyfish and particle swarm optimization algorithms to manage large data sets effectively. It optimizes support vector machine parameters and extracts decision rules, improving predictive accuracy and trading simulation compared to existing methods. Incorporating both technical indicators and financial news enhances performance, emphasizing the importance of integrating diverse data for better market predictions [17].

This study examines the impact of data preprocessing methods on prediction accuracy in big data analysis. Time series-based processing, notably using a proposed Time-series Recurrent Neural Network (TRNN), enhances prediction efficiency, especially in stock price forecasting. TRNN integrates trading volume and employs sliding windows for data processing, extracting trends and turning points while compressing data. Compared to conventional models like RNN and LSTM, TRNN demonstrates superior efficiency and accuracy, with potential applications beyond finance [18].

This paper [19] investigates the potential of video news presented by market specialists as a predictor for stock prices. Granger causality analysis and Pearson correlation coefficient tests establish the causal relationship strength. Sentiment analysis using TextBlob API and Google Cloud Natural Language API for video news is compared. SVM, LR, CNN, and LSTM models are assessed for sentiment analysis of S&P 500 stocks, with the most effective tool training a machine learning classification model. Empirical results validate a cause-and-effect relationship between the sentiment expressed in video news and the fluctuations observed in the stock market.

This research focuses on using Long Short-Term Memory (LSTM) neural networks for stock price forecasting, crucial in financial decision-making. By employing LSTM and deep learning techniques, it aims to build a robust predictive model. The study includes meticulous data collection, preprocessing, and model evaluation. Automated hyperparameter optimization enhances precision. Results highlight LSTM's efficacy in capturing stock market patterns, offering valuable insights for its application in finance, thereby improving prediction accuracy and reliability [20]. Addressing the intricate nature of predicting stock movements influenced by various factors, another paper suggests a blended ensemble deep learning approach specifically targeting the S&P 500 Index [21]. The model showcases a remarkable reduction in mean-squared error and increased precision rate, F1-score, recall and movement direction accuracy.

This study analyses over 34,000 news articles to understand how news sentiment influences

international equity markets. It finds limited spillover from news sentiment to stock markets, with stronger effects during periods of uncertainty, particularly related to Brexit [22]. Additionally, it investigates the steep stock market fall on March 13, 2020 (Black Friday), attributing it to investor fear of the COVID-19 pandemic and the Chinese stock market. The study reveals a significant impact of COVID-19 and China's market on global indices, suggesting a unidirectional relationship between China and global stock markets [23].

This study explores how social media serves as an attention driver for traditional media, influencing people to seek further information. Analyzing stock-related content on Sina Finance and Sina Weibo, the research finds that social media posts about a stock influence viewership of traditional news articles the following day. This effect is heightened with intense or positive sentiment and increased volume of verified social media posts [24]. Meanwhile, a novel stock prediction model, VMD-LSTMA+TCNA, is proposed to address challenges posed by stock series volatility. This model combines variational mode decomposition to reduce volatility, and a dual-channel attention network to capture both long-term dependencies and short-term patterns. Experiments on US and Hong Kong stock markets demonstrate the model's superior robustness and accuracy compared to existing methods [25].

In summary, the studies highlight the evolving landscape of predicting stock movements through news sentiments. From employing deep learning methodologies to gauging volatility predictability, these approaches reveal the intricate link between market dynamics and news related to finance [26]. As technology advances, these findings contribute to ongoing discussions on stock market prediction, emphasizing potential improvements in accuracy and forecasting insights. The literature survey underscores the evolution of research in stock market prediction, encompassing sentiment analysis, deep learning, and diverse data integration. These studies collectively offer valuable insights into challenges and opportunities, opening avenues for well-informed decision-making amid the dynamic financial terrain.

3.Methodology

3.1. Tools Used

For building classification models, Python, a popular programming language known for its simplicity and versatility, was utilized. Python offers a wide array of libraries and tools specifically designed for machine learning and data analysis tasks, making it an ideal choice for constructing sophisticated classification models.

Meanwhile, the user interface, which serves as the front end of the application, was developed using a combination of JavaScript, HTML, CSS, and JSON. JavaScript is a widely-used programming language primarily employed for web development, allowing for dynamic and interactive user interfaces. HTML and CSS are fundamental technologies used for structuring and styling web pages, while JSON (JavaScript Object Notation) facilitates the exchange of data between the server and the client.

To streamline the development process and enhance functionality, Flask, a lightweight and flexible web application framework in Python, was utilized. Flask simplifies the creation of web applications by providing tools and utilities for routing, handling requests, and managing sessions, allowing developers to focus on application logic rather than boilerplate code.

Additionally, the D3 library (Data Driven Documents) in JavaScript played a crucial role in developing the web application. D3 is a powerful and versatile library for data visualization, enabling the creation of interactive and visually appealing charts, graphs, and other visualizations. By leveraging D3, the application can present data in a clear and intuitive manner, enhancing user experience and comprehension.

3.2. Data Collection

For this research, we obtained data from a Kaggle challenge hosted by Two-Sigma, which is a company that manages investments. The dataset contains information about stocks, with details for 3,510 different stocks. This information is divided into two parts: news data from Thomson Reuters and market data from Intrinio. The market data initially had a lot of rows and columns—specifically, it had 4,072,955 rows and 16 columns. These columns included things like the prices of stocks when they opened and closed, the volume of stocks traded, and the returns on investments. On the other hand, the news data had even more rows and columns—it had 9,328,749 rows and 35 columns. This data includes the results of sentiment analysis, which is a way of understanding the emotions expressed in the news. For example, it looks at things like how many words and sentences are in each news article and whether the overall tone of the article is positive, negative, or neutral. It's worth noting that the market dataset appears smaller because there can be multiple news articles related to each stock on any given day. The process flowchart is depicted in Figure 1. The market data comprises the following features as shown in Table 1:

Table 1. Features of Market Data.

Feature Name	Description of Feature
time	indicates the current time (in market data, all values are taken at 22:00 UTC)
assetCode	asset's unique id
assetName	name of the asset, which may be for a group of assetCodes
universe	boolean value denoting whether or not the asset on a day will be put in scoring
volume	trading volume in shares for a day
close	close price which is not adjusted for splits or dividends for the day
open	open price which is not adjusted for splits or dividends for the day

These are the features based on returns: returnsClosePrevRaw1, returnsOpenPrevRaw10, returnsOpenPrevRaw1, returnsOpenNextMktres10, returnsOpenPrevMktres1, returnsClosePrevRaw10, returnsClosePrevMktres1, returnsClosePrevMktres10, returnsOpenPrevMktres10. The returns in the market data can be interpreted by the following properties:

- Returns are consistently computed either from close to close or open to open, representing the transition between the closing time of one trading day and the opening time of another.
- Returns are either market-residualized (Mktres), accounting for market fluctuations, or raw, reflecting unadjusted data against any benchmark.
- Returns can be computed over any arbitrary interval, with a focus on 1-day and 10-day horizons in this context.

Returns are tagged with 'Next' if they are forward looking, 'Prev' if they are backward looking.

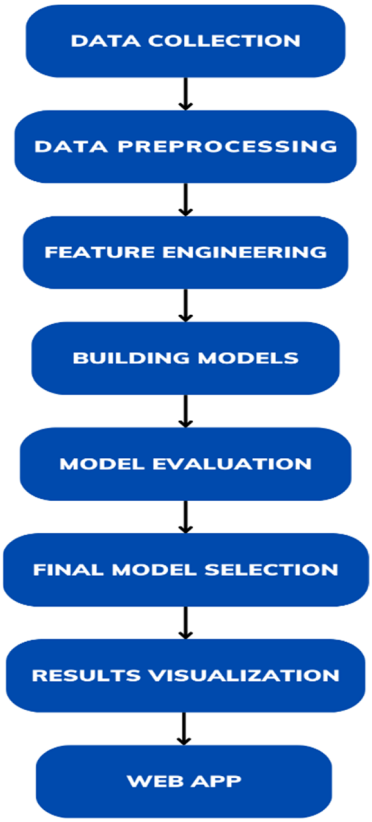

Figure 1. Flow chart of the proposed methodology.

The news data encompasses features detailed in Table 2. Notably, the news data includes additional features of noveltyCount and volumeCounts calculated over 12 hours, 1 day, 3 days, 5 days, and 7 days, respectively. This comprehensive approach to data collection and feature extraction forms a robust foundation for subsequent analysis and modeling in the study.

Table 2. Features of News Data.

Feature Name	Description of Feature
time	UTC timestamp showing the availability of data on the feed
sourceTimestamp	UTC timestamp of creation of the news item
firstCreated	UTC timestamp of the item's first version
sourceId	each news item has a source id
headline	the news headline
urgency	comprehend types of stories as alerts (1), articles (3)
takeSequence	The news piece's sequence number, starting at 1, differs for articles and alerts within a story.
provider	organisation which provided the news item
subjects	company identifiers and topic codes associated with this news piece
audiences	identifies which news item belong to the desktop news product
bodySize	size of the present version of the story body in characters
companyCount	the list of companies in the news piece are included
headlineTag	headline tag of the news item given by Thomson Reuters
marketCommentary	general market conditions are indicated as a Boolean value
sentenceCount	number of sentences in the news item
wordcount	number of lexical tokens in the news item
assetCodes	code for the assets listed in the item
assetName	name of the asset
firstMentionSentence	score is given based on the news item placement, headline (1), the initial sentence within the story's main content (2), next sentence of the body (3) and so on. If the news item asset is not there then 0
relevance	relevance of the news article denoted by a decimal number
sentimentClass	sentiment score with respect to the asset
sentimentNegative	–1
senimentNeutral	0
sentimentPositive	1
sentimentWordCount	the sentiment of the word count relevant to the asset
noveltyCount	novelty of the content on a particular asset within a news item
volumeCounts	volume of news for each asset

3.3. Data Preprocessing

The dataset used in this study was originally sourced from Kaggle and covered the years 2007 to 2016. However, the authors narrowed their focus to the period from 2010 to 2016 to exclude the influence of the 2008 financial crisis. Training the model involved using data from 2010 to 2015, while testing utilized data from 2016. Both market and news data underwent thorough cleaning and feature engineering, with a primary emphasis on the latter.

1) Data Preprocessing on News Data: During the preprocessing of news data, the authors calculated the delay, which represents the time between when a news item is created and when it becomes available. Rows with a delay of one day or more were removed to ensure the timeliness of the data. Additionally, rows with blank headlines were excluded as they were deemed to have little impact on predicting stock prices. Rows with an urgency level of 2, which accounted for less than 0.001% of the data, were also removed as they were considered insignificant.

2) Data Preprocessing on Market Data: In the preprocessing of market data, the authors first addressed outliers in open and close prices by removing rows where the difference between the close price of one day and the open price of the next exceeded 50%. Null values in market-adjusted columns were then filled using imputation methods, replacing them with raw values from the same row. Finally, outliers based on returns were identified and removed, specifically stocks with returns exceeding or falling below 50%.

By meticulously preprocessing the data in this manner, the authors ensured the integrity and quality of the dataset. This rigorous approach lays the foundation for meaningful analysis and modeling. Furthermore, their focus on mitigating the impact of historical events, such as the 2008 financial crisis, and refining the quality of the data underscores a robust methodology aimed at extracting valuable insights from the dataset.

3.4. Feature Engineering

The news data utilized in this study comprises several key features essential for understanding and analyzing market dynamics. One such feature is urgency, which categorizes news items as either articles or alerts on a scale ranging from 1 to 3. Additionally, the sentiment class feature provides insight into the overall sentiment toward a stock, with values of 1 indicating positivity, –1 indicating negativity, and 0

representing neutrality. Integrating this information with stock price data necessitates merging numerical and categorical attributes. To achieve this, the authors strategically combined specific features of the news data, such as urgency and sentiment class, to create new features conducive to integration with market data. This approach minimizes information loss and enhances synergy between the two datasets, enabling a more comprehensive analysis of market trends.

An illustrative example of this transformation is the creation of the "wordcount_3_1" feature, which denotes the word count of a news item with an urgency level of 3 and a positive sentiment class of 1. Additionally, several other transformed features were generated to capture various aspects of the news data, including word count, sentence count, sentiment word count, first mention sentence, relevance, rel_FirstMention, and rel_SentCount. These features provide detailed insights into the content and relevance of news items, facilitating a deeper understanding of their impact on stock price movements.

To minimize the influence of market data, only two features from the market data set were used to train the model: returnsOpenPrevMktres1_dir, representing the shift in market-adjusted stock prices from the preceding day, and returnsOpenPrevMktres10_dir, indicating the shift in market-adjusted stock prices from the 10th preceding day. By focusing on these specific features, the authors aimed to isolate the effects of market data and emphasize the significance of news-related variables in predicting stock price movements.

To consolidate the news data at a daily level, it was grouped based on asset code and date, allowing for the extraction of key features from all news items for each day. Subsequently, a merged dataset was created by aggregating the transformed news data with the daily market data. This merged dataset serves as the foundation for training and testing the predictive models, enabling a comprehensive analysis of the relationship between news sentiment and stock market dynamics. The above transformation is performed on the original features and the derived feature set is as following:

- wordCount: the count of words in the news item
- sentenceCount: the number of sentences in the news item
- sentimentWordCount: the number of lexicons in a news item which are relevant to an asset.
- firstMentionSentence: the sentence where the asset is mentioned initially.
- relevance: this value denotes the relevance of the news item to the asset between 0 and 1
- rel_FirstMention: indicates where in a news item, the asset or stock was first found.
- rel_SentCount: indicates the proportion of words in the news item which had sentiments affecting a stock.

To keep the effect of market data to minimal, only two features from market data are used to train the model:

- returnsOpenPrevMktres1_dir: the shift in market-adjusted stock prices from the preceding day.
- returnsOpenPrevMktres10_dir: the shift of the market adjusted stock price during previous 10th day.

Feature Importance: In the realm of machine learning, understanding feature importance is crucial for interpreting the predictive power of different attributes. Much like random forests, Gradient Boosting Machines (GBMs) offer a method to ascertain the importance of features once decision trees are constructed. These importance scores shed light on how influential each feature was in the creation of boosted decision trees within the model. Essentially, features that significantly contribute to decision-making across the trees are deemed more important. To visualize this importance, the XG-Boost algorithm generates a feature importance graph, providing a ranked list of attributes used for prediction. This graphical representation aids in comparing attributes and discerning their relative significance in the predictive process.

Determining feature importance involves evaluating the impact of each feature split point on performance metrics, taking into account the number of observations influenced by each node in the decision trees. Performance metrics may encompass measures of purity, such as the Gini index, or more specific error functions like Mean Square Error (MSE) or Mean Absolute Error (MAE). These values are calculated across all boosted trees constructed during the training process and are subsequently averaged within the model. By assessing feature importance, researchers and practitioners gain valuable insights into the underlying mechanisms driving predictive performance, allowing for informed decision-making and model optimization.

After analyzing the plot presented above, the authors have drawn several noteworthy conclusions regarding the relationship between news sentiment and stock market dynamics. Firstly, they observed that individuals tend to respond more to news articles than to alerts, suggesting that articles may have a greater impact on market sentiment. Additionally, the analysis revealed that positive sentiment exerts a more influential effect on the stock market compared to negative sentiment, indicating that optimistic news tends to drive market movements to a greater extent. Furthermore, the authors found that news

articles alone can influence stock price movements for a limited duration, with the effect gradually diminishing over time. This suggests that while news may initially impact market sentiment, its significance diminishes as time progresses, highlighting the transient nature of news-driven market movements.

Moreover, the analysis revealed that the fraction of sentences or words within a news article related to a particular asset holds greater importance than the overall sentence or word counts. This indicates that the relevance of content to specific assets plays a critical role in influencing market sentiment and subsequent price movements. Lastly, the relative position of where an asset is first mentioned within a news article was found to be more important than the actual sentence number of that first appearance. This suggests that the context and prominence of an asset's mention within a news article can significantly impact market perception and response.

Overall, these conclusions offer valuable insights into the nuanced relationship between news sentiment and stock market behavior, providing researchers and investors with a deeper understanding of the factors driving market dynamics.

3.5. Building Classification Models

Initially, the authors embarked on their research endeavor by adopting a holistic approach, encompassing all available features from both market and news data to construct their predictive model. However, upon detailed analysis, they discerned that the most influential features driving model performance predominantly originated from market data. This observation was incongruent with their primary objective, which aimed to predict stock price movements using sentiments derived from news data. Consequently, the authors refined their methodology by excluding features from market data and focusing solely on leveraging two engineered features from market data alongside all features from the news data for model training.

In the realm of neural networks, the authors opted to employ Multi-Layer Perceptron (MLP) and Recurrent Neural Network (RNN) architectures. The selection of MLP was grounded in previous research indicating its efficacy in handling tabular data, time series prediction, and classification tasks, all of which are directly applicable to stock price prediction tasks that inherently involve time series data. MLP's capability to capture intricate patterns in data made it a suitable candidate for modeling stock price movements based on historical trends and sentiment analysis derived from news data.

On the other hand, the authors also explored the utilization of Recurrent Neural Networks (RNNs). RNNs are well-suited for processing sequential data, making them particularly adept at handling time series data such as stock prices. By incorporating memory elements, RNNs can capture dependencies and temporal patterns present in sequential data, thereby enhancing their predictive capabilities. In the context of stock price prediction, RNNs offer the potential to capture complex relationships between past market behaviors and future price movements, making them a valuable addition to the modeling toolkit.

In addition to neural network architectures, the authors also considered gradient boosting machines as part of their modeling approach. Specifically, they focused on LightGBM (LGBM) and XGBoost, two state-of-the-art algorithms widely used in predictive modeling tasks. LGBM is renowned for its efficient training speed and scalability, making it particularly suitable for handling large datasets commonly encountered in financial markets. Its leaf-wise splitting strategy enables it to efficiently explore the feature space, resulting in faster convergence and improved predictive performance. However, it's important to note that LGBM's rapid training speed may come at the cost of potential overfitting, especially when dealing with imbalanced datasets or noisy data.

On the other hand, XGBoost, while not as fast as LGBM, offers stability and robustness in predictions. It employs a gradient boosting framework that sequentially builds an ensemble of weak learners, iteratively refining the model's predictions with each subsequent iteration. XGBoost's versatility and effectiveness have made it a popular choice in various machine learning competitions and real-world applications. Its ability to handle both regression and classification tasks, coupled with its interpretability and ease of use, makes it an attractive option for predicting stock price movements.

Throughout the iterative process of model refinement, the authors meticulously evaluated the performance of each modeling approach, considering both its strengths and limitations. By conducting comprehensive analyses of the predictive outcomes for both positive and negative movements in stock prices, the authors gained valuable insights into the efficacy of each methodology in capturing the nuanced dynamics of the financial markets.

In summary, the authors' modeling approach involved a careful selection of machine learning techniques tailored to the unique challenges posed by stock price prediction tasks. By leveraging a combination of neural network architectures such as MLP and RNN, along with gradient boosting algorithms like LGBM and XGBoost, they aimed to harness the predictive power of both historical market data and sentiment analysis derived from news data (Figure 2). Through thorough

experimentation and analysis, the authors aimed to develop a robust predictive model capable of accurately forecasting stock price movements, thereby providing valuable insights for investors and stakeholders in the financial markets.

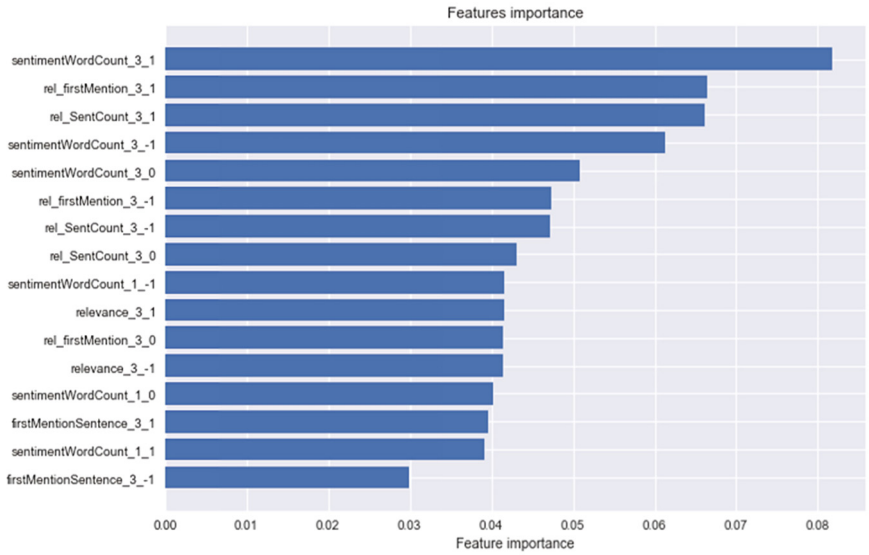

Figure 2. The feature importance graph given by XG-BOOST.

4. Findings and Analysis

4.1. Model Evaluation

The authors began their modeling process by using Recurrent Neural Networks (RNNs) to predict outcomes. They utilized two different types of RNN architectures, each designed to capture different aspects of the data. They summarized the training procedure and results in Table 3.

In the first RNN, the architecture included Gated Recurrent Units (GRU) and dense layers with Rectified Linear Unit (ReLU) and Sigmoid activation functions. The prediction process involved four hidden layers, and the achieved accuracy for this setup was around 68.9%. This success shows that the chosen architecture effectively captured the temporal patterns within the dataset.

For the second RNN, a similar approach was used, with dense layers featuring ReLU and Sigmoid activation functions. However, this configuration had an expanded architecture with five hidden layers. Despite the increased complexity, the accuracy achieved for this model was 58.22%. Comparing the two RNN architectures highlights the dataset's nuanced nature and the importance of customizing model configurations.

The detailed breakdown of the training process, including specific layers, activation functions, and the number of hidden layers for each RNN, offers valuable insights into the authors' modeling decisions. The accuracy metrics provide a quantitative measure of the models' predictive performance. This thorough examination enhances readers' understanding of the RNN-based modeling approach, setting the stage for a comprehensive evaluation of predictive capabilities.

Table 3. Training through RNN and corresponding accuracy.

Attempt	Layer	No. of Hidden Layers	Activation Functions	Accuracy
1	Gated Recurrent Unit (GRU) Dense	4	Rectified Linear Unit (ReLU) Sigmoid	68.9%
2	Dense	5	ReLU Sigmoid	58.22%

Following that, the authors employed a Multilayer Perceptron (MLP) to train their predictive model using Python. To determine the number of hidden layers, they systematically calculated it by dividing the shape of the training set by 50, which was then assigned as the variable for hidden layer sizes. They used the Rectified Linear Unit (ReLU) as the activation function in this process. The authors then conducted predictions for various time frames, including the next day, 2 days, 4 days, 6 days, and 10 days in advance. To provide clarity and evaluation, they presented the accuracy of both the training and

testing phases using MLP in Table 4. Through thorough testing, the authors found that the MLP model achieved an accuracy of 81.6% for next-day predictions and 54% for 10-day predictions. These results demonstrate the robustness of the MLP model in accurately forecasting stock price movements for both short and extended time frames. Additionally, they provide valuable insights into the model's predictive capabilities across various time horizons.

Table 4. Training and test accuracy of MLP.

Day	Train Accuracy	Test Accuracy
1	81.9%	81.6%
2	76.5%	76.4%
4	69.4%	69.3%
6	64.8%	64.6%
10	55%	54%

Subsequently, the authors employed LGBM as their chosen algorithm. To enhance the model's performance, a meticulous parameter tuning process was conducted to identify optimal settings for training the model using Python. The resulting optimal parameters, outlined in Table 5, were then input into the LGBM model to facilitate training. This comprehensive tuning and training approach led to the development of a well-optimized model. Upon evaluating the model's performance, the authors observed an accuracy rate of 53.08%. This accuracy metric indicates the proportion of correctly predicted outcomes, providing insight into the model's effectiveness in capturing patterns and making accurate predictions based on the chosen parameters.

Table 5. Optimal parameters of LGBM.

Parameter	Values of the Parameter
n_estimators	2500
max_depth	6
colsample_bytree	0.859
min_child_samples	72
sub_sample	0.876
reg_lambda	0.1

Given the less-than-expected accuracy from the LGBM model, the authors opted for the XG-Boost model for training in Python. The process commenced with parameter tuning on the training set to identify the most effective configuration. Given that the data used for predicting stock price movements is inherently time series data, the authors employed the TimeSeriesSplit functionality from scikit-learn. This was incorporated within the RandomizedSearchCV procedure, utilizing 5 folds to ensure comprehensive cross-validation. TimeSeriesSplit is particularly advantageous in time series data analysis as it maintains temporal constraints when creating folds, acknowledging the sequential nature of the data. To elaborate on the parameter tuning process, RandomizedSearchCV is employed to explore the parameter space and identify the optimal parameters for the XG-Boost model. The process involves systematically searching through a specified number of combinations, evaluating their performance, and selecting the set of parameters that result in the highest predictive accuracy. The use of TimeSeriesSplit within this cross-validation framework ensures that the temporal order of the data is preserved, enhancing the model's ability to generalize well to unseen time points.

The optimal parameters obtained through this rigorous process are detailed in Table 6. These parameters, crucial for the performance of the XG-Boost model, are selected based on their ability to enhance predictive accuracy in the context of time series data. The authors have thoughtfully presented these parameters along with their descriptions, providing transparency into the configuration that yielded the best results for anticipating stock price movements.

- n_estimators: no. of trees to build
- subsample: percentage of data to be sampled before constructing the tree
- max_depth: tree's maximum depth
- colsample_bytree: the proportion of columns to be sampled before constructing the tree
- min_child_weight: minimum value of sum of instance weight required in a child
- gamma: minimum loss minimisation needed to further partition the leaf node of a tree
- reg _lambda: regularization term on weights which determines how conservative the model is
- reg_alpha: the term of regularization on weights which determines how conservative the model is

- learning_rate: this is used to prevent overfitting in a model

Subsequently, the identified parameters are applied to the test set, culminating in the calculation of results using the XG-Boost model. Upon implementation, the authors observed an equilibrium in the proportions of true positive and negative, false positive and negative values for the next 10-day prediction interval. The obtained accuracy after the comprehensive training and testing with the XG-Boost model is explicitly detailed in Table 7. The evaluative metrics include a true positive rate of 55%, a true negative rate of 53%, and an overall accuracy of 54%.

Furthermore, a series of predictive models were systematically constructed to forecast stock price movements 1 to 10 days in advance. Notably, the model exhibited a commendable accuracy of approximately 77.01% for next-day predictions. However, this predictive accuracy gradually decreased with each successive day in the forecast horizon. The nuanced progression of prediction accuracy across different time frames provides a granular understanding of the model's performance over varying prediction intervals, offering valuable insights into its strengths and limitations across different temporal scopes.

Table 6. Optimal parameters of XG-boost.

Parameter	Value of the Parameter
n_estimators	5000
max_depth	8
subsample	0.7
colsample_bytree	0.8
min_child_weight	10
gamma	2
reg_lambda	1
reg_alpha	2
learning_rate	0.01

Table 7. Training and test accuracy of xg-boost.

No. of Days	Training Accuracy	Validation Accuracy
1	76.19%	77.01%
2	71.85%	72.73%
3	68.46%	69.01%
4	65.76%	66.26%
5	63.36%	63.75%
6	61.36%	61.62%
7	60.13%	59.65%
8	59.55%	59.85%
9	54.83%	53.13%
10	54.37%	53.40%

The chosen final model, following a thorough training of various models, is XG-Boost. The preference for XG-Boost over MLP stems from several key considerations. GBMs, like XG-Boost, are rooted in decision trees rather than intricate hidden layers, rendering them significantly more interpretable than neural networks. Their interpretability is crucial for understanding the decision-making process of the model. Moreover, GBMs, including XG-Boost, exhibit noteworthy speed and efficiency in comparison to neural networks, fitting quickly and requiring less computational time.

Although MLP, a type of neural network, demonstrated comparable accuracies to XG-Boost, the decision to prioritize the latter was influenced by the interpretability and computational efficiency advantages of GBMs. MLP, while proficient in certain tasks, involves more complex computations due to its hidden layers, leading to a higher computational burden. Given the almost equivalent performance of both models, the pragmatic choice of XG-Boost as the final model is grounded in its superior interpretability, computational efficiency, and comparable accuracy. This decision underscores the importance of selecting models not only based on predictive performance but also considering their interpretability and computational efficiency, especially in contexts where these factors hold substantial weight.

4.2. Results Visualization

Firstly, the accuracies for the next day prediction of all the models are compared and are shown in Table 8.

Table 8. Accuracy comparison**.**

Model	Accuracy
Recurrent Neural Network (RNN)	68.9%
Light Gradient Boosting machine (LGBM)	53.08%
Multi-layer Perceptron	81.6%
Extreme Gradient Boosting Machine (XG - Boost)	77.01%

The comparison of accuracy for predicting stock prices ten days in advance or for the next ten days is detailed in Table 9, where the performance of both Multilayer Perceptron (MLP) and XG-Boost models is showcased. As these two models exhibited the highest accuracy rates, they were selected for conducting the predictions. In addition, the training and test accuracies of the XG-Boost model are graphically represented in Figure 3, providing a visual depiction of the model's performance across different prediction horizons.

Table 9. The accuracy comparison for next 10 days.

Model	Next Day	2 Days	4 Days	6 Days	10 Days
MLP	81.6%	76.4%	69.3%	64.6%	54%
XG-Boost	77.01%	72.73%	66.26%	61.62%	53.40%

Upon close examination, the authors noted a gradual decrease in accuracy from predicting the next day's stock prices to forecasting eight days ahead. However, there was a significant drop in accuracy when attempting to predict stock prices for the tenth day. This observation suggests that while both MLP and XG-Boost models offer reliable predictions for shorter time frames, their accuracy diminishes as the prediction horizon extends, particularly beyond the eighth day. This finding underscores the importance of considering the limitations and challenges associated with long-term stock price forecasting, as indicated by the notable decrease in predictive accuracy observed for the tenth day.

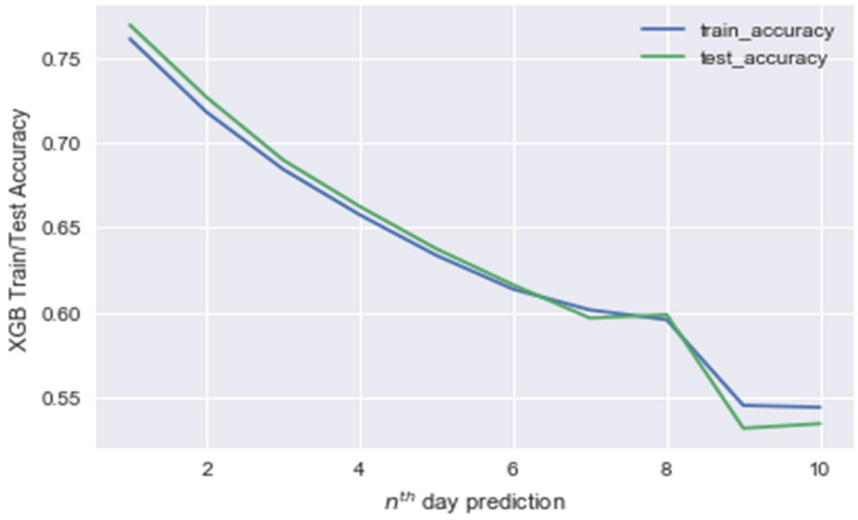

Figure 3. Training and test accuracy of XG-BOOST.

The authors have presented the true positive and negative rates of the XG-Boost model in a graphical format, depicted in Figure 4. Through their analysis, they noted interesting trends in these rates over various prediction horizons. Initially, both true positive and negative rates exhibit a gradual decrease up to the sixth day of prediction. However, beyond the sixth day, there is a notable variation in the true positive and negative rates.

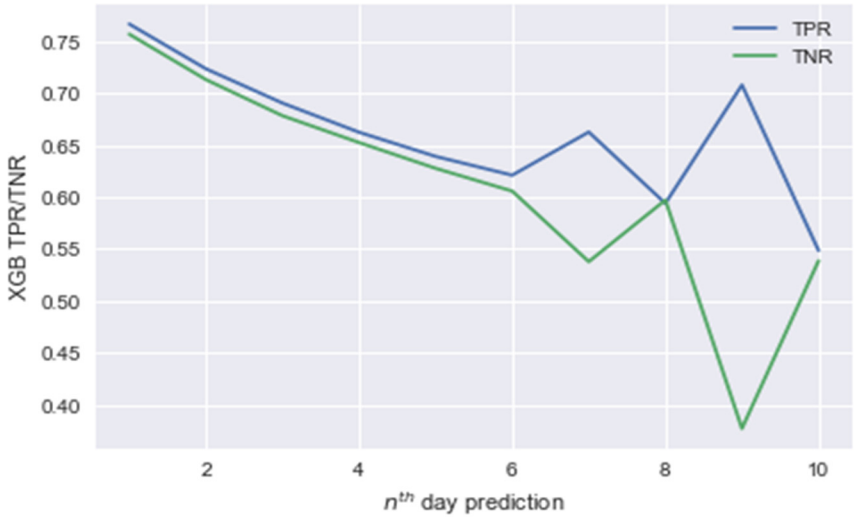

Figure 4. The true positive rate and negative rate of XG-BOOST.

For predictions extending to the seventh day, the true positive rate experiences an increase while the true negative rate decreases. Conversely, in the eighth-day prediction, the true positive rate decreases, and the true negative rate increases, ultimately converging at a certain point. Moving on to the ninth-day prediction, the true positive rate shows an increase, followed by a decrease in the tenth-day prediction, and vice versa for the true negative rate.

These observations indicate the dynamic nature of predictive performance over longer time horizons. The fluctuations in true positive and negative rates highlight the evolving nature of stock price movements and the challenges associated with accurately forecasting them beyond a certain point. Understanding these trends can provide valuable insights for refining predictive models and improving their performance over extended prediction periods.

This research presents an innovative approach by introducing an interactive web application developed by the authors using Flask and Data Driven Documents (D3). Unlike many previous studies, which often lack visualization of results, this application aims to provide users with a dynamic platform for visualizing stock price and volume data over time, while also exploring the impact of news on stock price movements.

The interactive web application consists of two main pages, each serving a distinct purpose in facilitating user interaction and understanding of the underlying data. The first page acts as an entry point, offering an introduction to the topic and providing guidance on how to navigate and utilize the application effectively. Additionally, this page offers insights into the importance of news features in predicting stock price movements, thus setting the stage for deeper exploration within the application.

Upon accessing the web application, users are greeted with a user-friendly interface that guides them through the functionalities available. The front page provides a brief overview of the application's objectives and features, ensuring that users have a clear understanding of its purpose and capabilities. Furthermore, it includes instructions on how to interact with the application, ensuring that even users with limited technical knowledge can navigate and utilize its features effectively.

In addition to serving as an introductory hub, the front page of the application offers valuable insights into the significance of news features in predicting stock price movements. By highlighting the relationship between news content and stock price fluctuations, users gain a deeper understanding of the factors driving market dynamics. This information empowers users to make more informed investment decisions by considering the impact of news events on asset valuations.

Moving beyond the introductory phase, users can delve deeper into the data through the application's second page, which is dedicated to visualizing specific asset graphs. This page, depicted in Figure 5, comprises six sections, each offering unique insights into the underlying data.

The first section of the second page provides an overview of the selected asset, including its name, symbol, and key performance metrics. This serves as a reference point for users, enabling them to quickly identify the asset of interest and understand its historical performance. The second section offers a graphical representation of the asset's historical price movements over time. By visualizing price trends, users can identify patterns, trends, and anomalies that may impact future price movements. This graphical representation enhances users' understanding of market dynamics and enables them to make more informed investment decisions.

In the third section, users can explore the relationship between news events and stock price movements. This interactive feature allows users to select specific news events and observe their impact on the asset's price trajectory. By correlating news events with price fluctuations, users gain valuable insights into the factors driving market sentiment and asset valuations. The fourth section provides a comprehensive analysis of trading volume associated with the selected asset. By visualizing volume trends over time, users can gauge market activity and identify periods of heightened trading activity or liquidity. This information enables users to assess market sentiment and liquidity conditions, informing their investment decisions accordingly.

The fifth section offers advanced technical analysis tools, including indicators such as moving averages, RSI, and MACD. These technical indicators help users identify potential entry and exit points, assess trend strength and momentum, and anticipate future price movements. By incorporating technical analysis tools, users can refine their investment strategies and enhance their trading performance. Finally, the sixth section offers customizable features that allow users to personalize their analysis experience. This includes the ability to adjust timeframes, select specific data points for analysis, and customize chart settings to suit individual preferences. By providing users with the flexibility to tailor their analysis experience, the application enhances usability and ensures that users can extract maximum value from the available data. The Figure 5 is the second page of the app which has six sections in it which are explained in detail based on the figure below:

1. The interface provides a comprehensive list of stocks, allowing users to select a specific stock to access relevant information and insights.
2. Users can track the movement trend of stock prices, with tooltips providing additional details such as expected and predicted stock price movements. The predicted arrow indicates the direction forecasted by the model, while the expected arrow represents the actual movement observed.
3. A ticker is prominently displayed, showcasing both user volume and actual stock prices for a chosen date. Users can interact with the line plot by hovering over it, which dynamically updates the displayed date.
4. A legend located at the top of the interface enables users to select their preferred time period for viewing stock price patterns, offering flexibility and customization.
5. A blue box feature allows users to specify a custom time window for analysis and exploration. Additionally, users can scroll over the blue box to navigate through different time frames.

6. Users can access a list of news items corresponding to the selected date of interest. Each news item is accompanied by sentiment values (1 for positive, –1 for negative, and 0 for neutral) and urgency values (1 for alerts and 3 for articles). The color of the news item indicates its impact on stock price movement, with green representing positive news and red indicating negative news. This feature provides users with valuable insights into how news events influence stock prices.

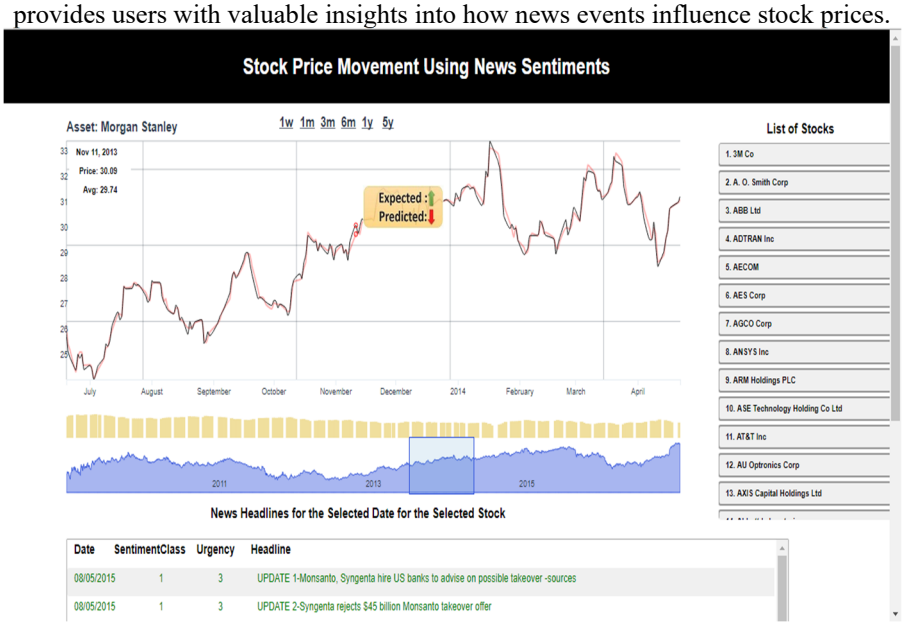

Figure 5. Flask app for visualisation.

5. Conclusions

In contemporary research, there is a notable shift towards forecasting stock prices by scrutinizing historical market data. Past endeavors have yielded commendable achievements, with some studies achieving an accuracy of approximately 70% in predicting stock prices concerning short-term pharmaceutical stock trends. However, our approach diverges by incorporating news data as a predictive factor, which has enabled our model to attain accuracies of around 77% for next-day predictions. This underscores the pivotal role of news data in influencing movements in stock prices, suggesting its potential as a valuable resource for investors seeking to make well-informed long-term investment decisions. This emphasis on news data is particularly significant given its sustained impact over extended periods, contrasting with the sporadic fluctuations often observed in market data.

Looking forward, there is ample opportunity for future research to delve deeper into the nuances of news data and its implications for stock price movements. One avenue for exploration involves investigating the influence of specific keywords within news articles on stock prices. By identifying key terms or phrases that correlate with significant fluctuations in stock prices, researchers can gain valuable insights into the underlying mechanisms driving market dynamics. Similarly, analyzing the impact of different news sources on stock price movements could offer further clarity on the factors shaping investor sentiment and market behavior. Such investigations could be facilitated through the utilization of news APIs, which provide access to a vast array of news articles and sources for analysis.

It is important to acknowledge the limitations of the dataset utilized in this study, which was sourced from Kaggle and spans until 2016. While this dataset offers a valuable historical perspective, future investigations could benefit from examining the model's performance with more recent data. By incorporating data from recent years, researchers can ensure that their models remain relevant and applicable to current market conditions. This expanded scope of research holds promise for refining predictive strategies, enhancing the accuracy of stock price predictions, and providing actionable insights for investors navigating the dynamic landscape of the stock market.

Our research underscores the importance of incorporating news data into predictive models for forecasting stock prices. By leveraging insights from news articles, investors can gain a more comprehensive understanding of market trends and make informed decisions about their investment strategies. Moving forward, there is a wealth of opportunities for further exploration and refinement, including investigating the impact of specific keywords and news sources on stock prices, as well as evaluating the model's performance with more recent data. Through continued research and innovation, we can continue to advance our understanding of the complex interplay between news data and stock market dynamics, ultimately empowering investors to navigate the markets with confidence and clarity.

Author Contributions

Conceptualization, V.M.S., S.A. and P. T.; methodology, V.M.S.; software, V.M.S.; validation, V.M.S., S.A. and P. T.; formal analysis, V.M.S.; investigation, V.M.S.; resources, V.M.S.; data curation, V.M.S.; writing—original draft preparation, V.M.S.; writing—review and editing, V.M.S.; visualization, V.M.S.; supervision, S.A. and P. T.; project administration, V.M.S., S.A. and P. T.;

Funding

This research received no external funding.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Data Availability Statement

Dataset is publicly available in Kaggle and is present in the reference number 1.

References

1. Two Sigma. Using News to Predict Stock Movements (2018). [Online]. Available at: https://www.kaggle.com/c/two-sigma-financial-news (Accessed on date 15 March 2020).
2. W.P. Shoong, S.A. Asmai, N.Z. Zulkarnain, T.C. Chuan. "An Improved LSTM technique using Three-Point Moving Gradient for Stock Price Forecasting". *International Journal of Computer Information Systems & Industrial Management Applications* 14 (2022).
3. A.I. Kabir, S. Vyas, S. Mitra, M.M. Uddin, and J. Jakowan "Evaluating the machine learning based momentum stock trading strategies with back-testing: An emerging market perspective. " In *AIP Conference Proceedings*, vol. 2919, no. 1. AIP Publishing, 2024.
4. A.l.-Alawi, A. Ismail and N. Alshakhoori. "Stock Price Prediction Using Artificial Intelligence: A Literature Review. "In *Proceedings of the 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*. IEEE, 2024.
5. S. Verma, S.P. Sahu, T.P. Sahu. "Wavelet decomposition-based multi-stage feature engineering and optimized

ensemble classifier for stock market prediction. " *The Engineering Economist* (2024): 1-26.
6. D. Ma, D. Yuan, M. Huang, L. Dong. "Vgc-gan: A multi-graph convolution adversarial network for stock price prediction". Expert Systems with Applications 236 (2024): 121204.
7. Z. Chang and Z. Zhang. "*The Composite Multi-Level Prediction Model for Stock Market Time Series*." Artificial Intelligence and Human-Computer Interaction. IOS Press, 2024. 177-186.
8. L. Wenjie, Y. Ge and Y. Gu. "Multi-factor stock price prediction based on GAN-TrellisNet." *Knowledge and Information Systems* (2024): 1-22.
9. L. Antonio Caparrini, J. Arroyo and J. Escayola Mansilla. "S&P 500 stock selection using machine learning classifiers: A look into the changing role of factors." *Research in International Business and Finance* (2024): 102336.
10. D.K. Sharma, H.S. Hota and A.R. Rababaah. "Forecasting US stock price using hybrid of wavelet transforms and adaptive neuro fuzzy inference system." *International Journal of System Assurance Engineering and Management* 15, no. 2 (2024): 591-608.
11. J. Zheng, D. Xin, Q. Cheng, M. Tian and L. Yang. "The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance." *arXiv* arXiv:2402.17194 (2024).
12. P.H. Vuong, L.H. Phu, T.H. Van Nguyen, L.N. Duy, P.T. Bao, T.D. Trinh. "A bibliometric literature review of stock price forecasting: From statistical model to deep learning approach." *Science Progress* 107, no. 1 (2024): 00368504241236557.
13. A. Khanpuri, N. Darapaneni and A.R. Paduri "Utilizing Fundamental Analysis to Predict Stock Prices." *EAI Endorsed Transactions on AI and Robotics* 3 (2024).
14. V. Kanaparthi. "Robustness Evaluation of LSTM-based Deep Learning Models for Bitcoin Price Prediction in the Presence of Random Disturbances." (2024).
15. L. Zeyu. "Stock Price Prediction Using Machine Learning Techniques." *Highlights in Science, Engineering and Technology* 85 (2024): 1122-1126.
16. M. Tajmazinani, H. Hassani, R. Raei and S. Rouhani. "Modeling stock price movements prediction based on news sentiment analysis and deep learning." *Annals of Financial Economics* 17, no. 01 (2022): 2250003.
17. R.J. Kuo, T.H. Chiu. "Hybrid of jellyfish and particle swarm optimization algorithm-based support vector machine for stock market trend prediction." *Applied Soft Computing* (2024): 111394.
18. M. Lu, X. Xu. "TRNN: An efficient time-series recurrent neural network for stock price prediction." Information Sciences 657 (2024): 119951.
19. F. Alzazah, X. Cheng and X. Gao. "Predict market movements based on the sentiment of financial video news sites." In *proceedings of the 2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pp. 103-110. IEEE, 2022.
20. Y. Yifan. "Forecasting Stock Price: A Deep Learning Approach with LSTM And Hyperparameter Optimization." *Highlights in Science, Engineering and Technology* 85 (2024): 328-338.
21. L. Yang and Y. Pan. "A novel ensemble deep learning model for stock prediction based on stock prices and news." *International Journal of Data Science and Analytics* (2022): 1-11.
22. A. Koch, T.L.D. Huynh and M. Wang. "News sentiment and international equity markets during BREXIT period: A textual and connectedness analysis." International Journal of Finance & Economics 29, no. 1 (2024): 5-34.
23. K. Najaf and A. Chin. "The impact of the China Stock market on global financial markets during COVID-19." *International Journal of Public Sector Performance Management* 13, no. 1 (2024): 100-114.
24. J. Ren, H. Dong, A. Popovic, G. Sabnis, J. Nickerson. "Digital platforms in the news industry: how social media platforms impact traditional media news viewership." *European Journal of Information Systems* 33, no. 1 (2024): 1-18.
25. Y. Liu, S. Huang, X. Tian, F. Zhang, F. Zhao and C. Zhang. "A stock series prediction model based on variational mode decomposition and dual-channel attention network." *Expert Systems with Applications* 238 (2024): 121708.
26. S. Veena Madhuri, A. Sirisha and T. Prathima. "AI-Infused Finance: Predicting Stock Prices through News and Market Data Analysis" *ISDA* (2023).