**DAILY NEWS FOR STOCK MARKET PREDICTION**

***Avinash A***

*Division of Mathematics*

*School of Advanced Sciences*

*Vellore Institute of Technology*

*Chennai Tamil Nadu– 600127*

*avinash.a2023@vitstudent.ac.in*

***Dr. David Raj Micheal***

*Division of Mathematics*

*School of Advanced Sciences*

*Vellore Institute of Technology*

*Chennai Tamil Nadu– 600127*

*davidraj.micheal@vit.ac.in*

**Abstract:-**

This project explores the relationship between public sentiment derived from financial news and social media and its impact on stock market performance, specifically focusing on the Dow Jones Industrial Average (DJIA). By employing sentiment analysis techniques on textual data, we aim to quantify sentiment and correlate it with stock price movements. The analysis utilizes various machine learning algorithms to predict stock price trends based on sentiment scores. The results indicate that sentiment can serve as a significant predictor of stock market behavior, providing valuable insights for investors and financial analysts. This study contributes to the growing field of financial analytics by demonstrating the efficacy of sentiment analysis in forecasting stock performance.

**Keywords:-** Sentiment Analysis,Stock Prediction,DJIA,Machine Learning,Financial Analytics,Text Mining,Natural Language Processing (NLP)

**Objectives:-**

The primary objective of this project is to gather and preprocess financial news and social media data relevant to the DJIA to extract meaningful sentiment features. It aims to implement sentiment analysis techniques to derive sentiment scores from the textual data, utilizing tools such as Natural Language Processing (NLP). Additionally, the project seeks to analyze the correlation between sentiment scores and DJIA stock price movements to identify potential predictive patterns. Another key objective is to develop and train machine learning models, such as regression and classification algorithms, to predict stock prices based on sentiment indicators. The performance of these predictive models will be evaluated using appropriate metrics, with a focus on comparing their effectiveness against traditional stock prediction methods. Finally, the project intends to provide insights and recommendations for investors on how to leverage sentiment analysis for informed trading decisions.

**Introduction:-**

The project on sentiment analysis for stock prediction, particularly focusing on the Dow Jones Industrial Average (DJIA), emphasizes the growing importance of sentiment in financial markets. With the rise of social media and online news platforms, public sentiment increasingly influences stock prices, prompting investors and analysts to recognize that market movements are affected not only by fundamental economic indicators but also by collective sentiment. This project aims to explore the relationship between sentiment analysis and stock price prediction by utilizing machine learning techniques to analyze sentiment from various textual sources, such as financial news articles and social media posts. By quantifying sentiment and correlating it with stock market behavior, the project seeks to uncover patterns that could improve predictive accuracy. The findings are expected to provide valuable insights for investors, enabling more informed decisions in a dynamic financial landscape. Ultimately, this research aims to demonstrate the potential of sentiment analysis as a tool for enhancing stock market prediction.

**Review Literature:-**

**Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J., & Lam, W. (1998)** This paper explores a data mining system designed to predict stock market movements by integrating both numerical time series data and textual information from financial news articles. By leveraging sentiment analysis and event extraction, the system improves prediction accuracy, offering more context for stock price movements. It provides real-time forecasts before major Asian markets open, giving investors a competitive edge. This approach highlights the value of combining structured data (like stock prices) with unstructured data (such as news articles) to enhance financial predictions, though challenges like processing complex language and adapting to market changes remain.**Dang, L. M., & Duong, D. M. (2016)** This paper presents a system that predicts daily stock market directions by combining time series analysis with advanced text mining techniques applied to financial news articles. With the rapid growth of online financial news, investors face vast amounts of information that can influence trading decisions. The system aims to quickly process this news data, providing insights to investors before markets react fully to new information. By integrating sentiment analysis and historical stock data, the model achieves up to 73% accuracy in predicting stock trends, highlighting the value of combining textual data with time series analysis for more precise financial forecasting. **Vargas, M. R., de Lima, B. S. L. P., & Evsukoff, A. G. (2017)**This paper investigates the use of deep learning models for predicting intraday directional movements of the S&P 500 index by combining financial news headlines and technical indicators. It explores the strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for analyzing complex patterns in this data. The findings indicate that CNNs are effective at capturing semantic details from text, while RNNs excel at modeling temporal dependencies and context, which are crucial for stock forecasting. The proposed method demonstrates improvements over similar studies, emphasizing the advantage of deep learning in financial prediction tasks.**Khan, W., Malik, U., Ghazanfar, M. A., Azam, M. A., Alyoubi, K. H., & Alfakeeh, A. S. (2019)** This paper explores stock market prediction by integrating social media and financial news data with machine learning and deep learning methods to improve accuracy beyond using historical data alone. After feature selection and spam tweet reduction, the study finds that social media data achieves 80.53% accuracy, while financial news reaches 75.16%. The ensemble of random forest classifiers provides the highest accuracy at 83.22%, proving effective in predicting stock movements influenced by external factors.**Vargas, M. R., dos Anjos, C. E. M., Bichara, G. L. G., & Evsukoff, A. G. (2018)** This paper examines stock price prediction using deep learning models that integrate financial news titles with technical indicators. Two technical indicator sets are compared, with a focus on using CNNs for analyzing news semantics and LSTMs for modeling temporal data patterns. The study compares a hybrid CNN-LSTM model (SI-RCNN) with a standalone LSTM model (I-RNN) for technical indicators, finding that financial news significantly improves prediction stability, while different sets of indicators show minimal impact on performance.**Vargas, M. R., dos Anjos, C. E. M., Bichara, G. L. G., & Evsukoff, A. G. (2018)** This paper provides a comprehensive survey on stock market prediction, addressing challenges in extracting and analyzing news events and historical data amidst high market volatility. It reviews key methodologies in data preprocessing, feature extraction, and deep neural network-based prediction techniques, emphasizing structured text features and opinion extraction. The survey offers a framework for news-sensitive stock prediction, detailing strengths, limitations, and future research directions for enhancing prediction accuracy. **Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019)** This paper enhances stock price prediction accuracy by combining a large dataset of time series stock data with over 265,000 related financial news articles, analyzed using deep learning models. By correlating stock movements with news sentiment, the study addresses the limitations of prior models that used smaller datasets, which impacted accuracy. Utilizing cloud computing for model training and real-time predictions, this approach highlights the importance of extensive data and computational resources in predicting volatile stock prices. **Jiang, W. (2020)** This paper surveys recent advancements in using deep learning models for stock market prediction, focusing on various data sources, neural network architectures, evaluation metrics, and implementation practices. By categorizing recent works, it aims to help researchers stay updated and facilitate reproducibility of existing studies as baselines. The survey also identifies key future research directions to guide further development in this evolving field.

**Methodology**

**Data Preprocessing:** The first step involves gathering and cleaning the dataset, which includes financial news and stock price data. Missing values are addressed through imputation or removal, while text data undergoes tokenization, removal of stop words, and lemmatization to standardize the text. Sentiment analysis techniques, such as using pre-trained models or sentiment lexicons, are applied to convert text into sentiment scores. The resulting numerical features are then normalized or standardized to ensure they are on a similar scale. Finally, the dataset is split into training and testing sets to facilitate model evaluation.

**Linear Discriminant Analysis (LDA):** LDA is employed as a dimensionality reduction technique to enhance the model's performance. The methodology involves calculating the mean vectors for each class and the overall mean, followed by computing the within-class and between-class scatter matrices. The eigenvalues and eigenvectors of these matrices are determined to identify the most significant linear combinations of features that maximize class separability. The reduced feature set is then used for classification tasks.

**Random Forest:** Random Forest, an ensemble learning method, is applied to improve prediction accuracy. The methodology entails constructing multiple decision trees using bootstrapped samples of the training data. Each tree is trained on a random subset of features to introduce diversity. The final prediction is made by aggregating the outputs of all trees, typically using majority voting for classification tasks or averaging for regression. Hyperparameter tuning, such as the number of trees and maximum depth, is performed to optimize model performance.

**Multilayer Perceptron (MLP):** The MLP is implemented as a feedforward neural network consisting of an input layer, one or more hidden layers, and an output layer. The methodology involves defining the architecture of the network, selecting activation functions (e.g., ReLU or sigmoid), and initializing weights. The model is trained using backpropagation and a suitable optimizer (e.g., Adam or SGD) to minimize the loss function. Regularization techniques, such as dropout or L2 regularization, may be applied to prevent overfitting.

**Multilayer Perceptron with Windows:** In this approach, the MLP is adapted to work with sliding windows of data. The methodology involves creating overlapping subsets of the time series data, where each window serves as an input sample. The MLP is then trained on these windows to capture temporal dependencies and patterns. The model's architecture and training process remain similar to the standard MLP, but the input shape is adjusted to accommodate the windowed data.

**Linear Discriminant Analysis with Windows:** LDA is similarly adapted to utilize windowed data for classification. The methodology involves creating overlapping windows of the time series data and computing the sentiment scores for each window. The LDA process is then applied to these windowed features, allowing the model to leverage temporal patterns for improved classification. The model is trained on the windowed dataset, and predictions are made based on the learned linear combinations of features.

**MODEL EVALUATION**

**LDA:-**

The LDA model demonstrates good classification performance with an overall accuracy of 0.83, indicating effective class distinction. While it correctly identifies many true positives and negatives, there are notable false negatives and positives, suggesting areas for enhancement. The classification report reveals balanced performance, with precision and recall values around 0.83 to 0.84, though the balance between class-specific precision and recall could be optimized further. Thus, the model is fairly effective, but there remains potential for fine-tuning to improve its overall performance.

**Random Forest:-**

The Random Forest Classifier has a moderate performance with an overall accuracy of 51%. It shows a higher recall and f1-score for class 1, indicating better identification of positive instances, while class 0 has lower recall and precision, suggesting difficulties in correctly identifying negative instances. The confusion matrix reveals a relatively high number of false positives compared to true negatives, indicating a tendency to misclassify negative instances as positive. Overall, the model performs reasonably but requires further refinement and optimization to improve its accuracy, especially in balancing precision and recall across both classes.

**Multilayer Perceptron:-**

The Multilayer Perceptron (MLP) model demonstrates a moderate performance with an overall accuracy of 50%. It effectively identifies a significant number of positive instances but struggles with a high number of false positives, indicating frequent misclassification of negative instances as positive. Class 0 shows lower precision and recall compared to Class 1, highlighting an imbalance in the model's performance. This suggests that while the MLP model is somewhat effective, it requires improvement in balancing its ability across different classes and reducing false positives to enhance its overall reliability and accuracy.

**Multilayer Perceptron with Windows:-**

The Multilayer Perceptron (MLP) model with windows shows moderate performance with an overall accuracy of 52%. It effectively identifies positive instances but struggles with a high number of false positives and false negatives, indicating frequent misclassifications. Class 0 has lower precision and recall, while Class 1 shows slightly better performance. The model's precision, recall, and F1-score averages around 0.52 suggest that while it performs reasonably well, there is significant room for improvement in balancing its ability across different classes and reducing misclassifications to enhance its reliability and accuracy.

**LDA with Windows:-**

The Linear Discriminant Analysis (LDA) model with windows shows a strong overall performance with an accuracy of 89%. It effectively identifies both positive and negative instances, having a good balance with precision and recall values around 0.88 to 0.90 for both classes. However, there are still some misclassifications, as indicated by 11 false positives and 11 false negatives. Despite these, the model demonstrates high reliability and effectiveness, with only minimal room for improvement to further enhance its performance.

**Conclusion:-**

Overall, the LDA and LDA with windows models demonstrate strong classification performance with high accuracy and balanced precision and recall values, indicating effective class distinction and minimal misclassification. Conversely, the Random Forest and Multilayer Perceptron (MLP) models show moderate performance with lower accuracy, struggling particularly with false positives and imbalanced class-specific metrics. While the MLP with windows improves slightly in terms of accuracy, it still faces significant issues with misclassification. These findings suggest that while LDA-based models are highly reliable and effective, further refinement and optimization are needed for Random Forest and MLP models to enhance their overall performance and reliability.

**References:-**

1) Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J., & Lam, W. (1998). Daily stock market forecast from textual web data. CiteSeer. Retrieved from https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/30606048/24e671e2-cfd0-449d-8e78-01e9b59905a1/p1.pdf

2) Dang, L. M., & Duong, D. M. (2016). Improvement methods for stock market prediction using financial news articles. In Proceedings of the 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (pp. 125-129). IEEE. https://doi.org/10.1109/NICS.2016.7725636

3) Vargas, M. R., de Lima, B. S. L. P., & Evsukoff, A. G. (2017). Deep learning for stock market prediction from financial news articles. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 2017, 1-6. <https://doi.org/10.1109/SMC.2017.8122733>

4) Khan, W., Malik, U., Ghazanfar, M. A., Azam, M. A., Alyoubi, K. H., & Alfakeeh, A. S. (2019). Stock market prediction using machine learning classifiers and social media, news. International Conference on Computing, Networking and Communications (ICNC), 1-6. https://doi.org/10.1109/ICCNC.2019.8685462

5) Vargas, M. R., dos Anjos, C. E. M., Bichara, G. L. G., & Evsukoff, A. G. (2018). A deep learning model combining convolutional and recurrent layers for stock price movement prediction. Journal of Financial Technology, Volume(Issue), Page range.

6) Vargas, M. R., dos Anjos, C. E. M., Bichara, G. L. G., & Evsukoff, A. G. (2018). Event extraction and sentiment analysis for stock market prediction. PeerJ Computer Science, 4, Article 490. https://doi.org/10.7717/peerj-cs.490

7) Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 205-208. https://doi.org/10.1109/BigDataService.2019.00035

8) Jiang, W. (2020). Applications of deep learning in stock market prediction: Recent progress. Preprint submitted to Elsevier Journal. https://arxiv.org/abs/2003.01859