Trend Prediction of DJIA index based on News Extraction from Yahoo Finance

Komal Batool
Department of Mathematics
NED University of Engineering & Technology

Ubaida Fatima, PhD
Department of Mathematics
NED University of Engineering & Technology

Mirza Faizan Ahmed, PhD
Department of Economics & Management Sciences
NED University of Engineering & Technology

ABSTRACT
Decision making in a financial world is a very challenging task for any investor as it can leads towards a very heavy loss as well as very higher returns. Therefore, proper understanding of market behavior is required. It is found in research that movement of prices in financial market is random in nature and depends on multiple factors. In this research sentiment-based prediction of DJIA (Dow Jones Industrial Average) index is performed to forecast the future direction of the indices. The objective behind this research is to analyze if the market is sensitive to news or not and if the web news data contributes in the movement of the market. Five different classification models of machine learning are used which include decision tree, random forest, support vector machine, K-Nearest Neighbor and logistic regression. It is observed that KNN is the best predictive model among all for our dataset with the accuracy of 70%. The results are validated on NASDAQ composite and proved that KNN outperforms other considered classifiers.

General Terms
Sentiment Analysis, Machine Learning, Classification Techniques

Keywords
Trend forecasting, Web news, Stock market analysis, KNN.

1. INTRODUCTION
Stock market prediction is a very attractive domain for the researchers as it helps the traders for decision making. Traders in financial markets trade with the intention of earning maximum possible return with least possible risk. Therefore, the decision-making criteria for the investors are based on estimated future return and risk associated with that return [1].

The forecasting of future returns or risks of financial markets is a challenging task as the movement of the prices depends on several controllable and uncontrollable factors. Financial markets are sensitive to the political and economic conditions of the country, it depends on external factors, opinions and sentiments of traders along with other technical indicators [2].

In order to forecast the future behavior of financial market, several econometrics and machine learning models are used which include both shallow and deep learning modelling. These models are good to understand the insights of movements in price or return to have an efficient prediction. Both regression and classification can be used for the prediction of financial market.

In this research, the international trade market 'Dow Jones Industrial Average (DJIA) index is forecasted. The future direction of the market is predicted that whether the closing price index will go upwards in the future or it remains unchanged or declines. Machine learning classifiers are used for the estimation of future direction.

1.1 DJIA Index
Dow Jones Industrial Average is an American stock index which includes 30 companies of United States. It is one of the most quoted financial markets that include companies of different sectors like industrial, financial, telecommunication, health care and pharmaceuticals, retail, food and many others. The dynamic of this market plays a very significant role in the world economy [3].

1.2 Machine Learning Classifiers
Real-world data has various types on the basis of which procedure of data analysis and mode development is decided. In order to study real world data, both regression and classification can be used depending on the nature of the target variable [4]. Classification using machine learning is performed when the target variable is categorical in nature i-e it contains multiple classes. Machine learning has multiple classification algorithms which are used for data analysis and prediction. The machine learning classifiers include K- Nearest neighbor (KNN), Support Vector Machine (SVM), Ada Boost, Decision Tree, Random Forest and so on. In the section below, an overview of those classifiers is given which are used in this study.

1.2.1 Decision Tree
Decision tree (DT) is a machine learning classifier that firstly selects the features that should be incorporated in the training of the model using any of the algorithms like GINI or ENTROPY which measures the importance of the features. On the basis of importance of the features, this model designs a tree which include nodes (root node, internal node and leaf node) and branches, which is the most important step in building a model [5].

1.2.2 Random Forest
Random forest (RF) is an ensembled form of decision tree model which is designed by merging multiple decision trees to improve the accuracy of the model. It is also included in CART (Classification And Regression Tree) models that is used for both classification and regression. Firstly multiple decision trees are used for the prediction, then by receiving the highest votes for the predicted value the result of random forest is obtained [6].

1.2.3 K- Nearest Neighbor
KNN or K-Nearest neighbor is a classification technique that works by measuring the distance among the datasets using Euclidean and Manhattan distances [7]. KNN is used widely where data description as a feature vector is not possible so a

similarity measure is done for classification [8][9].

1.2.4 Support Vector Machine

SVM is a machine learning classifier that designs a kernel to separate the classes. SVM is efficient to understand both linear and non-linear relationship among the features [10].

1.2.5 Logistic Regression

Logistic regression (LR) is used for two-class classification or for Boolean classification when we have two possibilities in a target variable. It designs a model on the basis of probability of occurrence of one event and the probability of occurrence of second event will be 1 less than the previous probability which is measured by natural logarithm [11].

2. LITERATURE REVIEW

As discussed in previous section, prediction of financial market can be performed using multiple approaches. This is because the movement in financial market is stochastic in nature and it depends on multiple factors. The relation among these factors is complex and non-linear in nature and is therefore difficult to handle [12].

Textual based prediction of financial market has been widely used for the forecasting purpose. This approach determines the impact of textual data on the movement in prices of financial instruments. Textual data may include tweets, discussion in public forums, political or economic news and so on [13]. [14] performed sentiment-based approach for stocks, cryptocurrencies, and Forex (Foreign Exchange Market) analysis. [15] classifies the emotions and finds that the accuracy in forecasting the stock market prices is improved if emotional index of financial news is used to design the training model. In 2017, Rahman and the team used machine learning algorithm to forecast the stock price based on financial news data. The trained model was capable to recognize the emotions from the financial news for the prediction of stock market. It was found that predictive of the stock market is improved if financial news is considered for training a model [16]. Financial news has been used by Manzoor et.al to study the effect of sentiments on financial market. It was concluded that the financial news and sentiment analysis can improve the predictions of financial market. It was also observed during their study that that positive news affects markets positively while negative news has a negative effect on the market [17]. [18] attempted to forecast the trend of financial market using time series analysis and text mining and succeeded to achieve the accuracy of 73%.

3. METHODOLOGY

3.1 Introduction to Methodology

Daily news in a form of textual data is collected from yahoo finance through web scraping using python library 'beautiful soup'. Along with that, daily closing price of DJIA index is fetched from investing.com. Six months daily closing price of DJIA index is collected from November 2023 to April 2024.

3.2 Data Modification

3.2.1 Data from Financial Market

In order to identify the direction of the daily price of DJIA index, firstly change in price is identified by calculating the difference of two corresponding prices as shown in **Error! Reference source not found.**). The change in price is then used to identify the direction of the price to convert the time series data into classes as shown in **Error! Reference source not found.**).

$$C_t = Y_t - Y_{t-1} \quad (1)$$

Where C_t is change in price at time 't'. Y_t is closing price at time 't' and Y_{t-1} is closing price of DJIA index at time 't-1'.

$$D_t = \begin{cases} 0 & \text{if } Y_t \leq 0 \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

Where D_t is direction of price at time 't'. It has two possible values; 0 and 1.

3.2.2 Data from Web News

News data is extracted from yahoo finance on daily basis. This collected data is processed to convert into usable form. In order to perform sentiment analysis, python library NLTK (Natural Language Tool Kit) is used. Using sentiment intensity analyzer negative, neutral and positive polarities are assigned to each news headline. Fig 1 indicates the overall positive, negative and neutral news collected from November 2023 to April 2024. Total count of negative, neutral and positive news on each day is computed and merged with the change in price of DJIA index and is stored as a data frame. Fig 2 shows the first five rows of final prepared dataset that is initially extracted from investing.com (closing price of DJIA) and yahoo finance (daily news).

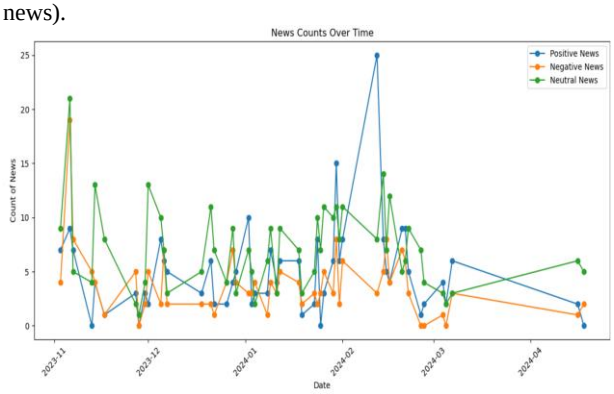

Fig 1: Graphical representation of each news type with respect to time

	date	Neutral	Negative	Positive	DJIA Direction
0	03/11/2023	9	4	7	1
1	06/11/2023	21	19	9	0
2	07/11/2023	5	8	7	1
3	13/11/2023	4	5	0	1
4	14/11/2023	13	4	4	1

Fig 2: First five rows of prepared dataset

3.3 Exploratory data Analysis

In order to understand the insights of the prepared dataset, exploratory data analysis (EDA) is performed. The objective behind EDA is to study the dataset visually for better understandings. The box plot in Fig 3 compares news counts across Neutral, Negative, and Positive sentiments for two Dow Jones Industrial Average (DJIA) directions (0 and 1. Neutral news has the highest variability, while Negative and Positive news show smaller distributions. Outliers are present, and Positive news counts are slightly higher for DJIA direction 1.

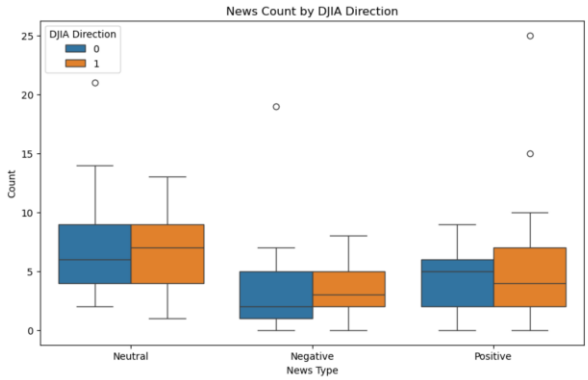

Fig 3: Box-plot of news type against trend of DJIA

3.4 Training of a Model

For training a model, dataset is split. 80% of the data is used for training purpose and rest is for testing the model. Since the target variable has two classes; 0 and 1, therefore classification models are used for the prediction of trend of DJIA index. Five machine learning classifiers are used for training the predictive model which includes decision tree, logistic regression, KNN, support vector machine and random forest.

3.5 Testing of a Model

The designed model is tested using 20% of remaining dataset which was not utilized in training of a model. Confusion matrix is designed in order to calculate precision, recall and accuracy of the model. Fig 5 represents the complete work flow our research.

4. FINDINGS

Five classification models are trained for the prediction of DJIA index future direction. Using the test dataset, the confusion matrices are designed to evaluate the performance of each trained classification model. The accuracy is then calculated using the formula as shown in Equation (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Where, TP and TN is true positive and true negative respectively. FP is for false positive and FN is false negative.

Using the calculated accuracy as given in Table 1 all training models are compared to identify the best one for our dataset. It is observed that decision tree gives the least accuracy of 50% whereas K-Nearest Neighbor outperforms among other classifiers with the maximum accuracy of 70% [Fig 4].

The same models are trained on NASDAQ composite for the validation of the models and it is observed that KNN again outperforms in prediction of the trend of NASDAQ.

Table 1. Obtained Accuracy of all trained models

S.No	Model Name	Accuracy of DJIA	Accuracy of NASDAQ
1.	Decision Tree	0.50	0.45
2.	Logistic Regression	0.60	0.55
3.	Random Forest	0.60	0.54
4.	SVM	0.55	0.55
5.	KNN	0.70	0.56

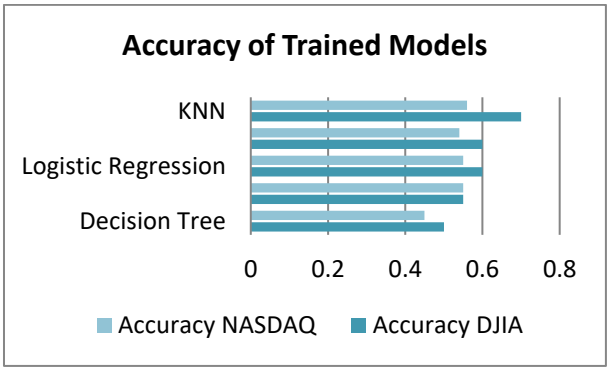

Fig 4: Accuracy of trained models

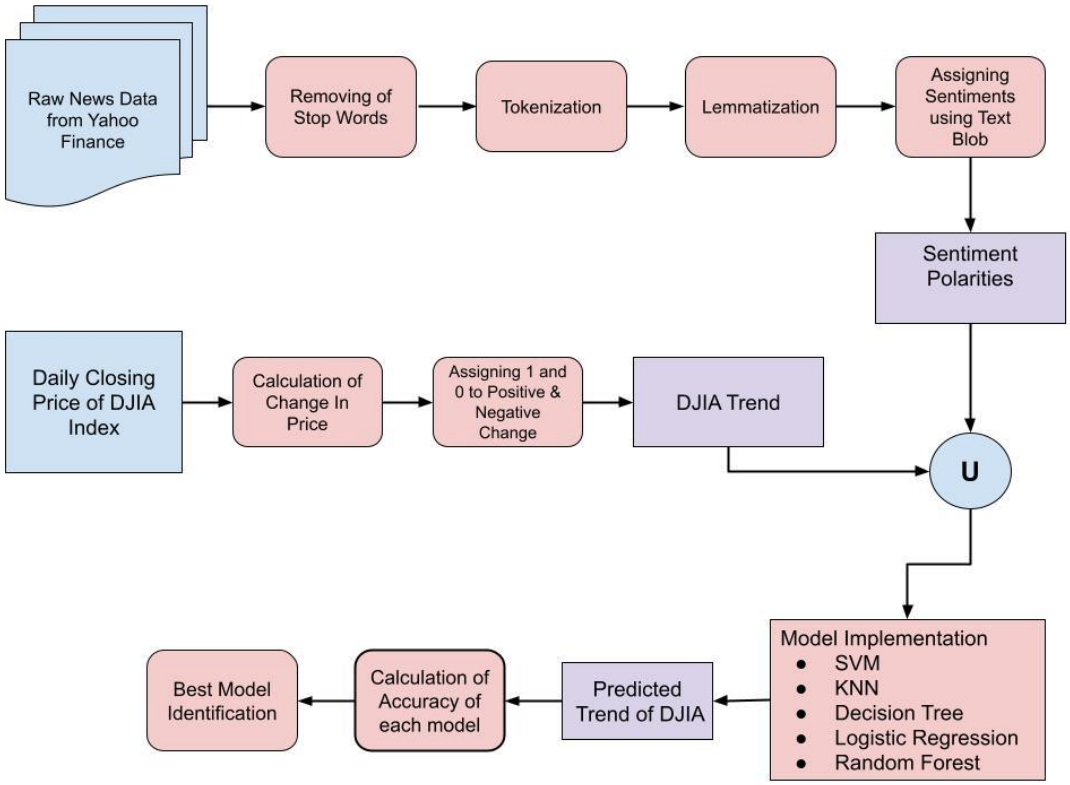

Fig 5: Proposed Work Flow

5. CONCLUSION

As financial market is sensitive to news, therefore impact of good or bad news has an impact on any financial market, particularly equity market. In this research, sentiment analysis is done to forecast the direction of one of the international stock markets, DJIA index. News sentiment based stock market forecasting is an efficient approach as it captures the effect of overall condition of any economy on a market. In this study, five machine learning classifiers are used to forecast the future movement of DJIA index based on news collected from yahoo finance. The classifiers include random forest, logistic regression, KNN and decision tree. It is found that KNN is the most efficient classifier in this particular scenario to forecast the future direction of DJIA index.

6. FUTURE WORK

As in our research, machine learning models have been used for trend forecasting. The trend prediction can also be performed using several other approaches like using hybrid modelling or deep learning models. Incorporating such models may improve the predictivity of the trend of DJIA index as these models are efficient to dig out complex relationship among the features.

7. ACKNOWLEDGMENTS

We would like to express our gratitude to our university, NED University of Engineering & Technology.

8. REFERENCES

[1] Batool, Komal, Mirza Faizan Ahmed, and Muhammad Ali Ismail. "A Hybrid Model of Machine Learning Model and Econometrics' Model to Predict Volatility of KSE-100 Index." *Reviews of Management Sciences Vol* 4.1 (2022)

[2] Mankar, Tejas, et al. "Stock market prediction based on social sentiments using machine learning." 2018 international conference on smart city and emerging technology (ICSCET). IEEE, 2018.

[3] Soloviev, V.N., A. Bielinskyi, and V. Solovieva. Entropy Analysis of Crisis Phenomena for DJIA Index. in ICTERI Workshops. 2019.

[4] Fatima, Ubaida, Saman Hina, and Muhammad Wasif. "A novel global clustering coefficient-dependent degree centrality (GCCDC) metric for large network analysis using real-world datasets." Journal of Computational Science 70 (2023): 102008.

[5] 2Song, Y.Y. and Y. Lu, Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry, 2015. 27(2): p. 130-5.

[6] Biau, G. and E. Scornet, A random forest guided tour. TEST, 2016. 25(2): p. 197-227.

[7] Wang, Yiwen, et al. "Improved KNN-based Stock Price Prediction." Academic Journal of Computing & Information Science 7.6 (2024): 38-43.

[8] Cunningham, P. and S.J. Delany, K-nearest neighbour classifiers-a tutorial. ACM computing surveys (CSUR), 2021. 54(6): p. 1-25.

[9] Siddartha Reddy, A., et al. "Stock Market Trend Prediction Using K-Nearest Neighbor (KNN) Algorithm." (2024).

[10] Awad, M. and R. Khanna, Support Vector Machines for Classification, in Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, M. Awad and R. Khanna, Editors. 2015, Apress: Berkeley, CA. p. 39-66.

[11] LaValley, M.P., Logistic Regression. Circulation, 2008. 117(18): p. 2395-2399.

[12] Chang, P.-C., et al., A neural network with a case based dynamic window for stock trading prediction. Expert Systems with Applications, 2009. 36(3, Part 2): p. 6889-6898.

[13] Bharathi, Shri, and Angelina Geetha. "Sentiment analysis for effective stock market prediction." International Journal of Intelligent Engineering and Systems 10.3 (2017): 146-154.

[14] Farimani, Saeede Anbaee, Majid Vafaei Jahan, and Amin Milani Fard. "From text representation to financial market prediction: A literature review." Information 13.10 (2022): 466.

[15] Bi, J., Stock market prediction based on financial news text mining and investor sentiment recognition. Mathematical Problems in Engineering, 2022. 2022(1): p. 2427389.

[16] Ab. Rahman, A.S., S. Abdul-Rahman, and S. Mutalib. Mining Textual Terms for Stock Market Prediction Analysis Using Financial News. in Soft Computing in Data Science. 2017. Singapore: Springer Singapore.

[17] Manzoor, N., D.S. Rai, and S. Goswami. Stock Exchange Prediction Using Financial News and Sentiment Analysis. in Proceedings of Integrated Intelligence Enable Networks and Computing. 2021. Singapore: Springer Singapore.

[18] Dang, Minh, and Duc Duong. "Improvement methods for stock market prediction using financial news articles." 2016 3rd National foundation for science and technology development conference on information and computer science (NICS). IEEE, 2016.