Integrated Technical and Sentiment Analysis Tool for Market Index Movement Prediction, comprehensible using XAI

Harit Bandi[Φ]
Department of Electronics and Telecommunication
Sardar Patel Institute of Technology
Mumbai, India
[Φ]harit.bandi@spit.ac.in

Suyash Joshi[∞]
Department of Electronics and Telecommunication
Sardar Patel Institute of Technology
Mumbai, India
[∞]suyash.joshi@spit.ac.in

Siddhant Bhagat[σ]
Department of Electronics and Telecommunication
Sardar Patel Institute of Technology
Mumbai, India
[σ]siddhant.bhagat@spit.ac.in

Prof. Dayanand Ambawade[Π]
Department of Electronics and Telecommunication
Sardar Patel Institute of Technology
Mumbai, India
[Π]dd_ambawade@spit.ac.in

***Abstract*—The Stock market has always been an attractive investment avenue. Though the investor has to perform technical analysis, sentiment analysis, and fundamental analysis in order to make informed decisions in stock market investments. This paper proposes an end-to-end system to simplify the process of investment research for the user by providing a holistic view of market index trends based on technical as well as news analysis. Various time-series algorithms have been tested and consolidated by news-based analysis to generate comprehensive data-driven recommendations. An interactive dashboard for the same gives insights for the short term trend of Nifty50 which is chosen for the hypothesis of this paper. The news classification is performed on news related to the index and it is powered by the principle of Explainable AI, incorporating the notion of ethical usage of data-based decision methods.**

***Index Terms*—Explainable AI (XAI), technical analysis, news classification, stock market index, Nifty50**

I. Introduction

Stock market is an investment avenue wherein majority of investors tend to lose money instead of gaining. The primary reason for this is the lack of analysis which this financial instrument demands. The fact is many investors willing to invest in the stock market do not have enough time and necessary knowledge to get positive returns from their investments consistently.

This paper presents a solution to this problem through a short term trend recommendation system which not only recommends the user what is to be done but also explains why it is to be done. The emerging concept of Explainable AI combines the human capabilities of decision making with the arduous task of data crunching performed by machines.

II. Literature Survey

The fact that financial news can lead to fluctuations in stock prices is well introduced. Linyi Yang et. al. in [1] have used the concept of event embedding to capitalise on this fact. They have used Sentence Embedding Natural Language Processing (NLP) Techniques to process a week-long data stream until the day of prediction. Also incorporated is a day-embedding feature which accounts for event embedding on different days of the week. The concept of input and output attention layers, regulated using Gated Recurrent Units (GRU) is used to affect the solution. The ability to provide a human-understandable explanation for every result given by the model (on the lines of XAI to discover hidden layer's approach) is attempted. All related news, and not only target company news are taken into account to make predictions. Taken into account are the effects of "time of the day" and "day of the week" parameters to make a more informed decision. However, this paper has not utilised any quantitative model for prediction and hence leaves a scope for a greater generalization.

Technical, technological as well as Sentiment analysis techniques are basically all that can be done to ensure an exhaustive prediction engine. In [2], such a solution has been discussed. The paper delivers solutions of two kinds. Firstly, prediction of positive/ negative day and secondly, generating buy/ sell calls on Nifty for profit-making decisions. For the prediction methods, technical analysis using Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI) and Exponential Moving Averages (EMA) are considered. For technological indication, back-propagated neural network models are considered. The combination of these two with the output of Sentiment Analysis on popular information sources, gives the results for the first part. The next part of

the paper is executed using the concept of ”support” and ”resistance”, on an intraday basis. The novelty of this paper is the integration of a wide range of predictors and being able to deliver single-weighted prediction system.

The use of Explainable AI technique on time series analysis has been conducted in [3] and the use of this user friendly technique has been discussed. In this paper a selected XAI method creates explanations for every sample of the test data. Based on the time point relevance by the explanations, the test data gets changed by the evaluation and verification methods. The newly created test set gets predicted by the model, and the quality measure is calculated for the comparison. The novelty of this paper is in the various XAI methods on time series and the first evaluation of selected methods on a variety of real-world benchmark datasets. Two sequence verification methods and a methodology which evaluate and check XAI explanations on time series automatically has been presented.

Machine learning classifiers can be useful for making stock market predictions based on the news and media outlets. Such work has been presented in [4] and stock sentiment analysis has been performed. In this paper, algorithms on social media and financial news data have been used to discover the impact of this data on stock market prediction accuracy for 10 subsequent days. For improving performance and quality of predictions, feature selection and spam tweets reduction are performed on the data sets. Some specific experiments are performed to find such stock markets that are difficult to predict and those that are more influenced by social media and financial news. Segregation of specific stocks into classes is the novelty factor in this research. Stocks influenced by social media and financial news have been classified which enhances the specificity and improves the accuracy of the research.

Trusting a machine learning model and before that interpreting its prediction go hand in hand. The paper [5] delineates a novel method to explain the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction in a non-redundant way, framing the task as a submodular optimization problem. This framework provides flexibility in explaining different models for text (e.g. random forests) and image classification (e.g. neural networks) only, facilitating no room for time series and sequential models. After getting insights from explanations, issues with the dataset can be identified and the model can be fixed with further feature engineering. The LIME framework provides with the decision factor for model choice, assessing trust, improving untrustworthy models, and getting insights into predictions.

The paper [6] presents a system for forecasting stock price returns. By using MKSVR (Multi-Kernel Support Vector Regression), the system quantitatively analyzes and integrates intraday market news and stock tick price. Experiments have been conducted by using the market news and tick data of the Hong Kong stock market over one whole year. Results have demonstrated that the MKSVR is capable of making the better use of hidden information in news articles and historical stock prices than the models that simply use news articles to forecast stock prices. MKSVR is a regression approach. The outputs of MKSVR are numerical values rather than user defined categories. This would overcome the signal strength bias. MKSVR learns the weights of sub-kernels and determines which information source is more effective in prediction, where the weights could be interpreted as the extent to which one information source contributes to the prediction.

III. Methodology

The methodology of this research project can be classified into three categories. The first is data collection and preparation of final usable data sets through data pre-processing. The second is using technical analysis for price prediction based on the collected technical data. The final category pertains to the sentiment analysis performed on news headlines and their descriptions. The insights from technical and sentiment analysis are delivered through an interactive dashboard. The above mentioned process is explained in detail further.

A. Data Collection

The most primary requirement to predict the trend of an index is to collect data of appropriate size and granularity. The size of the data set plays a crucial role in training of the model while high granularity improves accuracy. The data sets having high granularity are paid and have significant costs. Only the primary factors such as OHLC (Open-High-Low-Close prices) are available for free. Using the technique of web scraping, the issue of high cost for data was mitigated as real time data was collected from Yahoo Finance.

To achieve the necessary granularity for technical data collection, a web scraping script for both index value and news headlines related to Nifty50 has been developed. BeautifulSoup is a python library used for parsing HTML (HyperTextMarkup Language) documents. Using this library, the necessary data was collected and processed as per requirement to tabulate it into the final data set.

1) Price Data: The index price data consists of various columns such as date, timestamp, index value, change in index value and percentage change in index value as compared to closing price of previous trading session. A sample snippet of the data frame can be observed in Fig. 1. A data point was recorded for every half a second which satisfies the need of granularity and enhances the accuracy of predictions made by technical analysis.

2) News Data: The news related to the Nifty50 index emerge not only during the trading session but also before the opening and post the closure of market hours. Hence, the web scraping script is designed such that it runs throughout the day and fetches the headline, short description, and timestamp of every news. The data was collected from ET Markets and a snippet of news data can be seen in Fig. 2.

B. Data Preprocessing

The process of data collection was optimized such that it would require minimum processing to convert it into the final data set. The technical data recorded timestamp for each data

	Date	Time	Nifty50	Change(Value)	%Change
0	2020-09-21	14:03:40.577134	11,370.35	-134.60	-1.17%
1	2020-09-21	14:03:41.101858	11,370.35	-134.60	-1.17%
2	2020-09-21	14:03:41.628413	11,370.35	-134.60	-1.17%
3	2020-09-21	14:03:42.149413	11,369.60	-135.35	-1.18%
4	2020-09-21	14:03:42.708271	11,369.60	-135.35	-1.18%
5	2020-09-21	14:03:43.164392	11,369.75	-135.20	-1.18%
6	2020-09-21	14:03:43.632253	11,369.75	-135.20	-1.18%
7	2020-09-21	14:03:44.120868	11,369.75	-135.20	-1.18%
8	2020-09-21	14:03:44.773351	11,371.70	-133.25	-1.16%
9	2020-09-21	14:03:45.291663	11,371.45	-133.50	-1.16%
10	2020-09-21	14:03:45.805292	11,371.45	-133.50	-1.16%
11	2020-09-21	14:03:46.295751	11,371.45	-133.50	-1.16%
12	2020-09-21	14:03:46.822195	11,372.35	-132.60	-1.15%
13	2020-09-21	14:03:47.301415	11,372.35	-132.60	-1.15%

Fig. 1. Scraped technical data

	Heading	Timestamp	Description
0	Lupin Ltd. shares up 0.12% as Nifty gains	Sep 08, 2020, 11.28 AM	A total of 31,930 shares changed hands on the ...
1	Shares of Indraprastha Gas Ltd. as Nifty gains	Sep 08, 2020, 11.44 AM	On the technical charts, the 200-day moving av...
2	Power Finance Corporation Ltd. shares falls 1...	Sep 08, 2020, 12.05 PM	A total of 80,765 shares changed hands on the ...
3	Share price of Berger Paints (India) Ltd. rise...	Sep 08, 2020, 12.10 PM	A total of 11,037 shares changed hands on the ...
4	Share price of Crompton Greaves Consumer Elect...	Sep 08, 2020, 12.21 PM	A total of 2,178 shares changed hands on the c...
5	Divi's Laboratories Ltd. shares down 1.38% as ...	Sep 08, 2020, 12.41 PM	A total of 16,468 shares changed hands on the ...
6	Share price of Crompton Greaves Consumer Elect...	Sep 08, 2020, 12.21 PM	A total of 2,178 shares changed hands on the c...
7	Divi's Laboratories Ltd. shares down 1.38% as ...	Sep 08, 2020, 12.41 PM	A total of 16,468 shares changed hands on the ...
8	Havells India Ltd. shares down 0.79% as Nifty ...	Sep 08, 2020, 01.07 PM	A total of 17,820 shares changed hands on the ...
9	Ambuja Cements Ltd. shares down 1.0% as Nifty ...	Sep 08, 2020, 01.23 PM	A total of 28,535 shares changed hands on the ...
10	Shriram Transport Finance Company Ltd. shares ...	Sep 08, 2020, 01.54 PM	A total of 82,120 shares changed hands on the ...
11	Share price of ACC Ltd. falls as Nifty strengt...	Sep 08, 2020, 02.10 PM	A total of 21,444 shares changed hands on the ...
12	Share price of Bharti Infratel Ltd. falls as N...	Sep 08, 2020, 02.15 PM	A total of 686,477 shares changed hands on the...
13	Sensex erases intraday gains, ends 52 points l...	Sep 08, 2020, 04.15 PM	CLOSING BELL: Sensex erases intraday gains, en...
14	Share price of Bharti Infratel Ltd. falls as N...	Sep 08, 2020, 02.15 PM	A total of 686,477 shares changed hands on the...
15	Sensex erases intraday gains, ends 52 points l...	Sep 08, 2020, 04.15 PM	CLOSING BELL: Sensex erases intraday gains, en...

Fig. 2. Scraped news data

point up to an accuracy of 6 decimal places which was not necessary for the final data set. In case of the news data, the headlines and short descriptions were concatenated to get the entire text under one column. The text had to undergo a long process of lower casing, removal of stop words, removal of punctuation, removal of non-useful words, and lemmatization to make it usable for Explainable AI model. The news were labelled as bullish, bearish, or neutral based on the nature of the news.

IV. Experimental Results and Discussions

A. Data Sets

The final technical dataset contains more than 28000 data points having columns as illustrated in Fig. 1. The data points were recorded during the market hours from 9:15am to 3pm on Sept 8, 2020.

The news dataset consists of more than 1000 news from the same website which are related to Nifty50. The model was trained on all the news except those emerging during the prediction time zone.

B. Price prediction using historical data

A large number of regression techniques are available at our disposal, each method having its own shortcomings. Also, there are a number of factors which are a characteristic of the source datasets which ascertain the applicability and efficiency of these models. The research project compares different models and their performance to finalise the most optimum model.

*1) **Artifical Neural Networks (ANN):*** An artificial neural network [7] [8] is a biologically inspired computational model formed from hundreds of single units, artificial neurons, connected with coefficients (weights) which constitute the neural structure. They are also known as processing elements (PE) as they process information. Each PE has weighted inputs, transfer function and one output. PE is essentially an equation which balances inputs and outputs. ANNs are also called connectionist models as the connection weights represent the memory of the system.

As is evident from the organisation of a ANN system, they rely on making sense of complex relationships between predictor variables. Similar to all Machine Learning techniques, ANNs can also be used for Supervised as well as Unsupervised learning. The goal in supervised learning is to predict one or more target values from one or more input variables. Supervised learning is a form of regression that relies on example pairs of data: inputs and outputs of the training set. In unsupervised training, the network is provided with inputs but not with desired outputs. The system itself must then decide what features it will use to group the input data. This is often referred to as self-organization or adaptation. The self-organising behaviour may involve competition between neurons, co-operation or both [9]. ANNs are particularly found to be useful in supervised learning tasks with a back-propagation learning rule. The performance evaluation of a ANN model for our project is depicted in Table I. The plot of actual normalised data points versus the predicted values by model are shown in Fig. 3.

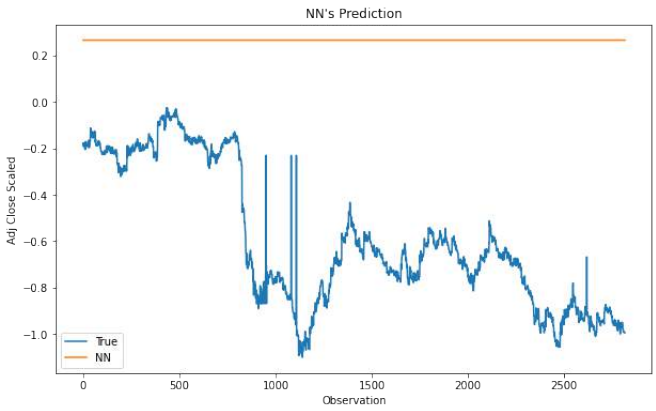

Fig. 3. Normalised plot for actual vs predicted values for ANN Model

TABLE I
PERFORMANCE EVALAUTION OF ANN MODEL

Test Parameter	Train Set	Test Set
R^2 Score	-0.002	-8.379

*2) **ARIMA**:* Exponential smoothing and ARIMA are two of the most widely used time-series forecasting methods. In this section discusses the ARIMA model. The acronym ARIMA stands for Auto-Regressive Integrated Moving Average. There are seasonal and Non-seasonal ARIMA models that can be used for forecasting. A non-seasonal ARIMA model is a function of three variables.

$$ARIMA(p, q, d)$$

P = Periods to lag for eg: (if P= 3 then the three previous periods of our time series in the autoregressive portion of the calculation) are used. P helps to adjust the line that is being fitted to forecast the series.

Q = This variable denotes the lag of the error component, where error component is a part of the time series not explained by trend or seasonality.

D = An ARIMA model transforms a time series into stationary one(series without trend or seasonality) using differencing. D refers to the number of differencing transformations required by the time series to get stationary.

The Autoregressive Integrated Moving Average (ARIMA) model converts non-stationary data to stationary data before working on it. Using a ARIMA model-based forecasting method requires predefined processing to ensure prediction accuracy. Firstly, it is necessary to check if a series is stationary or not, because time series analysis only works with stationary data. For this purpose, *Augmented Dickey-Fuller Test* is conducted. Next, separating seasonality and trend from the series may be required. Next, the data is split into training and testing sets. Performance plots will be analysed to finalise on the value of p,q and d variable values and see the performance of the diagnostic plots, namely, the *Standardized Residual plot*, *Histogram incorporated estimated density plot*, *Sample Quantiles vs Theoretical Quantiles* and *Correlogram*.

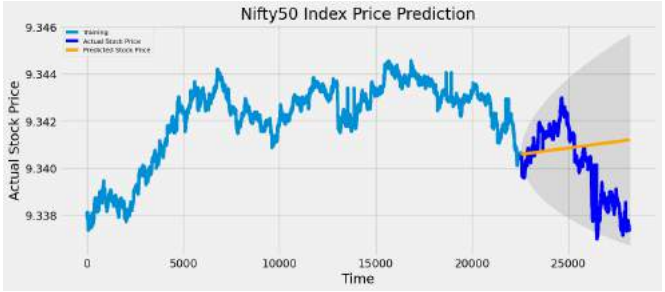

Fig. 4. Prediction using ARIMA Model

The final step would be to forecast the predicted variable on the test dataset keeping 95% confidence level. Some disadvantages due to nature and frequency of data points lead to poor predictions when this model is used. The outcome of these steps can be summarised via the diagnostic plots shown in Fig. 5. Also, the plot of the training and test data points versus the data points predicted by the ARIMA model, indicated by different colour schemes are shown in Fig. 4.

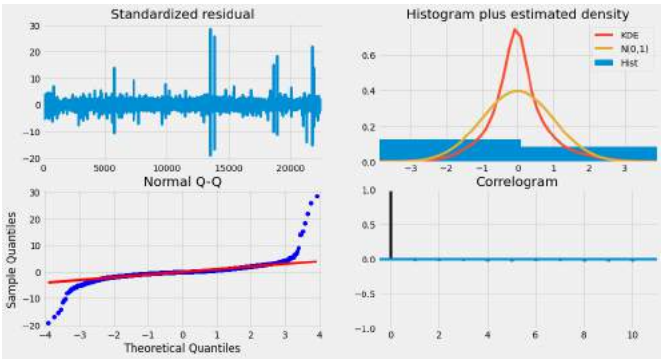

Fig. 5. Diagnostics Plots for ARIMA Model deployed

Model Performance for our use case in shown in Table II.

TABLE II
PERFORMANCE EVALAUTION OF ARIMA MODEL

Test Parameter	Value
Mean Absolute Error	0.00159
Mean Squared Error	3.586 x 10^{-6}
Root Mean Square Error	0.0019
Mean Absolute Percentage Error	0.00016

*3) **LSTM**:* Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning.

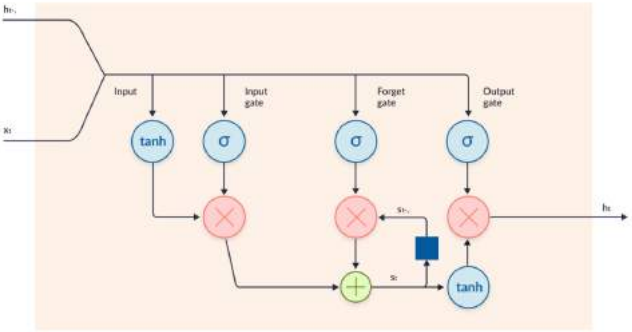

Fig. 6. LSTM Cell Structure

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the

flow of information into and out of the cell. A basic LSTM cell structure is shown in Fig. 6.

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.

The advantage of an LSTM cell compared to a common recurrent unit is its cell memory unit. The cell vector has the ability to encapsulate the notion of forgetting part of its previously stored memory, as well as to add part of the new information. To illustrate this, one has to inspect the equations of the cell and the way it processes sequences under the hood.

Common application areas are unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic. Performance evaluation of a LSTM model for our case is shown in table III. Also, the plot of the training and test data points versusthe data points predicted by the LSTM model, indicated bydifferent colour schemes are shown in Fig. 7.

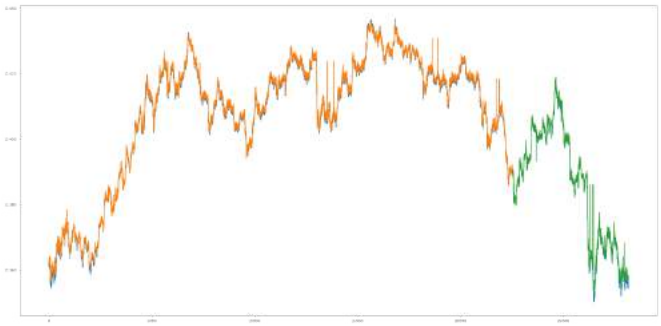

Fig. 7. Prediction plot using LSTM Model

TABLE III
PERFORMANCE EVALAUTION OF LSTM MODEL

Test Parameter	Train Set	Test Set
Root Mean Square Error	0.83	1.33

C. News Text Classification with Explanations

The labelled news dataset is used to train a random forest classifier model. The pipeline includes Count Vectorizer followed by TF-IDF Transformer. The predictions of the model to differentiate between “Bearish”, “Neutral” and “Bullish ” are explained. Although this classifier achieves 81% accuracy, and one might trust the prediction based on this, the explanation for an instance may be made for quite arbitrary reasons. Hence, Local Interpretable Model-agnostic Explanations (LIME) framework is used to provide interpretable representation for text classification, a vector indicating the presence or absence of a word, even though the classifier may use more complex (and incomprehensible) features such as word embedding.

In order to ensure both interpretability and local fidelity, L(f, g, πx) needs to be minimized while having Ω(g) be low enough to be interpretable by humans. The explanation produced by LIME is obtained by the following: ξ(x) = argmin g ϵ G L(f, g, x) + Ω(g).

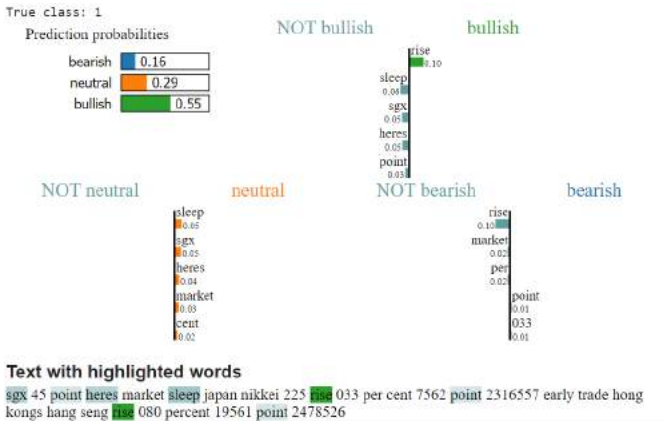

Fig. 8. Explanation for a ”Bullish” news correctly classified as ”Bullish”

The model predicts that the news snippet in Fig. 8 is “Bullish”, and LIME highlights the keywords in the news that led to the prediction. The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to. Rise is portrayed as contributing to the “Bullish” prediction, while “sleep” and ”sgx” is evidence against it. With these, a user can make an informed decision about whether to trust the model’s prediction.

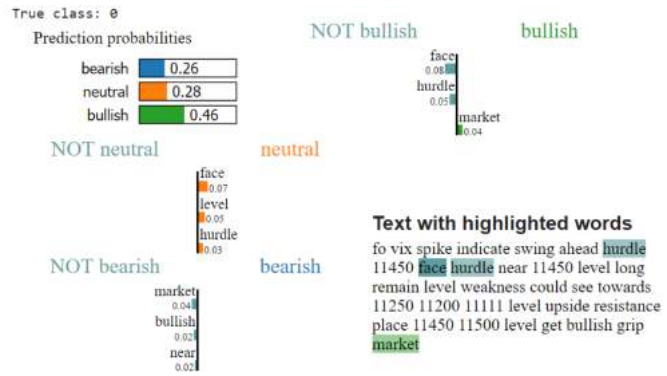

Fig. 9. Explanation for a ”Neutral” news wrongly classified as ”Bullish”

The news may be wrongly classified as well. For example, in the Fig. 9 the true class of the news is “Neutral” but the model wrongly predicted it as a “Bullish” new (words “market” and “bullish” have no connection to Neutral class). The word “market” appears in 161 out 1000 news in the training set, most of them in the class “Bullish”.

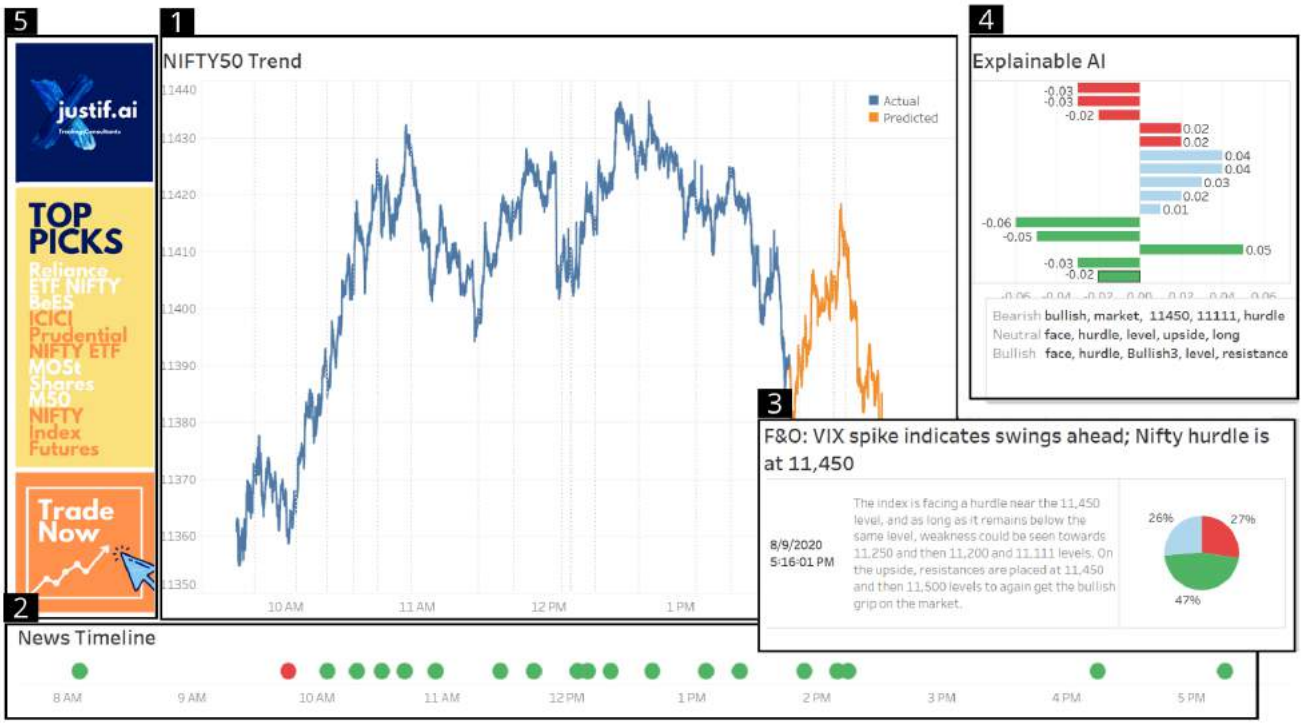

Fig. 10. Snippet of the end-user dashboard

D. Experimental Results and Discussions

It is evident that the classification is not completely reliable considering the fact that the classifier associates some frequently occurring words to a particular class with no solid relevance. Thus steps such as removal of frequently occurring irrelevant words were taken to improve classification accuracy.

E. Dashboard for end-user

The purpose of this project is to simplify the process of analysis for the end user. The interactive dashboard developed using Tableau provides informative insights in a lucid manner through a combination of visuals which can be seen in Fig. 10.

The first pane indicates the trend of Nifty50. The index value is plotted against time and segregated as actual and predicted. The actual part represents the real index value during the trading session while the predicted part represents the output of prediction by technical analysis. A user can simply hover the cursor to get basic details such as index value about the index. Vertical dotted lines in the dashboard indicate the emergence of a news at a particular time instant.

The second pane showcases various news along with their timestamps through green or red circles. The colour of the circles indicates whether the news is bullish or bearish and gives a quick idea of market sentiment to the end user.

The third pane can be accessed by clicking on any of the news data points in the News Timeline. The details of a particular news such as the headline, timestamp, news description can be seen in the pop-up visual. The visual also contains a pie chart which indicates the percentage split of the news being bullish, bearish or neutral using conventional intuitive colours.

The fourth pane explains to the user why a particular news has been classified as bearish, bullish or neutral. Each of the categories mentioned earlier displays top 5 influential key words for that class and they are highlighted in the tooltip. The contribution of their weights towards the news, either positive or negative, can be seen in the horizontal bar plots.

The fifth pane in the dashboard contains 3 images. Clicking on the first image redirects users to the website of the company. The second image displays top picks for the user and takes the user to NSE website to get further details. Clicking on the final image redirects users to a trading platform where they can open an account and start trading.

V. Conclusion and Future Scope

The paper primarily focuses on two pillars of stock market analysis which are technical and sentiment analysis. The hypothesis of this paper is focused on the Nifty50 index and thus does not directly depend on fundamentals of a particular company but a group of 50 companies which are constituents of the index. The hypothesis can be further extended to a particular company wherein incorporating fundamental analysis could be vital to achieve higher accuracy of predictions.

This paper simplifies the analysis of the stock market for an investor willing to invest in the stock market but lacking

the necessary knowledge and time to analyze before investing. The concept of Explainable AI lets the user make decisions on his own by transparently displaying the reasons behind the prediction. The user friendly dashboard provides insights which helps in decision making and saves time for the user which in turn benefits financial markets as a whole.

REFERENCES

[1] Yang, Linyi Zhang, Zheng Xiong, Su Wei, Lirui Ng, James Xu, Lina Dong, Ruihai. (2018). Explainable Text-Driven Neural Network for Stock Prediction. 441-445. 10.1109/CCIS.2018.8691233

[2] A. A. Bhat and S. S. Kamath, "Automated stock price prediction and trading framework for Nifty intraday trading," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-6, doi: 10.1109/ICCCNT.2013.6726675

[3] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A. Keim, "Towards A Rigorous Evaluation Of XAI Methods On Time Series", 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), DOI: 10.1109/ICCVW.2019.00516

[4] Wasiat Khan, Mustansar ali Ghazanfar, Muhammad Awais Azam, and Amin Karami, "Stock market prediction using machine learning classifiers and social media, news", March 2020 Journal of Ambient Intelligence and Humanized Computing, DOI: 10.1007/s12652-020-01839-w

[5] Ribeiro, Marco Singh, Sameer Guestrin, Carlos. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 97-101. 10.18653/v1/N16-3020

[6] Li, Xiaodong Huang, Xiaodi Deng, Xiaotie Zhu, Shanfeng. (2014). Enhancing Quantitative Intra-day Stock Return Prediction by Integrating both Market News and Stock Prices Information. Neurocomputing. 10.1016/j.neucom.2014.04.043

[7] J. Zupan, J. Gasteiger, Anal. Chim. Acta 248 (1992) 1–30

[8] J.M. Zurada, Introduction to Artificial Neural System, PWS, Boston, 1992

[9] Kustrin, Snezana Beresford, Rosemary. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis. 22. 717-27. 10.1016/S0731-7085(99)00272-1