**ABSTRACT**:

Stock market prediction has been one of the most interesting use cases of machine learning for a long time. Stock market prediction involves both fundamental and technical analysis of a particular stock. Fundamental analysis of stocks itself depends on multiple factors – physical factors, behavioral economics, news about a particular stock, demonetization, monetary policy, natural disaster, etc. Existing studies have shown that there has been a strong correlation between news article, blogs, and stock price momentum. The project aims to use text mining (Natural Language techniques) to predict stock market momentum and time-series data to further improve the accuracy of prediction. The time series model will use Long Short-Term Memory (LSTM) neural network to predict future stock close price which will cover the technical analysis of stock momentum. The hybrid approach of using both sentiment analysis and the LSTM model will result in high predictability of the stock market.

Keywords- NLP, LSTM, Neural Network, Stock Market, Sentiment Analysis.

## INTRODUCTION:

The stock market has a direct impact on the country's economy, financial institutions, and individual investors. The growth of the stock market of a particular country gives the projection of financial investments both domestic and international. However, due to the highly volatile nature of the stock market and dynamic information flow, the prediction of the stock market remains a difficult topic in the machine learning domain. Several attempts in the past have been made to accurately predict the momentum of the stock using Support Vector Machine (SVM), Linear regression, and other machine learning techniques but due to the limitation of these models the accurate prediction of stock momentum has not been achieved. Moreover, studies have shown that there has been a strong correlation between news articles and stock prices.

The project aims to adopt a hybrid approach where firstly sentiment analysis on multiple news feeds from different financial websites (Bloomberg News, Yahoo finance, google finance, MarketWatch, Reuters, finviz) will be performed. Financial data will be gathered by calling APIs, parsing the data, and finally performing sentiment analysis using NLP techniques. The compounded score obtained for a particular stock for different news feeds for a particular day will then be averaged to obtain the final score. Based on this average value predictions can be made for a stock momentum in future. Finally, the realized predictions can then be analyzed visually with the help of Matplotlib, Plotly, or Seaborn.

In the second stage, the time series historic data will be gathered from multiple sources will be used as input for Recurrent Neural Network (RNN). For the technical analysis, Recurrent Neural Network (RNN) algorithm LSTM will be used to analyze at least 60 days of stock price data. Long-Short Term Memory has been one of the most successful RNN architecture and since stock data is time-based meaning present-day prices depend on previous day prices, using RNN is the most suitable model for stock price prediction.

## PROBLEM STATEMENT:

Predicting stock market momentum is a difficult task, given the number of parameters involved. If one can predict the stock market momentum accurately one can easily increase the profits and reduce the losses. The stock market dynamic relies primarily on two analyses, fundamental and technical. The fundamental analysis covers the behavioral patterns of individual investors, financial institutions, and news circulating in the market. Hence it becomes extremely important to analyze the sentiments for performing fundamental analysis. Moreover, technical analysis is mostly performed on day-to-day stock prices. Since stock data is time-based it is useful to do a time-series analysis to perform technical analysis.

## GOALS AND OBJECTIVES:

To analyze and predict stock market data using a hybrid approach, combining Natural Language Processing (NLP) and Recurrent Neural Network (RNN) architecture.

**Objective 1**: Perform literature review on Natural Language Processing, Long-Short Term Memory, and Recurrent Neural Network architecture.

**Objective 2**: Data gathering and data pre-processing for sentiment analysis of news feeds.

**Objective 3**: Transforming data into data dictionary and calculating sentiment score

**Objective 4**: Collecting and transforming data for time series analysis (LSTM)

**Objective 5**: Splitting the data for the training and testing model.

**Objective 6**: Predicting 60 days future closing stock prices using a model built and using Adam optimizer.

**Objective 7**: Finally, comparing results obtained from both the results, NLP and Times Series Analysis, and visualizing stock momentum and price predicted using Matplotlib or seaborn library.

**Objective 8:** Creating an interactive dashboard using Microsoft Power BI to visualize the results in detail.

## LITERATURE REVIEW:

Over the period researchers have explored many ways and methods for predicting the stock market, but recent studies [1] have shown growing interest in using sentiment analysis for stock forecasting. Some studies have also proven a strong correlation between stock prices, news articles, and investor's sentiments. Broadly sentiment analysis can be categorized into lexicon-based analysis and machine learning-based analysis. Lexicon-based analysis can be further classified into corpus-based, and dictionary-based. Machine Learning based analysis can also be classified into supervised and unsupervised.

Dev Shah et al [2] in their research used a dictionary-based approach where they used a python library and transformed the text corpus into numerical vectors. The approach used finally converts the words as positive and negative and assigns a sentiment score for each token of the word. One of the major drawbacks of this approach was that the tokens used for calculating sentiment scores have no weight assigned to them which can lead to false sentiment prediction.

In [3] the author has used bag-of- word approach to perform sentiment analysis using Twitter data and news articles. In this approach, the author labeled the tweets and news articles' headlines as positive, negative, or neutral. For pre-processing data, the author used tokenization, lemmatization, and n-gram methods also the use TF-IDF weighting schema is being used. Although, this method was successfully implemented the bag-of-words approach it ignored sentence structure due to which it failed to capture sentiments. Also, the method failed to capture sentiments of informal text like slang and acronyms.

In 2015 Alostad and Davulcu [4] used a hybrid approach to perform sentiment analysis using N-gram feature extraction They built the document matrix by collecting news articles from the NASDAQ website and applied OpenNLP for sentence extraction. Alostad et al [4] also applied classification techniques like logistic regression and Support Vector Machine (SVM) over the n-gram document matrix to improve the predictability of their model. Although the hybrid method used by the author to predict the stock price sentiments was unique, the document-level sentiment analysis could not improve the accuracy of the prediction.

In 2015 kai et al [5] proposed an LSTM- based approach for predicting the Chinese stock market. They labeled 30 days stock data price using earning rate, which was calculated averaging the 3-day closing price. The model proposed uses a single layer for input/output, multiple LSTM layers, and a dense layer. The feature used for learning includes historic price data of 30 days and the historic price data for market indexes. One of the drawbacks of this approach was that it did not consider the impact of technical analysis and fundamental analysis (GDP, oil prices, etc.). Also, setting the right threshold for effectively excluding the extreme high and extremely low prices resulted in low accuracy for this method.

In 2019 Nikou et al [6] compared different machine learning algorithms such as Random Forest, Support Vector Machine (SVM), Deep Learning, etc. for efficient prediction of the stock market. The result indicated that the LSTM model, a special case of Recurrent Neural Network (RNN) performed better as compared to other models. Also, in 2018 Achkar et al [7] conducted a study to compare the LSTM neural network and Multilayer perceptron Neural Network which further confirmed that the percentage error in the case of LSTM was minimum.

## METHODOLOGY:

### Phase 1: Sentiment analysis on news data

1. **Data Pre-processing:** The data for sentiment analysis will be collected from multiple sources. The news headlines will be collected from(https://finviz.com/) using the python beautifulsoup module. The field which is not related to financial data will be removed during this phase also new articles headlines will be considered for now and not the entire article. The Ticker values will be separated based on timestamps. Since the new articles come with a different timestamp. For example, some of the news has timestamp as yyyy:mm:dd: hh: mm whereas some timestamps are in the format of yyyy:mm: dd. Similarly, tweets from Twitter will also be collected for specific tickers using “SNSCRAPE” library. The duplicate tweets will be removed using drop duplicate ()

function of pandas library. The data received after pre-processing and transformations will be stored in a dictionary or panda's data frame.

2. **Sentiment Analysis:** The data frame obtained from the previous step will hold the stock name, date, time, and articles related to stock. The sentiment analysis will be performed on the news articles using NLTK Vader lexicon. Vader uses the polatiy_scores () method to generate the compounded score for positive, negative, or neutral. The compounded value is a value between [-1, 1] where -1 means extremely negative and +1 means extremely positive. Next, the compounded score for all the articles for a particular stock will be done, and the mean value will be taken for it.

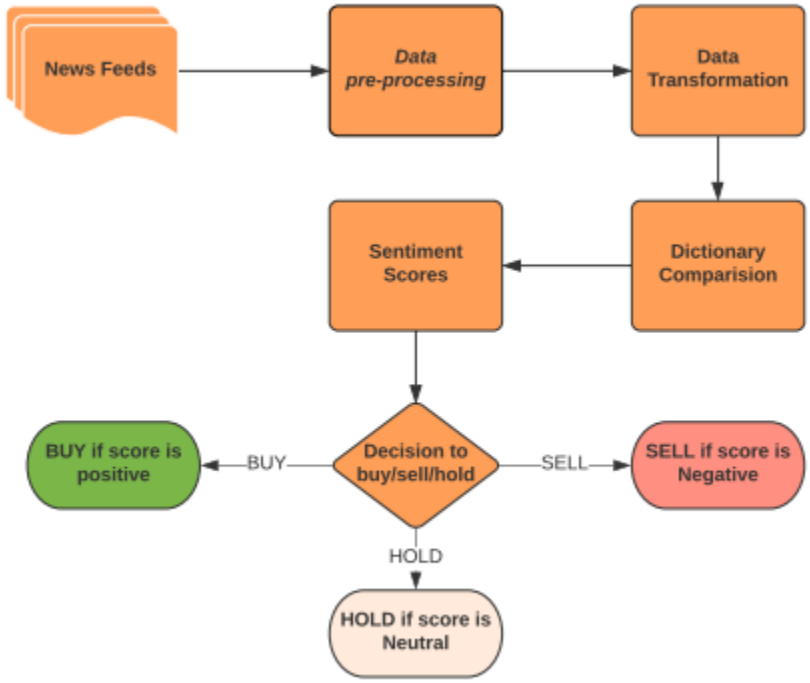


Fig [1]: Performing sentiment analysis on news feeds

**Phase 2: Long Short-Term Memory (LSTM)**

1. **Data Preparation and Scaling**

The input data for the LSTM model is prepared by collecting the historical stock price from January 2015- December 2020 for three specific tickers (Microsoft, Google, and Amazon) by calling API endpoints from Tiingo (https://api.tiingo.com/), TradingView, and Yahoo Finance. The input for the LSTM model will consider only stocks closing and opening

price. LSTM is sensitive to the scale of the data, so the input will be transformed using MinMax scaler between [0,1] with the help of the NumPy library.

2. **Dataset Splitting and Model Construction**

The input data from the previous step will be further split into training and test data. Initially, multiple training and test will be taken into consideration to find the best possible split. To obtain better accuracy during model fitting different values for parameters will be taken into consideration for example Epoch keeps the count of how many times the data is passed through the neural network. In this project since the data is huge the epoch value is taken as 40. Also, the project will make use of the Adams optimizer which is mostly used for building RNN architecture. The detailed flow of model construction is shown in fig [2]

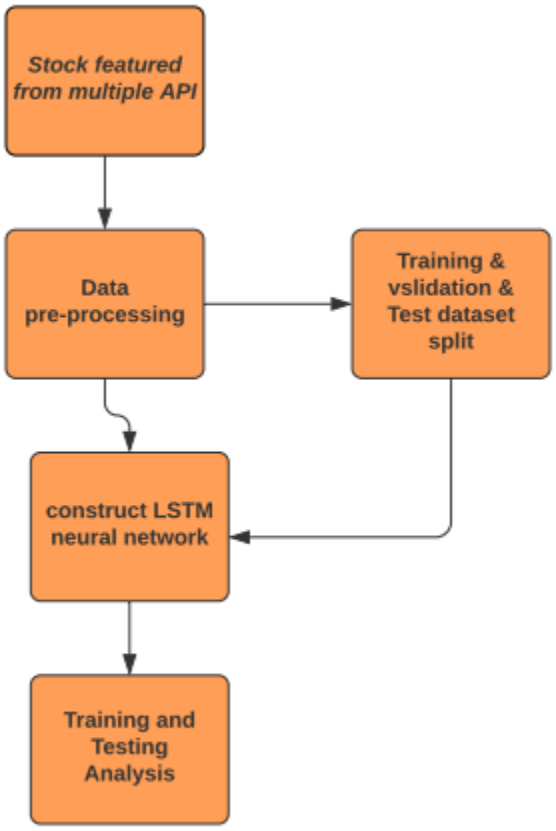


Fig [2] Model construction process

3. **Data Pre-Processing and Model Evaluation**

This phase will consider taking custom timesteps for hyperparameter tunning of LSTM Model and converting train and test data to independent (x-train, y-train) and dependent (x-test, y-test) features. The train and test data will be evaluated separately by running the LSTM model. The project will use Root Mean Squared Error (RMSE) as the key

performance index to evaluate train and test data. The lower the difference between rmse_train and rmse_test the better the model will be. Finally, the output of the train and test data will be plot using Matplotlib which will show the prediction graph of the stock price for test data.

## Phase 3: Combining Sentiment Analysis and Long Short-Term Memory (LSTM)

### 1. Data Collection and Pre-Processing

The data collected from sentiment analysis from phase 1 using news headlines and VADER (valency Aware Dictionary and Sentiment Reasoner) library will be used to calculate the compounded score. For calculating the sentiments of articles/news content another Python library Text Blob [10] will be used which will generate a polarity score ranging from [-1, +1] where +1 represent the positive sentiments and -1 represent negative sentiments. The combination of both compound score and polarity score will help improve the accuracy of sentiment analysis.

Finally, the historical stock data prepared for LSTM model during phase 2 will be combined with sentiment analysis data. The new data set constructed and will then be transformed and scaled to be used as input for new LSTM Model.

### 2. Model construction and Evaluation

The LSTM model constructed during this phase will be similar to the LSTM model constructed during phase 2 with the key difference of combined input data from phase 1 and phase 2. The combined output from Sentiment analysis and LSTM will help improve the accuracy of the LSTM model constructed during this phase. Fig [3] shows the detailed flow of model construction.

For evaluating LSTM Model, the dataset prepared from the previous step will be split into train (65%) and test (35%) dataset for model training and prediction analysis purposes. The Key Performance Indicator (KPI) for phase 3 will be RMSE (Root Mean Squared Error) or MAE (Mean Absolute Error). Root mean squared error will represent the average difference between predicted and actual values. Smaller RMSE and MAE will show that the stock prices predicted are closer to actual stock prices. Finally, the model with smaller RMSE will provide better prediction accuracy.

The comparison will be made using LSTM Model constructed (phase 2) without using sentiment score and the LSTM model constructed (phase 3) using sentiment score. The RMSE score for both phase 2 and phase 3 will be compared. This will further prove the correlation of sentiment analysis on stock prices.

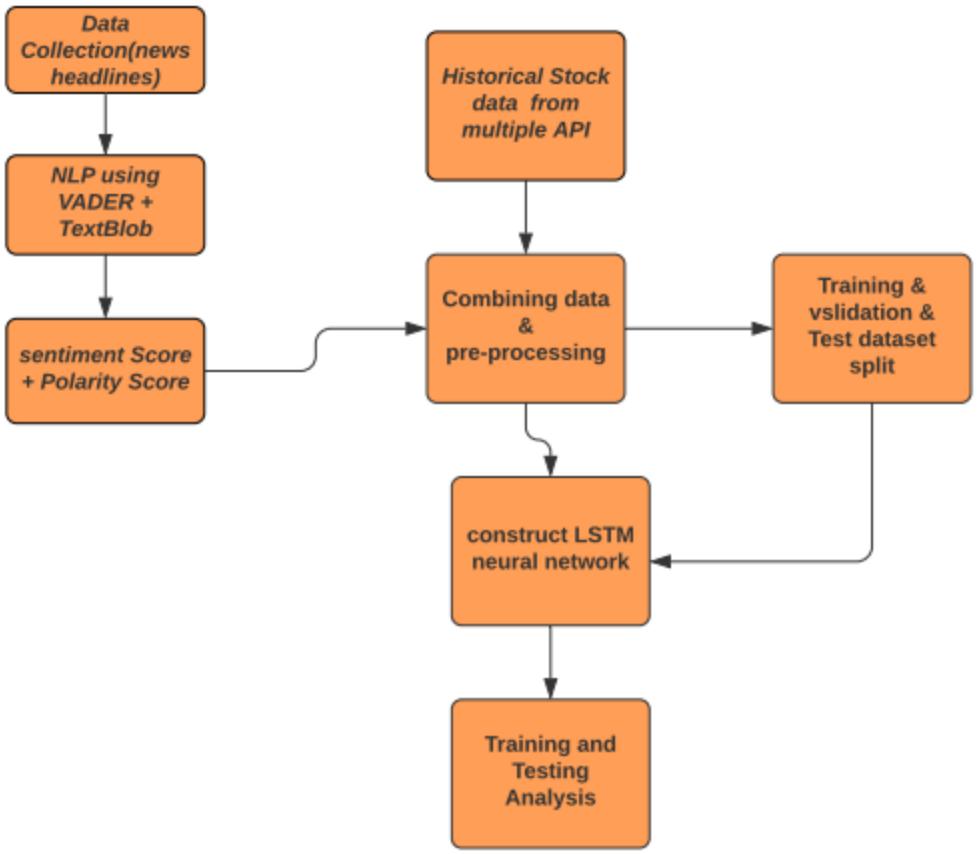


Fig [3] LSTM Model with Sentiment analysis

**Phase 4: Data Visualization**

The dataset prepared during phase 1, phase 2, and phase 3 will be visualized using the python matplotlib library. The LSTM model evaluation (RMSE) during phase 2 and phase 3 for training and test data will also be visualized using matplotlib. The visualization will help analyze the difference between phase 2 and phase 3 LSTM models. Finally, a dashboard will be presented using Microsoft PowerBI to further visualize the stock(ticker), historical price, and predicted price of 60 days. Based on the price momentum shown in the PowerBI dashboard an informed decision can be made to buy or sell a particular stock.

**DELIVERABLES:**

1. A report explaining in detail the methods used and results achieved from sentiment analysis of stock market data.
2. An interactive dashboard (Microsoft Power BI) for data visualization of the results obtained from NLP and LSTM.

3. A compiled and executable Python code used for performing sentiment analysis and Long Short-Term Memory Model.
4. An Instruction Manual and readme file explaining the step-by-step process and working of Python code and Deep Learning model.
5. A PowerPoint presentation explaining, in brief, the importance and relevance of the project.

**TIMELINE:**

| No. | Task | Start Date | End Date |
|---|---|---|---|
| 1 | Pre-Proposal | 07/21/2021 | 07/30/2021 |
| 2. | Literature review and proposal | 08/02/2021 | 08/16/2021 |
| 3. | Data preparation, and pre-processing for sentiment analysis | 08/16/2021 | 08/23/2021 |
| 4. | Preform Sentiment analysis on news data | 08/24/2021 | 09/01/2021 |
| 5. | Data collection and processing for LSTM architecture | 09/02/2021 | 09/09/2021 |
| 6. | Dataset split and parameters setting | 09/10/2021 | 09/13/2021 |
| 7. | LSTM model construction. | 09/13/2021 | 09/27/2021 |
| 8. | Model prediction and analysis | 09/29/2021 | 10/07/2021 |
| 9. | Build dashboard visualizing results from NLP and LSTM Model | 10/08/2021 | 10/21/2021 |
| 10. | Initial draft of project report | 10/22/2021 | 10/29/2021 |
| 11. | Final draft of project report and presentation | 10/30/2021 | 11/17/2021 |
| 12. | Submission | | 12/01/2021 |

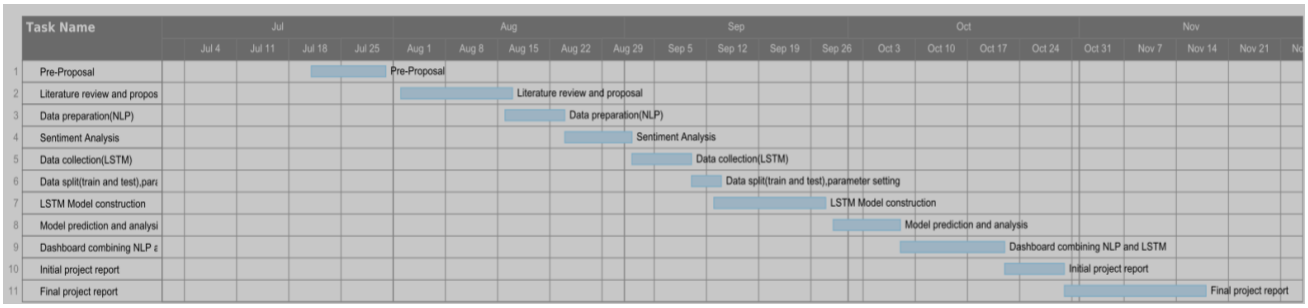

Fig [4] Gant chart showing project timeline

## CONCLUSION:

In this project, predictive analysis is performed on stock market data using both sentiment and time series analysis. The project uses multiple news sources to perform sentiment analysis using the VADER library in addition to other lexicon approaches. The benefit of using such a hybrid approach is that it can work easily on live stream data and is suitable for the versatile analysis of data. The final prediction will show a strong correlation between the news articles, social media posts, and stock prices. The LSTM model build will further verify the significance of public sentiment for stock market prediction. The proposed method can predict almost 80 % of market variance using simple sentiment analysis and LSTM. In the future, the proposed method and model build can be further extended to the forecast of cryptocurrency since research shows that there is a huge change in the momentum of cryptocurrency with news sentiments.

## REFERENCES:

[1] Hegde, M. S., Krishna, G., & Srinath, R. (2018). An Ensemble Stock Predictor and Recommender System. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1981–1985. https://doi.org/10.1109/ICACCI.2018.8554424

[2] Shah, D., Isah, H., & Zulkernine, F. (2018). Predicting the Effects of News Sentiments on the Stock Market. *2018 IEEE International Conference on Big Data (Big Data)*, 4705–4708. https://doi.org/10.1109/BigData.2018.8621884

[3] Pasupulety, U., Abdullah Anees, A., Anmol, S., & Mohan, B. R. (2019). Predicting Stock Prices using Ensemble Learning and Sentiment Analysis. *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 215–222. https://doi.org/10.1109/AIKE.2019.00045

[4] Alostad, H., & Davulcu, H. (2015). Directional Prediction of Stock Prices Using Breaking News on Twitter. *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 523–530. https://doi.org/10.1109/WI-IAT.2015.82

[5] Chen, K., Zhou, Y., & Dai, F. (2015). A LSTM-based method for stock returns prediction: A case study of China stock market. *2015 IEEE International Conference on Big Data (Big Data)*, 2823–2824. https://doi.org/10.1109/BigData.2015.7364089