National College of Ireland

Optimal Timing for Stock Trading in USA: Data Driven Signals for Buy and Sell

MSc Research Project
Data Analytics

Girija Madireddy
Student ID: x21235929

School of Computing
National College of Ireland

Supervisor: Dr.Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Girija Madireddy
Student ID:	x21235929
Programme:	Data Analytics
Year:	2023-24
Module:	MSc Research Project
Supervisor:	Dr.Athanasios Staikopoulos
Submission Due Date:	31/01/2024
Project Title:	Optimal Timing for Stock Trading in USA: Data Driven Signals for Buy and Sell
Word Count:	8835
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Girija Madireddy
Date:	26th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimal Timing for Stock Trading in USA: Data Driven Signals for Buy and Sell

Girija Madireddy
x21235929

Abstract

Stock market forecasting is an interesting topic since there are a lot of variables that can affect prices in the future, along with unanticipated noise. Still, the ability to analyze stock market trends could be vivacious to researchers and investors, and thus has been of continued attention. It is tremendously imperious to distinguish the equilibrium between financial perceptions and econometrics, learning tactics, computational astuteness, technical indicators, and their assimilation in order to analyze and study the arena of interest. In this research, graph theory is introduced as an approach. This method uses Spatio-temporal relationship data among diverse stocks by modelling the stock market as a composite network. Long Short-Term memory networks is combined with this graph-based approach to form a hybrid model. Graph Convolutional Neural networks (GCNs) are recognized for their capacity to exploit spatio-temporal relationships among diverse stocks, while LSTMs excel in handling sequential time series data for prediction modelling. An experiment had been conducted to add sentiment analysis of daily news headlines along with stock price data as input to the model. These deep learning models are equated in contrary to a conventional statistical time series model to assess both computational efficiency and prediction accuracy. This research work concluded that GCN-LSTM model without sentiment analysis is performing better in predicting individual companies stock movement, while the method of aggregation needs to be updated to weighted voting instead of majority voting when calculating overall DJIA movement. Deep learning models are performing far better when compared with statistical ARIMA model since they leverage structural and temporal information from data.

1 Introduction

Stock market is a perilous constituent of global economy. Stock market provides a stage for companies to increase capital by issuing shares to public. This wealth can be used for expansion, innovation and operations which can lead to financial growth and job formation. Investors can participate in achieving financial goals of the company and gain part of profits in the form of dividends. Stock market is often considered as leading indicator of economic health. Throughout the stock market's history, traders and investors have tried to accurately forecast stock prices with the goal of making money by buying cheap and selling high. The best time to purchase or sell assets for the greatest profit has been the subject of much research and has long been a central issue in the study of economics Dwarakanath et al. (2022).

The stock market is prejudiced by numerous factors including political actions, financial indicators, natural adversities and communal trends. This complexity and volatility requires continuous study to understand and predict market movements accurately. The Efficient Market Hypothesis advocates that prices completely replicate all accessible information at any given time. However, anomalies and inefficiencies do occur, and studying these can provide opportunities for investors. Accurate stock market forecasting is essential to preserving financial stability since it allows one to recognize primary threatening signs of possible economic crises or market uncertainty. The solid form of market productivity and the high grade of noise in economic time series prediction make these challenges extremely difficult Fama (1970). The current value of an investment, whether made now or in the future, and the selling price in the future are two prices that every investor must take into account. Nevertheless, investors continue to examine historical pricing patterns in order to inform their choices about current and future investments. Some investors steer clear of rapidly growing stocks or indexes because they expect a correction, while others stay away from losing equities because they think the downturn will continue.

An imaginary investment portfolio with securities that emphases on specific sector or combination of various sectors in the fiscal market is an Index. Total market performance can be gauged by looking at index performance. Index performance is used by portfolio managers as a standard against which to evaluate the success of their investment portfolios. To predict the direction of the market and make trading decisions, stock market traders thoroughly track the index movement. The stock market of United States of America is enormous and has the utmost effect on monetary markets internationally. Several international investors have substantial acquaintance to United States of America's equity market. Hence, any stock market instabilities in United States of America will strongly influence great number of stockholders globally. One of the utmost prevalent indices in U.S. stock market that will be tracked by numerous investors to measure the market performance is Dow Jones Industrial Average (DJIA). DJIA is a combination of major companies from different sectors such as Technology,Financial Service, Healthcare, Consumer staples, Industries, Energy and Materials. Companies are chosen for their impact on country's economy so that the index will reflect overall economy's performance.

Utilizing suitable techniques for forecasting stock market trends and accurately decoding the outcomes can enhance comprehension of market fluctuations and pinpoint the most advantageous moments for purchasing or selling shares. Traders often enter or exit the market at inopportune times, missing out on the full potential of their profit-making opportunities. The objective of this study is to assist traders, portfolio managers, and individual investors in increasing their earnings by precisely forecasting market trends and aiding in the determination of the best trading times. Despite extensive research in this field, there remains room to refine the precision of these predictions by using new technical evolution.

RQ : To what extent predictive analytics can analyse and forecast stock index direction to gauge overall market performance by combining news sentiment analysis, machine learning and deep learning techniques, in order to support investors to trade at optimal timing and gain more profits ?

The goal of this research work is to forecast DJIA index movement to give buy and sell signals to investors and traders by using Graph theory, deep learning and sentiment

analysis. This research work starts with collecting stock market historical time series data for DJIA index as well as 29 companies which are part of DJIA index. Daily news headlines from Reddit news is captured to gauge public sentiment and incorporate this sentiment analysis to aid model predictions. Graph theory is used to build structural relationships between companies and stock price data is used to capture temporal information. Models are used to predict overall DJIA movement to give signals of market raise or fall to traders and investors. This research objective is to analyse the usage of graph theory in stock market prediction and the importance of incorporating public sentiment when predicting market movement.

Earlier studies have not adequately captured the dynamic and complex relationships between different stocks. This research work introduced a graph based approach using Graph Convolutional Neural Networks to capture spatio-temporal relationships among stocks which in turn helps to capture intricate inter-dependencies with in the market. Traditional statistical methods like ARIMA cannot capture the sequential and non-linear nature of stock market data. By integrating Long Short-term Memory networks, this thesis leveraged their ability to handle sequential time-series data effectively. This will allow better modelling of temporal patterns in stock market data. The impact of external and Geo-political issues is huge in stock market movement. Not considering these factors made other works potentially missing out on key predictive signals. Proposed thesis incorporated sentiment analysis of news sources to provide holistic view of the factors influencing stock movements.

The remaining research paper is ordered as follows. A thorough analysis of preceding researches on this topic is detailed in section2. This section describes and discusses about major achievements and limitations of several individual research works. Features and methods taken into consideration to achieve the objectives of this research. Data sets and methodology used to solve the research question and design specifications of individual models, evaluation metrics are described in Section3. Section 4 discusses about ethical implications of the project. Detailed descriptions about hyper parameters used and implementation strategies are given in Section5, followed by critical review of model evaluations in section 6.

2 Related Work

The stock market characterizes a persuasive field for many organizers and stockholders to produce treasured predictions. Elementary knowledge of the stock market, joint with technical pointers, can be used to discover numerous features of the financial market. The influence of diverse events, financial news, and sentiments on investor selections and subsequently market trends have been noted. This kind of information can be exploited to articulate faithful prophecies and accomplish greater revenues. Computational intelligence has advanced to challenge the convolutions of the stock market, integrating diverse deep neural network (DNN) methodologies. This section delves into several research works related to stock prediction.

2.1 Stock Prediction – Statistical Analysis, Machine Learning and Deep Learning

Predicting the future movement of stock index prices has consistently been a fascinating area of exploration for both investors aiming to generate profits through stock trading and researchers striving to uncover hidden insights from the complex time series data of the stock market. Two class labels can be derived for stock market prediction problem, one for the decreasing movement and other for increasing movement and hence can be addressed as a binary classification. Historically, very vast range of classifiers has been established for this application. As the result of separate classifier differs for a varied dataset with respect to dissimilar performance measures, it is unreasonable to recognize a definite classifier to be the finest one. Hence, designing a competent classifier ensemble in its place is yielding growing attention from many researchers Dash et al. (2019).

Artificial neural networks are described by Chhajer et al. (2022) as network of numerical equations with input, output and hidden layers. Authors emphasizes how ANNs can efficiently handles data with high volatility, non-constant variance which makes them ideal for financial time series predictions. Chhajer et al. (2022) stated that Support vector machines as an influential machine learning tool for classification problems, especially effective in high-dimensional spaces. LSTMs are a type of recurrent neural networks which have the capability to store past information, which is crucial in predicting stock market movement. Even though each model has their advantages, they have certain limitations as well. LSTMs require extensive training and refinement which can be resource intensive. Quality and quantity of input data plays a critical role in success of any model and it is challenging to get significantly large datasets to overcome issues like over fitting. It is very challenging to interpret the results of Deep learning networks as they act like black boxes.

A hybrid method to combine two technical analysis strategies namely Triple Exponential Moving Average and Moving Average Convergence/Divergence with machine learning techniques is proposed by Ayala et al. (2021). Authors mainly concentrated on avoiding false trading signals and produce more robust signals based on asymmetric return distribution. Dow Jones Industrial Average and ibex 35 indices in stock market were used to test these methods and results showed that proposed hybrid approach yielded better overall competitive performance in producing trading signals. Without proper background in financial sector, it is extremely difficult to implement this model as it involves combination of technical analysis and machine learning. The model requires continuous optimization updates to remain effective which makes it to limit its practicality for some users. Even though it is tested on three major indices there are concerns about generalization as the best parameters for each algorithm vary depending on the analysed index.

A thorough comparative analysis on LSTM and DNN models in forecasting daily and weekly movements of a stock market index was done by Shah et al. (2018). Authors provided detailed description of network architecture, data processing, normalization strategies and configurations used in training and testing these models. This greatly helps further research in same domain. Their findings shows that LSTM model performed well for weekly predictions and it highlights the potential of LSTM in capturing long term dependencies in stock data. The model was tested on different stocks to prove the generalizability. Only price data was used for predictions which may not capture the

full spectrum of factors influencing stock market. There is an opportunity to improve model's accuracy by adding other variables like trading volume, market sentiment, or economic indicators.

2.2 Stock Prediction – Sentiment Analysis

Corporate buyouts and new product releases are highly impact-full on financial markets. Modelling relationships between these events from news articles and upcoming price movements has become a recent focus area. Matsubara et al. (2018) proposed a generative model of news articles with added condition as price movement to avoid over fitting. Harnessing the power of news sentiment processed from textual news to inform trading decisions is used by Feuerriegel and Prendinger (2016). They employed sentiment analysis methods to extract subjective tones from financial disclosures by utilizing data from regulated adhoc announcements. Statistical evidence that news trading be a gainful scenario was provided by authors, thus confirming the efficiency of established decision supportive system. The quality of news information and the risk of market noise are major practical challenges. The strategies focus on impact of new information instead of detecting unidentified variables or patterns, hence raise alarms about generalizability of the results.

Traditional prediction methods which rely completely on historical stock price data are insufficient due to highly volatile nature of stock market impacted by external factors like news and social media. Khan et al. (2020) investigated the impact of social media and news on stock market prediction. They employed a different approach for feature selection to reduce spam tweets in their dataset. Authors have done a comparative analysis of various machine learning algorithms to find a reliable classifier and achieved significant prediction accuracy of 80.53 percent using social media news and 75.16 percent using news data. Since their study heavily relies on social media and news data, the accuracy and reliability of predictions are contingent on the quality and representativeness of this data. If the data is biased and lacks in comprehensive coverage, predictions can become unreliable.

Converting news articles into distributed depictions by using Paragraph Vector and LSTM to model the temporal properties of past events on stock values to capture both content and time-sensitive nature of stock information was studied by Akita et al. (2016). Combining text and numerical data which often differ in their dimensionality was handled effectively by scaling the size of vectors in neural networks. Market simulation experiments were used to validate the approach and they proved that, this method outperformed numerical data only methods and bag-of-words based models. Researchers proposed that Moving averages and Moving averages convergence divergence can be added as extra features in future studies to increase model's profit-making capabilities. Their study mainly focused on correlations with in the same sector which may not generalize to different sectors in financial markets.

2.3 Stock Prediction – Graph Theory

Understanding the behaviour of stock markets by utilizing its underlying structure has been proposed by Rechenthin (2014). The concept of constructing graphs, considering individual stocks as nodes and their relationships as edges to analyse market dynamics

by examining inter-connectedness of various stocks and broader market structure was explored in their research. The graph-based structure is crucial in understanding the movements in one stock or a group of stocks where stocks can have ripple effects throughout the market, influencing other stocks and sectors. Authors asserted that this concept can be used by administrations to predict and prevent a financial catastrophe happened in 2008. They described that topology of constructed graphs represents the market. They have discussed the importance of dimensionality reduction in handling huge amount of financial data. Rechenthin (2014) deployed an ensemble model to cover similar stocks with in same sector and is optimized to select best classifiers for predicting future market direction.

A fresh approach to forecast stock prices consuming graph theory by utilizing spatio-temporal relationships between various stocks by modelling the stock market as a composite network is proposed by Patil et al. (2020). They have modelled stock price prediction problem as graph problem where stocks are represented as nodes and relationships between nodes are defined based on both correlation and causation. Researchers constructed a Graph Convolutional neural networks based on graphs through spatio-temporal relationships between companies, an aspect often ignored in many traditional time series forecasting models like ARIMA and LSTM. Their model can be generalized to any time series model where graph representations of data are available. Addressing exploding gradient problem with nodes of higher degrees in graph, which is a limitation of GCN is proposed as future direction by authors.

A combination of Graph Convolutional Network and Gated Recurrent Units to model stock market dynamics as a complex network of relationships is experimented by Ye et al. (2021). Authors want to account cross-effects among stocks instead of focusing on individual stocks. Multiple graphs were constructed using shareholding influence, industry relationships and topical news impact. These graphs were used to encode complex inter-stock relationships into a network structure. The proposed multi-GCGRU architecture allows comprehensive analysis of spatial correlations across stocks and temporal dependencies of stock prices. Combining multiple graphs and deep learning process could make interpreting the results and understanding model's behaviour as a challenging task.

2.4 Comparison of Recent Researches

This section highlights major research works on predicting stock market data. Figure1 shows a table of different research works along with their models, evaluation methods, best performed model and data considered.

Akita et al. (2016) implemented Long short term memory networks and other classifiers on NIkkei 225 index and evaluated using market simulation. LSTM gave best results among other models according to their research. Li et al. (2020) performed SVM and LSTM on Hang Seng Index and concluded that LSTM performance is better than SVM. Patil et al. (2020) proposed a novel approach using Graph Convolutional Neural networks on 30 companies data and proved that GCN is performing better than any other linear models and statistical models. Ye et al. (2021) implemented Multi GCN-GRU, a hybrid model and achieved high level of accuracy in predicting CSI500 index movement.

An inspiration had been taken from models described in related work for this research and a hybrid model comprising of Graph Convolutional Neural Netowrks and Long Short Term Memory Networks is implemented to predict index movement. An experiment

Author/Year	Models	Evaluation Method/Metrics	Best Performed Model	Stocks/Index	Data Period
Akita et al. (2016)	Long Short-Term Memory (LSTM) and Paragraph Vector Multi Layer Perceptron Support Vector regression with Radial Basis Function	Market simulation	LSTM	Nikkei 225	2001 - 2008
Li et al. (2020)	Support Vector Machines(SVM) Multiple Kernel Learning Long Short-Term Memory (LSTM)	Prediction Accuracy F1-Score	LSTM Loughran–McDonald Financial Dictionary	Hang Seng Index	2003 - 2008
Zhang et al. (2022)	Long Short-Term Memory (LSTM) with Transformer Encoder-based Attention Network (TEANet)	Accuracy Matthews correlation coefficient (MCC)	LSTM with Transformer Encoder-based Attention Network (TEANet)	88 Stocks with Highest capital market	2014 - 2019
Patil et al. (2020)	Graph Convlutional Neural Networks(GCN) Graph based Linear Models ARIMA	Root mean squared error (RMSE) Mean absolute percentage error (MAPE)	Graph Convlutional Neural Networks(GCN)	30 Stocks from various sectors	Minute Level data for 44 days
Ye et al. (2021)	Multi GCN-GRU(Graph Convolutional Neural Networks-Gated Recurrent Units)	Accuracy Matthews correlation coefficient (MCC)	Multi GCN-GRU	CSI300 and CSI500 Indices	2015 - 2019

Figure 1: Comparison of research works

had been conducted by adding news sentiment analysis as an additional feature to these models.

3 Research Methodology and Design Specification

Each and every investment in the stock market is aimed at maximising yield and minimising allied risk as it is a critical axle in every rising and booming economy. Because of this, various researches have been steered on market analysis using fundamental analysis through different computing algorithms and techniques Nti et al. (2020). In this research, modified knowledge discovery in database (KDD) methodology is used to predict optimal time for buying and selling stocks based on index movement forecast. Extracting meaningful and actionable knowledge along with finding patterns in data is major benefit of KDD methodology. KDD process can be applied to structured data, unstructured data and complex data types, which makes it adoptable for any domain. This section describes about research methodology and design specification of the project. Tools used to perform these tasks are detailed in Configuration Manual document.

3.1 Methodology Framework

The stock market time series data is enormously biased by various aspects such as deflation, guidelines of banks, inflation and fluctuations in economic policies of a country. Evolving a competent forecasting model to predict the future oscillations on the stock index price direction is continuously seemed as an additional stimulating job because price of a stock can increase or reduce overnight Dash et al. (2019). Figure 2 shows the knowledge discovery framework followed in this research. Various steps in methodology framework includes collecting stock related data and news headlines data, pre-processing of data, exploratory data analysis and feature engineering, modelling through applying deep learning techniques and sentiment analysis, Evaluating results, interpreting the decision making system and using these meaningful insights to take informed decisions. Following sections include detailed explanation about each stage in framework.

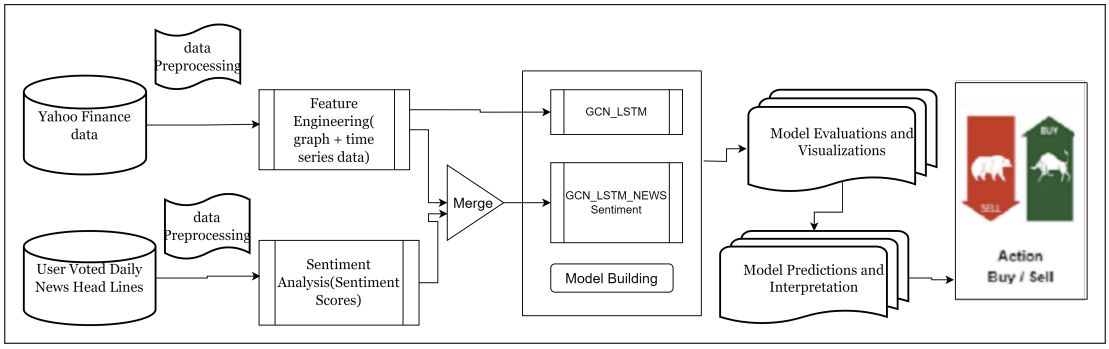

Figure 2: Research Methodlogy Framework

3.2 Data Collection

Collecting quality data is primary stepping stone for any Data Mining and Machine learning research work. In this research, primary data is considered as daily stock price of Dow Jones Industrial Average index (DJIA) from USA Stock market and individual stock price information of 29 companies that are included in DJIA index. Five years (2012 – 2016) of stock data for DJIA is downloaded using Python libraries from Yahoo finance. Yahoo Finance offers accurate historical data which is crucial in any time series analysis. Data is directly pulled from stock exchange feeds, hence accuracy and consistency is assured. Date, Open,High,Low, Close, Adj Close-Adjusted Closing price, Volume and ticker are the features that are extracted. Date indicates the day on which the stock was traded, Open indicates the opening stock price at market opening, High indicates the highest price that stock was traded on given day, Low indicates the lowest price that stock was traded on given day, Close indicates the price of stock at market closing, Adj Close is adjusted close price after considering splits and dividend distributions and Volume indicates the number of shares traded that day. Ticker is the individual company name when all companies stock data is combined to use in modelling.

Historical news headlines data is secondary dataset in this research work, which is collected from Kaggle public data repository Sun (2016). The data was originally crawled from Reddit world news channel. News headlines were ranked by Reddit users and top 25 highest voted news headlines were picked to capture public sentiment. The news data is considered for same period from 2012 to 2016 as historical stock price data. Sentiment analysis using News headlines is proved to be more influential than sentiment analysis on news content when it comes to stock market prediction Ding et al. (2014). The collected news headlines data is used in sentiment analysis to capture the positive or negative trend of public opinion and use it as a feature in modelling.

3.3 Data Pre-processing and Feature Engineering

Knowledge Discovery in Database process ensures thorough examination of stock price data which is structured and daily news headlines data which is unstructured data. Cleaning is an important step in data pre-processing stage. Missing values and NA values are checked and removed in historical stock price data. Multiple companies stock market data is used for Graph Convolutional Neural networks. A graph is constructed using correlation matrix built on Pearson correlation coefficient . Correlation is calculated between close prices of 29 companies where each company is represented as node. Threshold of

0.5 is defined to create an edge between two nodes. Features, labels and graph were converted into tensors to align with model design.

Neural networks including Long Short Term Memory tend to converge faster and perform better when input data is scaled. Scaling assures that all data points in time series are treated uniformly in terms of their relative changes. Historical stock price data is scaled to use it in Long Short Term Memory networks. LSTMs are designed for time series forecasting and they require data in sequence format where input is a series of past data points and output is a future data point. Stock price data is split into input-output pairs to train LSTM networks. Close price is considered as significant feature from time series data as model needs to predict next day's close price to understand market movement. Close price is normalized and scaled to values between 0 to 1 and created as features list. Labels are calculated using daily price differences with previous day's closing price and created as labels list. Both features and labels lists are converted into numpy arrays. Input data dimensions are checked to ensure that they are aligned with model input specifications.

Statistical distribution, median and standard deviation are calculated on time series data to understand data distribution. Date is transformed to 'datetime' format and is set as an index for time series analysis. Dickey-Fuller test is performed to understand the stationarity of time series data. Statistical predicting models assume time series data to be stationary, hence stationarity is tested and assured using data transformation. Auto-correlation and Partial auto-correlation plots were built to determine statistical parameters for modelling. News headlines data is verified for missing values, duplicate values and removed. Headlines for each day are combined and all special characters,extra spaces are removed to form only words to process for sentiment analysis. Subjectivity and Polarity are calculated using Text blob and positive score, negative score and compound scores are calculated using sentiment analyser.

3.4 Data Modelling – Design Specification:

There are behaviours to predict future proceedings and gain the prizes securely even though it is indefinite and inexact. The application of artificial intelligence and machine learning for stock market prediction is one such prospect. It is advisable to use artificial intelligence to do calculated predictions before investing as stock market is very volatile Chhajer et al. (2022). Five models are studied on stock prediction in this research work. Statistical ARIMA model, Long Short Term Memory networks , Long Short Term memory networks along with news sentiment analysis, Combination of Graph Convolutional neural networks and long short term memory networks and a hybrid model combining Graph Convolutional Neural networks, Long Short Term Memory along with News sentiment analysis. This section describes about design specification of each model.

3.4.1 Graph based Models

Patil et al. (2020) presented a novel strategy that uses graph theory to leverage the spatio-temporal interactions across different company stocks through showing the stock market as an intricate web. Their investigations showed that Graph-based techniques surpass conventional methods by integrating structural data for building the forecasting models. This research added an additional feature of news sentiment analysis to Graph based neural networks. For each stock, traditional time series models such the Long Short-Term

Memory (LSTM) employing recurrent neural networks and the Auto regressive integrated moving average (ARIMA) must be built. The linkage or unspoken connections between businesses in related industries will not be taken advantage of by these models.

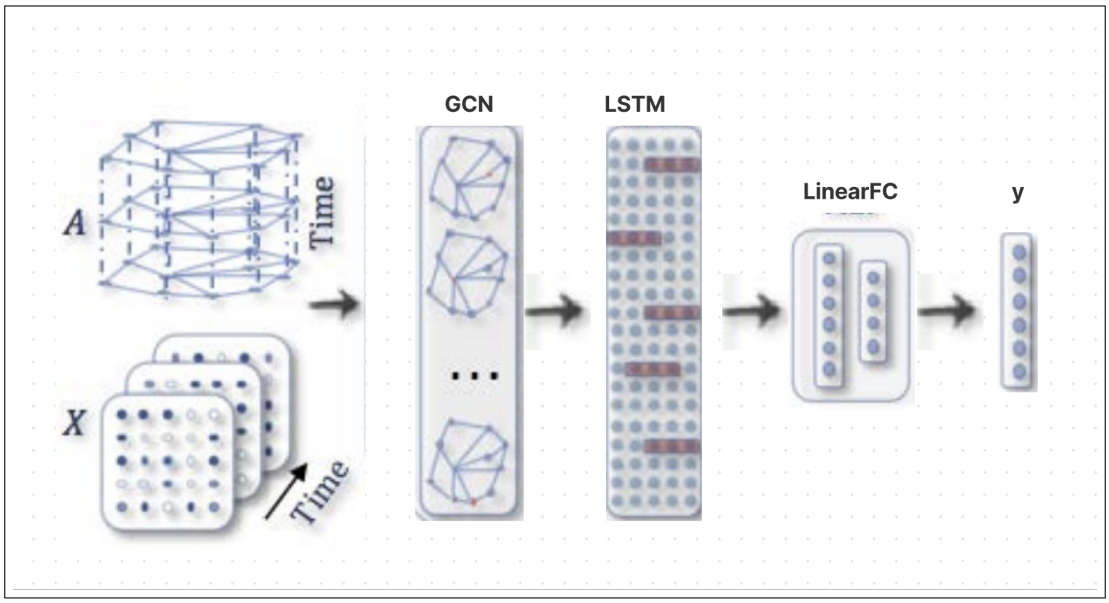

Figure 3: GCN-LSTM Model Architecture

Figure 3 illustrates the detailed design of GCN-LSTM model. This is the architecture of Graph convolutional neural networks. Feature matrix X is passed to the model along with adjacency matrix A. In this research feature matrix contain vector of historical stock time series data corresponding to each node. These inputs are passed to Graph Convolutional layer to learn spatial relationships among different data points. Then GCN output is reshaped and passed to LSTM to process time series data and learn temporal dependencies. LSTM output is passed to fully connected linear layer which maps it to number of output Y classes, in this scenario stock classification for each node. This architecture is powerful where spatial information among data points and temporal information of data are important for predictions. Relationship among companies are spatial information and time series sequence is temporal information. Adjacency matrix of the graph and time series for every node in the graph are two inputs to this model. Adjacency matrix is calculated using Pearson Correlation Coefficient. Time series data is normalized and scaled to have even range for all data points.

The collective confidence or trepidation of investors regarding the market or specific stocks is often reflected by public sentiment. Positive sentiment can lead to bullish behaviour and increased buying where as negative sentiment can lead to pressure of selling and falling prices. Investment decision can not always be taken based on financial data and rational analysis. Emotional reactions to the events occurred on daily basis can play a significant role in market movement. In this research, an attempt has been made to understand public opinion and incorporating this as additional feature. Model design is same as described above but sentiment scores calculated on daily news headlines had been added as additional feature vector in X. This feature will be same across all nodes from the graph and add advantage in predicting each company stock movement.

3.4.2 Deep Learning Models

A variant of deep recurrent neural networks called Long Short Term Memory (LSTM) is used to predict DJIA index prices. Model is developed using Keras framework in Python. Figure 4 illustrates general architecture of Long Short term Memory networks. In this research, model is created as linear stack of layers in Keras which is sequential. LSTM layer is first and core layer of the model. The number of neurons or units were defined as 125 as too higher units can complicate the model and lead to higher training periods and increase the chances of over fitting. The hyperbolic tangent(tanh) activation function is used for LSTM units which is common choice for these neural networks. Next is fully connected layer which serves as transitional layer between LSTM and output layer. We are predicting the stock price of DJIA index as final output in this model.

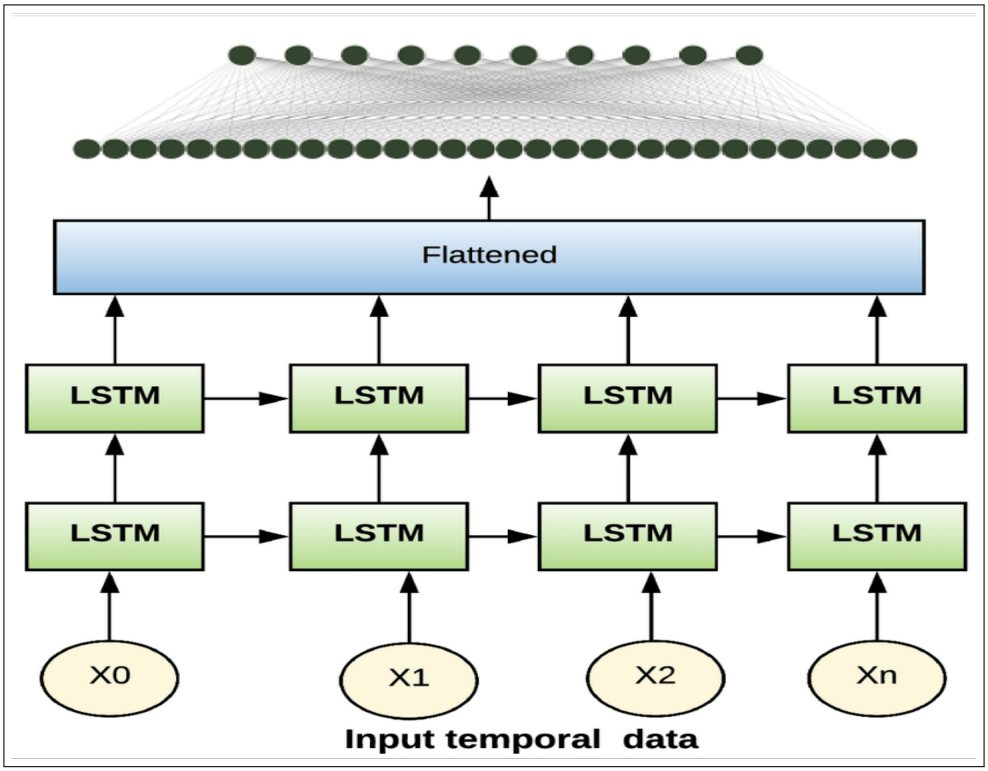

Figure 4: LSTM Model Architecture

Model is compiled using 'adam' optimizer which is considered as efficient stochastic optimization model that adapts learning rate during training. Mean squared error is considered as loss function as it penalizes larger errors and leads to more accurate predictions. Another experiment is performed in this research by adding news sentiment scores to LSTM model. Model architecture is adjusted to perform a classification task, where sentiments calculated from daily news headlines are added as additional features along with sequential time series stock data to predict stock index movement. Time series data will be helpful to learn historical temporal dependency and news sentiment scores will integrate public opinion. These two inputs will help model to consider long term historical data at the same time incorporating real time information to capture immediate reaction of public for daily events and announcements.

3.4.3 Statistical Time Series Model

Auto regressive integrated moving average (ARIMA), a well-established statistical model that can be applied to any kind of time series forecasting, is also studied in this work. ARIMA is an abbreviation that translates to: Auto-regression, or AR. This represents the connection between the time series lag parameters as variables. I: Consolidated. Integration of differencing is done to eliminate the stability over time series. If a time series is considered as stationary if it has root unit. The standard deviation and mean of a stationary time series are 1 and 0 respectively. MA stands for moving average. This uses a lag model for the residual error terms. Parameters derived from the model of moving average. Hence, p, d, and q are the three parameters of ARIMA, where p is number of lagged terms considered, d is number of times the time series is differenced to make it stationary and q is length of moving average window. When constructing the model, a component is not taken into account if its value is 0. This includes p, d, and q. Consequently, based on the parameter value, an ARMA, AR, MA, I, or ARIMA model can be constructed to match the specified time series.

3.5 Evaluation and Interpretation of Results:

Multiple models for each stock were built in this research. Graph based models were built by using 29 companies stock data and news sentiment analysis as classification problem. 29 companies stock classification labels are combined and aggregated using majority vote to get final classification label for DJIA index, since the objective of this study is to predict index movement. DJIA index labels will be generated from DJIA stock data and compared against aggregated labels generated in previous step. Classification models will be evaluated using accuracy , f1-score, precision and recall. Accuracy verifies the proportion of correctly classified instances. Precision checks the correctness of optimistic prediction where as recall is segment of positives that were properly recognized. F1 score is needed to measure the equilibrium among precision and recall in scenarios of imbalanced datasets. Confusion matrix will help us understand false positives and false negatives.

Deep learning LSTM models were built using DJIA index data as time series prediction regression models. Hence these models will be evaluated using root mean squared error. ARIMA model is built based on DJIA index stock time series data as regression model and will be evaluated using root mean squared error, absolute error and absolute percentage error. Mean squared error is extra delicate to outliers than mean absolute error. Root mean square error is useful for interpretation. Cross validation in case of both regression and classification models is performed using time based splitting. Time based split ensures that models were validated against unseen data during training.

4 Ethical Considerations of Research Project

Any endeavour involving data must inevitably and absolutely take ethical principles and data privacy into consideration. Data from daily news channel and publicly available stock market data were used in this study on stock market movement prediction. There is no sensitive information in datasets that could be harmful to a person or an organization. Bias and fairness is mindfully considered while designing algorithms. A great

deal of attention is paid to ensure that the research complies with ethical and regulatory requirements around the sharing of information that may affect stock prices.

5 Implementation

Five models were implemented as part of this research work. This section contains detailed description of each model implementation.

5.1 Graph based Models

Graph convolutional networks (GCN) are different from traditional statistical and deep learning models like ARIMA and Long Short Term Memory. GCNs are popular to capture spatio-temporal relationships between data points, in this research relationships between companies. Stock market data for 29 companies which are part of DJIA index is downloaded from yahoofinance using yfinance library. Five years data is considered from 2012 to 2016 where 2012 to 2014 is considered as training and rest is for validation. Each company has 626 data points in training data and 503 data points in testing data. Labels are generated for each data point and for each node.

Adjacency matrix is first input for GCN layer which is built based on correlation matrix of different companies close prices. Threshold is defined as 0.5. An edge between company nodes will be created if correlation between those companies exceeds the threshold. High positive correlation means the stocks tend to move in same direction. Edges between companies are created as unidirectional as relationship between company i to j is same as j to i. This edges list is converted to PyTorch tensor for it to be compatible for GCN layer. Figure 5 illustrates graphical representations of relationships between different companies in DJIA index. PyTorch geometric data object is created which comprised of input feature time series data along with edges data from graph.

Closing price is used as feature for 29 companies and each company have 626 days data. 64 gcn hidden units are used and 32 lstm hidden units are defined. Number of classes are 2, since this is binary classification. Adam optimizer is used with learning rate of 0.001. Cross entropy loss function is used which is suitable for binary classification. Model is trained over 200 and 500 epochs. The model processes the input data and loss is calculated using Loss function. Loss is back propagated and optimizer updates model parameters.

Initial GCN-LSTM model considered only historical stock time series data. Second model is developed by adding Sentiment analysis of daily news headlines as secondary feature considering public opinion is an impacting element of stock market movement. News headlines data is converted to lowercase to ensure same words in different cases are treated equally. Special characters and punctuation marks are removed as they do not add value while calculating sentiment scores. SentimentAnalyzer from Natural Language ToolKit is used to calculate sentiment scores and later these scores are aggregated on daily basis. The compound score is single metric that represents the overall sentiment of text, hence it is used as additional feature for model. Sentiment scores are also split on dates basis as these scores will be combined with closing prices from time series data. Feature list and labels are generated using sentiment scores and daily closing prices. Labels are generated based on day to day change in closing prices. If closing price increases compared to previous day, the label is set to 1 otherwise 0.

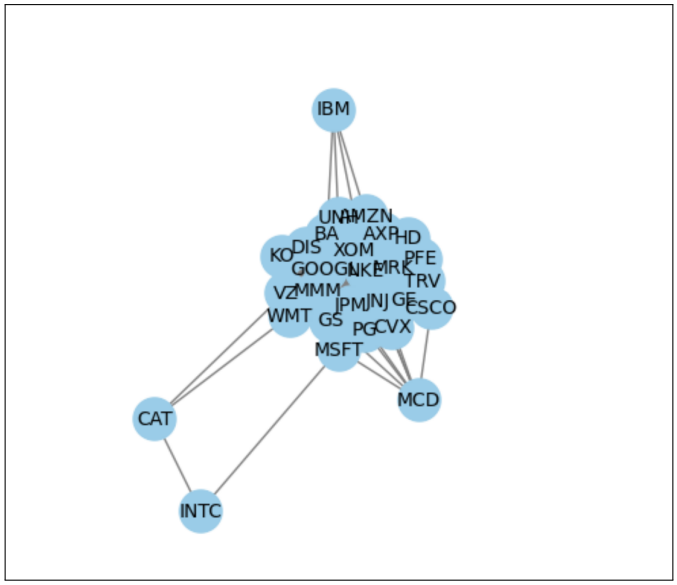

Figure 5: Graph representation of companies relationships from DJIA

Figure 6 illustrates sentiment scores trend over time. Sentiment scores range between 0 to -0.5 which indicates that news sentiment is negative throughout the period. There is a significant amount of volatility in the sentiment scores as indicated by vertical spikes through out the graph. There is no significant upward or downward trend in the sentiment scores and they fluctuate around the same range. Figure 7 illustrates scatter plots between closing prices and sentiment scores for each company. Widely distributed points suggests that there is a significant variation in both sentiment scores and stock prices over time period considered. There is no clear visible positive or negative correlation between closing price and sentiment scores for most of the companies.

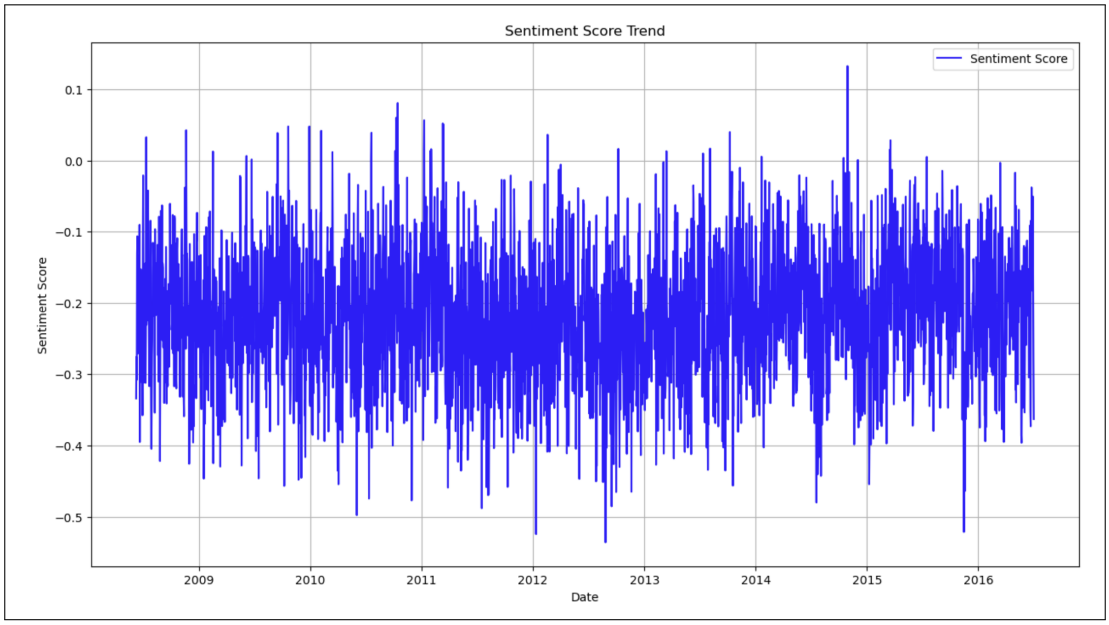

Figure 6: Time series plot of sentiment scores

Adjacency matrix is generated using correlations and converted to tensor to input to the model. Combined features and labels are also converted to tensors to be compatible with the model input. Training data and validation data are prepared separately as splitting tensors can create randomness in time series data. Input X dimensions are [29,626,2] which indicates 626 days of data for 29 companies with 2 features including closing price and sentiment score. Input A dimensions are [2,572] which represents a graph with 572 edges. Output labels are [29,626] representing 626 days data for 29 companies. 64 gcn hidden layers, 32 lstm hidden layers are used. Model is trained using 'adam' optimizer with learning rate of 0.001 and CrossEntropyLoss function. Adam optimizer combines the advantages of other optimizers AdaGrad and RMSProp. Model is set to training mode which enables features like dropout and batch normalization. Model is trained over 200 and 500 epochs. Loss is calculated and then backpropagated through the network to update the model parameters.

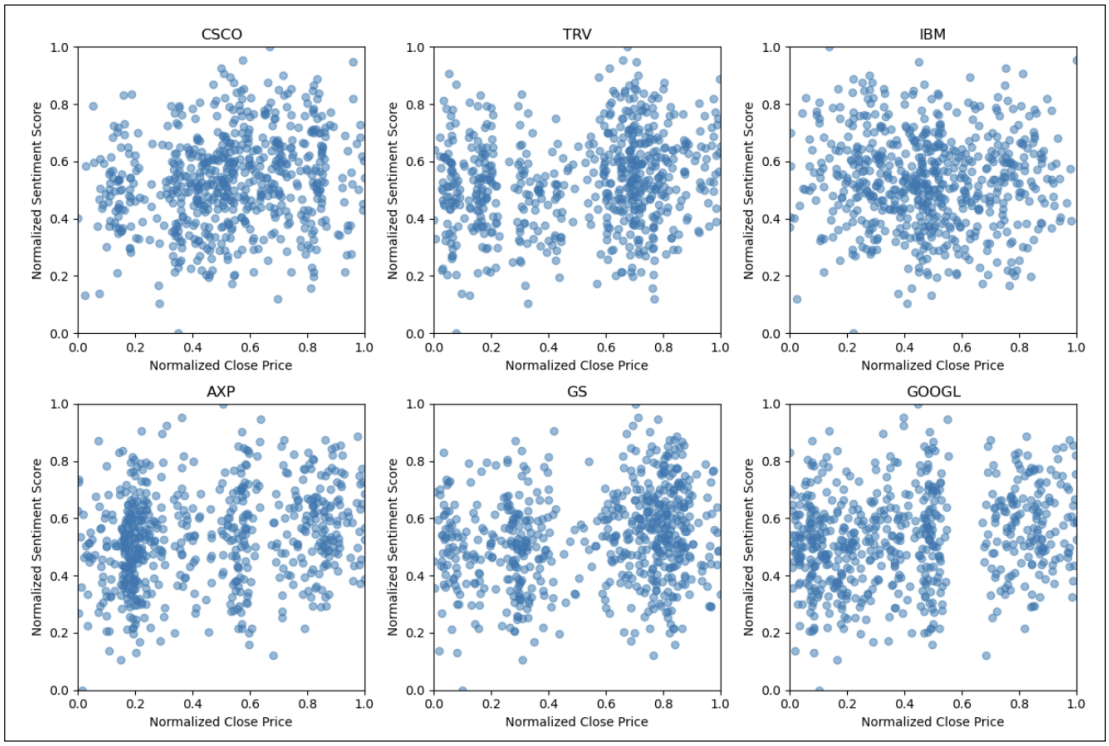

Figure 7: Scatter plots for closing prices and sentiment scores for different companies

5.2 Deep Learning LSTM Models

LSTM networks are a variant of Recurrent Neural Networks that can learn and remember over long sequence of data and are explicitly designed to deal with long term dependency problems, hence well suited for time series analysis. DJIA index stock time series data is downloaded from yahooFinance from 2008 to 2016. Training data is considered from 2008 to 2014 and rest is considered as test data. Time based split is done because sequence of data has significant importance in time series analysis. LSTM algorithms performs better when data is scaled to a data range of 0 to1, hence stock data is scaled using 'MinMaxScaler'. Optimizing algorithms can get better convergence when data is scaled and normalized. Time series sequence data is split into multiple input output pairs for

LSTM to learn from. Window size is given as 60, hence each input pair will contain 60 data points and immediately following data point is considered as corresponding output. Creating subsequence from time series data by window technique is important for LSTM because it expects input in the form of sequence and can learn to predict the next time step.

LSTM model architecture is built using Keras API from Tensorflow. LSTM layer is defined with 125 units indicating dimensionality of output space. Activation function is defined as 'tanh', which is standard for LSTM layer and helps in managing flow of gradients through the network. Input is defined as (60,1) which indicates window size as 60 and number of features as 1 since high price is being considered as a feature for this model. LSTM layer is followed by Dense layer with 25 neurons which will help to interpret the features learned by LSTM and map them into desired output size. Final layer is another Dense layer as this will give us prediction for next time step in sequence. Model is compiled with 'adam' optimizer and Mean Squared Error loss function. MSE loss function is suitable for regression problems which will guide the training process by measuring model's prediction accuracy. Model is trained for 15 epochs.

Second variant of LSTM deep learning model is developed by adding Sentiment analysis on daily news headlines from redditnews as an additional feature in LSTM model to incorporate public opinion in model predictions. Top 25 headlines data is merged as combined text and additional spaces, special characters are removed by using regular expressions. Subjectivity is calculated for combined news data using TextBlob to quantify the amount of personal opinion and factual information contained in the text. Polarity indicates the sentiment orientation in the text. This score is particularly useful to gauge public sentiment. SentimentIntensityAnalyzer is used to compute positive,negative,neutral and compound scores. Data is scaled and split by 80-20 percent as we have label for each day and want to predict the label based on stock price data along with sentiment data. First layer in this model is LSTM with 50 units with 'relu' activation function and it will return sequence to second LSTM layer. Second LSTM layer with 50 units and 'relu' activation function will pass output to Dense layer. Dense layer is defined with 'Sigmoid' activation function which is typical to binary classification problems. Model is compiled with 'adam' optimizer and 'binary crossentropy' loss function. Model is trained for 100 epochs.

5.3 Statistical ARIMA Model

Statistical ARIMA model has been created for this research as a baseline model for predicting DJIA stock price. Time series data is verified for stationarity using Dickey-Fuller test. Based on the test result, time series data is differenced to make it stationary since ARIMA model assumes data to be stationary. Figure 8 shows rolling mean and standard deviation of time series data after differencing. Relatively constant rolling mean and standard deviation indicate that time series data achieved stationarity. Test Parameter selection is most challenging and time taking process of ARIMA modelling. Akaike information criterion (AIC) is used to perform ARIMA with different input parameters (p,d,q) and find out best parameters with grid search. AIC explains how well the model with given parameters fits the time series data. Model with lowest AIC value : p-0 , d-1,q-0 is finalized to train and predict.

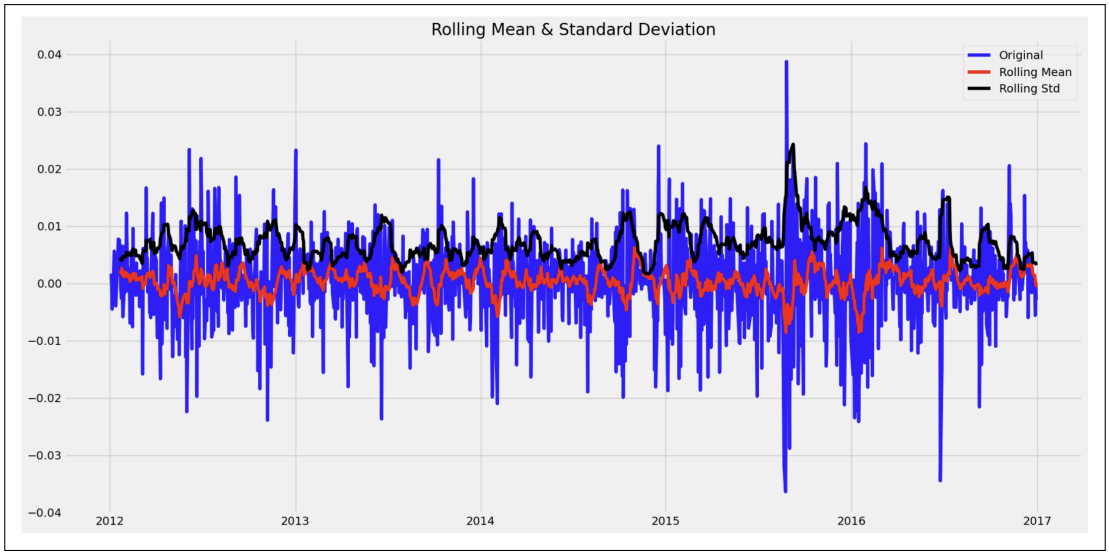

Figure 8: Rolling Mean and Standard Deviation of time series data after differencing

6 Evaluation

Evaluations of five models implemented on index stock movement prediction are discussed in this section.

6.1 Graph based Models

Model performance needs to be valuated on unseen test data. Test data is processed same as train data for GCN-LSTM model. Gradient calculations during validation are prevented to reduce memory consumption and speeds up computation. The validation loss is quantitative measure of model's prediction error. 64 percent validation loss depicts that there is still notable difference between actual and predicted values. However, model training loss is 62 percent which indicates model is generalizing well and not overfitting. 76 percent validation accuracy in complicated and volatile stock market prediction is considered a promising result. When model is trained for 500 epochs, a validation loss of 26 percent and validation accuracy of 94 percent is achieved which is very impressive and promising result. Model predictions are very close to actual values. The level of accuracy is very high especially in the context of stock market prediction which is challenging task.

Classification labels for each company stocks are aggregated to create a label for DJIA index. Majority voting system is applied to determine daily label for DJIA index. If more than half companies are having a label of 1 then DJIA label will be calculated as 1 otherwise it is 0. Accuracy of 51.88 percent indicates that model is performing slightly better than random guess which is better in complex system like stock market but this level of accuracy limits model's reliability in taking trading decisions. Precision of 51.88 percent indicates that model is more inclined to predict positive outcomes of index movement. 96 percent of recall with lower precision often points bias in model's prediction. High rate of false positives indicates a tendency to predict increase in DJIA price even on days it does not happen.

Sentiment scores are added as an additional feature in second model to incorporate public sentiment in model predictions. Model is trained for 200 and 500 epochs. Integ-

rating sentiment analysis lead to a decrease in performance with validation accuracy of 44 percent and validation loss of 88 percent. The training loss was slightly less which indicates that model may be slightly under fitting to validation data and not generalizing well. Increasing training duration with 500 epochs did not improve validation accuracy and showed very high validation loss. Table 1 shows validation accuracy and validation loss for different runs for both models.

Table 1: Evaluation Metrics of Graph Based Models

Model	Epochs	Validation Accuracy	Validation Loss	Training Loss
GCN-LSTM	200	76%	64%	64%
GCN-LSTM	500	94%	26%	26%
GCN-LSTM_NEWS SENTIMENT	200	48%	88%	45%
GCN-LSTM_NEWS SENTIMENT	500	47%	200%	20%

Sentiment scores, particularly when derived from daily news headlines, can introduce noise. If this noise is not related to stock price movements, it can confuse model, leading to decreased accuracy. This noise can cause the model to over-fit to irrelevant features in the training data, reducing its ability to generalize to unseen data. This behaviour is particularly problematic in complex models such as GCN-LSTM where capacity to fit the data is high. News sentiment does not always have a direct and immediate impact on stock prices. The impact might have varying lag times that the model does not account for. Model is trained to associate sentiment with immediate stock price movements without considering the lag period. Different lag periods need to be explored to capture the complex and non-linear relationship of news sentiment and stock prices. Performance of sentiment analysis tool is very crucial. SentimentAnalizer from NLTK is used in this research work. If SentimentAnalizer does not accurately capture the contextual sentiment of news headlines, it could lead to misleading features being fed into the model.

6.2 Deep Learning LSTM Models

Test data is prepared for LSTM model by using scaling and sequence creation. MinMaxScalar is used for scaling and sequence function is used to create sequences as input output pairs. Test data is processed through LSTM model and predictions are saved. Predictions are inverse scaled to transform them into their original scale. Both predictions and original values are inversely scaled so that they can be compared against each other. Figure 9 illustrates plot of actual and predicted values for DJIA index. Predicted prices are closely following the actual prices indicating that model is a good fit and able to capture overall trend in stock price movement. Lagging behaviour of time series in this plot which is common behaviour indicating the model is reactive instead of predictive in those periods. Model is handling volatility very well as seen in sharp rises and falls.

Model's performance is consistent over time as there is no apparent degradation in the quality of predictions through out the time period. Root Mean Squared Error is a standard way to measure the error of a model in predicting quantitative data. RMSE of 201.54 means that, on an average model prediction are 201.54 units away from real values. Since the stock prices are ranging from 16000 to 19000, this amount of RMSE is considered quite good. MAE measures the average magnitude of the error. 164.34 of MAE which is close to RMSE suggests that errors are more uniformly distributed.

Second experimented model is LSTM with news sentiment scores along with stock price data. This model is evaluated as classification problem. Increasing training period

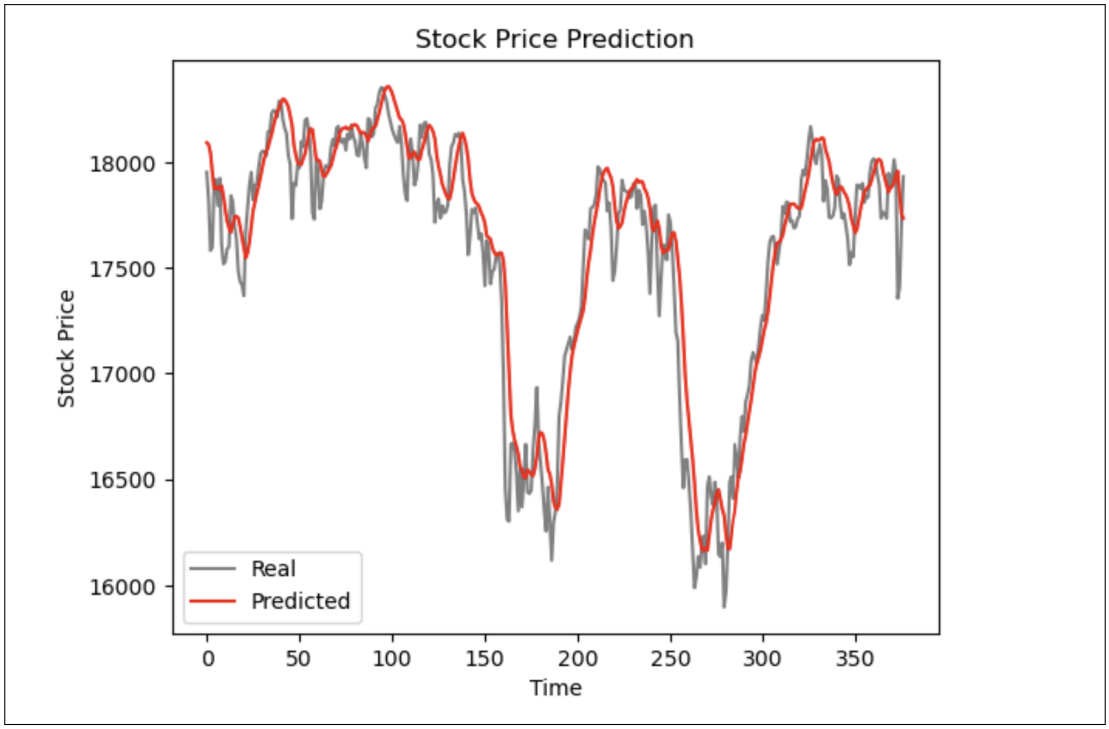

Figure 9: LSTM model predictions against actual values

from 10 to 100 epochs did not show any improvement in model accuracy. Model accuracy with 100 epochs is 57 percent which is slightly higher when compared with GCN-LSTM-sentiment model. Precision of 0.59 indicates that model is predicting positive classes correctly about 59 percent of the times. Recall of 0.84 indicates that around 84 percent of the times model is predicting actual positive classes as positive. Model is more lenient in predicting positive classes at the cost of more false positives.0.69 of f1-score indicates balance between precision and recall with tendency towards recall. There is an opportunity for significant improvement in model's performance.

6.3 Statistical ARIMA Model

The interpretation of RMSE and MAE heavily dependents on scale of data. Since DJIA stock price is between 16000 to 20000 , RMSE of 1066 is considered as reasonable performance. An MAE of 886 indicates that the absolute error of model's prediction is around that number. Figure 10 illustrates actual and predicted values plot for ARIMA Model. A flat predicted values line indicates that model has predicted a constant value throughout the forecast period. The actual data shows fluctuations and raising trend which is not captured by forecast. ARIMA models are linear and may not capture the non-linear patterns present in stock market data. Prices in Stock market is influenced by many factors such as economic indicators, company performance, investor sentiment and geopolitical events. Many factors among these will exhibit non-linear relationships.

ARIMA model assumes that the data is stationary which means, the statistical properties of data do not change over time. In this research work, data was differenced to achieve stationarity. While differencing can help achieve stationarity in mean, it might not account for changes in variance or the possibility of structural breaks in the series.The

huge difference between actual values and predicted values is a common occurrence in simple statistical models. Financial time series data such as stock price, are often said to follow 'Random Walk', indicating that past prices are not a reliable indicator of future prices. This undermines the fundamental premise of ARIMA model. Simple Statistical ARIMA model is unable to capture the trends, fluctuations and non-linear relationships present in complex DJIA time series data, hence the difference in actual and predicted values. In this research work, this model is developed as a baseline model to compare model results against deep learning models.

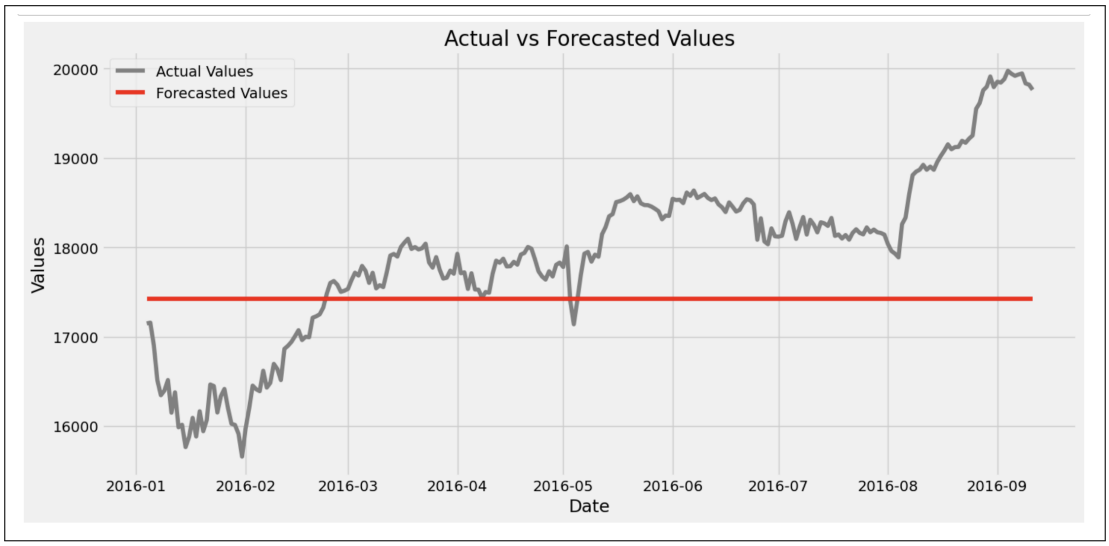

Figure 10: ARIMA Model – Actual and predicted values plot

7 Conclusion and Future Work

In this research, a hybrid model comprised of Graph convolutional neural networks (GCN) and Long short term memory (LSTM) is implemented and an experiment is conducted by adding daily news sentiment analysis to this model. Deep learning LSTM model and statistical ARIMA models were built to compare the performance of GCN models. GCN-LSTM models were performed on stock information of 29 individual companies which are part of DJIA index, using their spatial relationships as input along with stock data. Results were aggregated by using majority vote to calculate overall DJIA movement and compared against actual DJIA values. LSTM and ARIMA models were performed on DJIA stock data so that results can be comparable.

It has been observed that, GCN-LSTM model without sentiment analysis generalized well to new and unseen data proven with closely aligned training loss and validation loss. Model complexity and capacity are well balanced with the amount and nature of data that it was trained on. GCN-LSTM model showed robust performance when training epochs are increased. Sentiment analysis is introducing noise rather than useful insights, detracting the model's prediction capability. Further investigation and refinement are required to leverage sentiment data effectively where baseline GCN-LSTM model demonstrates a strong foundation.

While GCN-LSTM model showed promising results on individual stock prediction, the overall DJIA movement calculation using majority vote did not show reliable results. LSTM model performed well on DJIA data but model's performance got degraded when sentiment analysis is added. Both GCN-LSTM and LSTM models performed well when compared to statistical ARIMA model. Aggregation using majority vote while each company has different weights in DJIA index is causing misrepresentation of model predictions in case of GCN-LSTM model. A weighted aggregation considering the market capitalization or historical influence of each stock on the DJIA could be more representative and fully utilize nearly accurate predictions of GCN-LSTM. Additionally, this aggregation can be incorporated into model design, instead of calculating separately. Other methods of integrated sentiment analysis can be explored rather than direct input. Adding more technical indicators like Moving Average Convergence Divergence (MACD), Cumulative Return (CR) instead of direct Closing price of stock can be explored to add more insights from time series data.

Ensemble methods , Deep Reinforcement Learning (RL) and Graph Attention Networks (GATs) are other alternative approaches to those used in this research work. Combining predictions from multiple models or different configurations of neural networks is great approach to gain positive strengths of different models. Deep RL is well suited to use in developing trading strategies that adapt to new market conditions, as the model learns to maximize financial reward over time. The sequential decision-making process inherent in trading is very well aligned with deep RL. GATs are effective in capturing dynamic and complex interconnections between different stocks or market indices. They will allow to give more weight to most influential stocks which is very critical in accurate aggregation.

Acknowledgement

The author is grateful to Dr.Athanasios Staikopoulos for the supervision and guidance given on this research project implementation and author is greatly thankful to her family for providing moral backing in learning process. With the increasing computational power and advanced algorithms, ensemble methods can be trained on large datasets, providing more accurate and reliable forecasts.

References

Akita, R., Yoshihara, A., Matsubara, T. and Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information, *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, IEEE, pp. 1–6.

Ayala, J., García-Torres, M., Noguera, J. L. V., Gómez-Vela, F. and Divina, F. (2021). Technical analysis strategy optimization using a machine learning approach in stock market indices, *Knowledge-Based Systems* **225**: 107119.

Chhajer, P., Shah, M. and Kshirsagar, A. (2022). The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction, *Decision Analytics Journal* **2**: 100015.

Dash, R., Samal, S., Dash, R. and Rautray, R. (2019). An integrated topsis crow search based classifier ensemble: In application to stock index price movement prediction, *Applied Soft Computing* **85**: 105784.

Ding, X., Zhang, Y., Liu, T. and Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1415–1425.

Dwarakanath, K., Dervovic, D., Tavallali, P., Vyetrenko, S. and Balch, T. (2022). Optimal stopping with gaussian processes, *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 497–505.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work, *The journal of Finance* **25**(2): 383–417.

Feuerriegel, S. and Prendinger, H. (2016). News-based trading strategies, *Decision Support Systems* **90**: 65–74.

Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H. and Alfakeeh, A. S. (2020). Stock market prediction using machine learning classifiers and social media, news, *Journal of Ambient Intelligence and Humanized Computing* pp. 1–24.

Li, X., Wu, P. and Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong, *Information Processing & Management* **57**(5): 102212.

Matsubara, T., Akita, R. and Uehara, K. (2018). Stock price prediction by deep neural generative model of news articles, *IEICE TRANSACTIONS on Information and Systems* **101**(4): 901–908.

Nti, I. K., Adekoya, A. F. and Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions, *Artificial Intelligence Review* **53**(4): 3007–3057.

Patil, P., Wu, C.-S. M., Potika, K. and Orang, M. (2020). Stock market prediction using ensemble of graph theory, machine learning and deep learning models, *Proceedings of the 3rd international conference on software engineering and information management*, pp. 85–92.

Rechenthin, M. D. (2014). *Machine-learning classification techniques for the analysis and prediction of high-frequency stock direction*, The University of Iowa.

Shah, D., Campbell, W. and Zulkernine, F. H. (2018). A comparative study of lstm and dnn for stock market forecasting, *2018 IEEE international conference on big data (big data)*, IEEE, pp. 4148–4155.

Sun, J. (2016). Daily news for stock market prediction, version 1, `https://www.kaggle.com/aaron7sun/stocknews`. Retrieved [Date You Retrieved This Data].

Ye, J., Zhao, J., Ye, K. and Xu, C. (2021). Multi-graph convolutional network for relationship-driven stock movement prediction, *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp. 6702–6709.