Minor Project

Computer Science & Engineering
7th Semester

Title: News Sentiment Analysis for Stock Price Prediction

Computer Science and Engineering
University Institute of Engineering and
Technology Panjab University, Chandigarh –
160014, INDIA

Submitted By:

Saksham Agarwal (UE183090)

Saloni Goyal (UE183093)

Mentored By:

Dr Naveen Aggarwal

CERTIFICATE

I hereby certify that the work which is being submitted in this project work titled "**News Sentiment Analysis for Stock Price Prediction** " is in partial fulfilment of the requirement for the award of the degree of "Bachelor of Engineering in Computer Science and Engineering" submitted in UIET, Panjab University, Chandigarh, is an authentic record of my work carried out under the supervision of **Naveen Aggarwal** and refers to other researchers work which is duly listed in the reference section. The matter presented in this project work has not been submitted for the award of any other degree of this or any other university.

(**Saksham Agarwal - UE183090**)

(**Saloni Goyal - UE183093**)

This is to certify that the statements made above by the candidate are correct and true to the best of my knowledge.

Naveen Aggarwal
The designation CSE, UIET, Panjab University, Chandigarh – 160014

DEPARTMENT: COMPUTER SCIENCE AND ENGINEERING

VISION:

To be recognized as an international leader in Computer Science and Engineering education and research to benefit society globally.

MISSION:

- To move forward as frontiers of human knowledge to enrich the citizen, the nation, and the world.
- To excel in research and innovation that discovers new knowledge and enables new technologies and systems.
- To develop technocrats, entrepreneurs, and business leaders of the future who will strive to improve the quality of human life.
- To create world-class computing infrastructure for the enhancement of technical knowledge in the field of Computer Science and Engineering.

PROGRAMME:B.E. CSE (UG PROGRAMME)

PROGRAMME EDUCATIONAL OBJECTIVES:

I. Graduates will work as software professionals in an industry of repute.
II. Graduates will pursue higher studies and research in engineering and management disciplines.
III. Graduates will work as entrepreneurs by establishing startups to take up projects for societal and environmental causes.

PROGRAMME OUTCOMES:

A. Ability to effectively apply knowledge of computing, applied sciences and mathematics to computer science & engineering problems.
B. Identify, formulate, research literature, and analyze complex computer science & engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

C. Design solutions for computer science & engineering problems and design system components or processes that meet the specified needs with appropriate consideration for public health and safety, and cultural, societal, and environmental considerations.
D. Conduct investigations of complex problems using research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
E. Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to different computer science & engineering activities with an understanding of the limitations.
F. Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
G. Understand the impact of professional engineering solutions in societal and environmental contexts and demonstrate the knowledge of and need for sustainable development.
H. Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
I. Ability to function effectively as a member of a team assembled to undertake a common goal in multidisciplinary settings.
J. Ability to communicate effectively to both technical and non-technical audiences.
K. Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
L. Recognition of the need for and the ability to engage in life-long learning. The ability to successfully pursue professional development.

COURSE OUTCOMES (CO):

1. Identify research and development-oriented development-oriented projects based on problems of practical and theoretical interest
2. Explore, identify and use appropriate methodologies and tools for implementation of identified problem
3. Demonstrate solution at periodic intervals for suggestions and review
4. Write a report outlining the entire problem, including literature survey and various results obtained

Abstract

Stocks price prediction is a current hot spot with great promise and challenges. Recently, there have been many stock price prediction methods. In this project, we propose a stock price prediction method that incorporates multiple data sources and investor sentiment. Firstly, we collect historical transaction data from **Yahoo Finance** and News Headlines of a respective company from **Finhub** for a period of time. Preprocessing the historical datasets, we calculate the technical indicators. Then, we use the sentiment analysis method based on transformers **(BERT**) for the non-traditional data (news data), which can calculate the investors' sentiment index. Finally, we combine sentiment index, technical indicators and stock historical transaction data as the feature set of stock price prediction and adopt **Random Forest Regression** to predict the close price of the 5 listed companies including **Goldman Sachs, JP Morgan, Morgan Stanley, Amazon and Microsoft.** We evaluate the model on three performance metrics including Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. The experiments show that the predicted stock closing price is closer to the true closing price than the single data source, which is better than traditional methods.

Table of Contents

List of Figures

1. Introduction

Investing carries a certain degree of uncertainty and hence is risky. An investor might predict a stock's value to increase over time but may lose or gain money on that investment. Being able to predict the future price of a stock is a very powerful tool for an investor.

2021 has seen a substantial increase in the number of investors. According to a report by SBI, "*The number of individual investors in the market has increased by a whopping 142 lakh in FY21*". Being able to analyze the reaction of major events like changes in management and major acquisitions on the future movements of stock prices could be useful for an investor. Since it is not possible for every individual to analyze the above, the investment strategies of new investors therefore usually rely on the sentiment of successful investors. This method of analysis is usually called **fundamental analysis.**

The stock prices are not only affected by the news about the company but the historical transaction data including the opening price, highest price, lowest price, the volume also affect the stock prices. The prices are also largely impacted if a stock is either overbought or oversold which brings into the picture the technical indicators including Stochastic Oscillator (%K) and Relative Strength Index (RSI). This method of analysis is called **technical analysis.**

2. Related Work

We in this project aim to combine both fundamental and technical methods to produce a feature set for better prediction of stock prices. This approach has been used in past as per [3] but the sentiments for the news data were calculated by a model based on CNN but we here propose to use a BERT transformer, which reads the entire sequence of words at once. Therefore, it is considered bidirectional. This allows the model to learn the context of a word based on all of its surroundings.

To train and evaluate the model, we use Random Forest Regression algorithm which as per [5], gives the best results for stock price prediction among other Machine Learning Models.

3. Program Design

3.1 Technical Analysis

3.1.1 Historical Data Preprocessing

For calculating technical indicators, we first collected the historical transaction data from Yahoo Finance using the yfinance API. We chose five stocks of listed companies Including Goldman Sachs, J.P. Morgan Chase, Morgan Stanley, Amazon and Microsoft. The transaction data included Date, Open Price, High Price, Low Price, Close Price, Volume, Adj. Close Price and Stock Code. It is critical to remove unnecessary information and as Adj Close Price is not required for future prediction, we leave that.

We collected the data from 1 January 2021 to 31 November 2021 for each of the 5 stocks. For each stock, we get the data for 229 days.
Below is the sample of preprocessed historical data of stock.

	Date	Open	High	Low	Close	Volume
0	2021-12-01	3545.0	3559.879883	3441.600098	3443.719971	3745800
1	2021-12-02	3460.0	3492.699951	3423.750000	3437.360107	3236300
2	2021-12-03	3455.0	3469.870117	3338.600098	3389.790039	4032600
3	2021-12-06	3393.0	3473.909912	3338.689941	3427.370117	3443000
4	2021-12-07	3492.0	3549.989990	3466.689941	3523.290039	3320500

Fig. 1

3.1.2 Technical indicators calculation

The technical analysis method mainly analyses the stock price fluctuation of the company according to the historical stock trading data and charts, and the technical indicators are often used in the technical analysis method. We calculated 2 technical indexes, which are stochastic oscillator index (%K) and relative strength index (RSI). The stochastic oscillator index (%K) reflects the correlation between the price range and the closing price in a given period. The RSI is very suitable for the short-term volatility of stock prices. These technical indicators measure the timing of overbought and oversold to a certain extent, and can also affect the

decision-making of investors. Usually, investors will buy stocks when the market is oversold and sell stocks when the market is overbought, so these technical indicators can also affect the fluctuation of stock prices.

The technical indicators are calculated based on the historical trading data of stocks. Their calculation is shown in the formulas (1), (2), (3). The TA-lib algorithm, a common library for quantitative trading in Python, is used to calculate the technical indicators.

$$\%K = 100.(C - Lt/\,Ht - Lt) \qquad (1)$$

$$RSI = 100 - (100/(1 + RS)) \qquad (2)$$

$$RS = Avg.\ Gain/Avg.\ Loss \qquad (3)$$

,where C is the close price, the H_t and L_t respectively denote the high price and low price for the last t days, and the value of t is set to 7 in our experiments.

Below is the sample of technical indexes calculated for stock.

Date	Open	High	Low	Close	Volume	RSI	%K
2021-12-01	3545.0	3559.879883	3441.600098	3443.719971	3745800	50.097572	65.446477
2021-12-02	3460.0	3492.699951	3423.750000	3437.360107	3236300	50.097572	65.446477
2021-12-03	3455.0	3469.870117	3338.600098	3389.790039	4032600	50.097572	65.446477
2021-12-06	3393.0	3473.909912	3338.689941	3427.370117	3443000	50.097572	65.446477
2021-12-07	3492.0	3549.989990	3466.689941	3523.290039	3320500	50.097572	65.446477

Fig. 2

3.2 Fundamental Analysis

3.2.1 Text Data Preprocessing

For fundamental analysis, we collected the news data stocks using Finhub API. The Finhub API allows a user to collect the news data of any company for a maximum period of 1 year. We collected the news data of S&P 500 companies from 1 January 2021 to 31 November

2021. The API compiled the data in a database file. We further queried the data of 5 listed companies including Goldman Sachs, JP Morgan Chase, Morgan Stanley, Amazon and Microsoft.
The compiled dataset of all 5 companies contained 34,512 records. The raw data included DateTime, Headline, id, Image, Source, Summary and Stock Code. Removing the unnecessary details, we were left with DateTime, Headline and Stock Code.

The Dataset for each stock is preprocessed to remove headlines that were not in the English Language using the inbuilt Python Lang Detect library.
Below is the sample of preprocessed news data of stock.

	Date	Time	headline	related
0	2021-12-02	00:00:00	What May Prove Truly 'Transitory' Is FAAMG's R...	AMZN
1	2021-12-01	23:07:11	Looking Ahead to the Q4 Earnings Season	AMZN
2	2021-12-01	22:51:06	Basket Trading With ETPs Gives Tactical Advant...	AMZN
3	2021-12-01	21:32:59	Facebook's Meta Ticker Trades in Canada, Ignor...	AMZN
4	2021-12-01	19:12:26	Amazon eyes downtown D.C. for another full-ser...	AMZN

Fig. 3

3.2.2 Sentiment Analysis using BERT

BERT stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers.

We find the sentiment whether positive or negative for each stock using BERT with the support of ktrain. We used the implementation of pre-trained BERT provided by ktrain to classify the sentiments.
Using the dataset provided by Google, which included 2 files, ‘train.csv’ and ‘test.csv’, we first train the BERT model on ‘train.csv’ that had 25000 records. To train the model, we first found the optimal learning rate using lr_find and train the model at 2e-5 learning rate. Then we use the model on our preprocessed news data to find the sentiment of each headline of a stock.

Below is the sample sentiment analysis for each headline of a stock.

	Date	Time	headline	related	Sentiment
0	2021-12-01	20:13:23	Goldman Sachs picks its favorite under-the-rad...	GS	pos
1	2021-12-01	16:57:00	Goldman Sachs Group Inc. stock underperforms W...	GS	pos
2	2021-12-01	16:05:00	Why BioXcel Therapeutics Stock Is Slumping Today	GS	neg
3	2021-12-01	13:40:01	Is the Options Market Predicting a Spike in Go...	GS	neg
4	2021-12-01	10:31:00	Dow's nearly 400-point rally highlighted by ga...	GS	pos

Fig. 4

3.2.3 Sentiment Index Calculator

In the news platform, after an event occurs, the public will have some comments on the company, and news articles will also report. These texts have an impact on the stock movements to a certain extent. After the sentiment analysis of financial texts, each news or post is classified as positive or negative. However, this is only the result of text classification. We need to calculate the overall sentiment tendency of the public in a day based on the number of positive and negative texts. The sentiment index is calculated based on the ratio of the sum and the difference between the number of positive and negative texts. The sentiment index is defined based on formula (4).

$$Sentiment\ Index = (Mtpos - Mtneg)/(Mtpos + Mtneg) \quad (4)$$

, where Mtpos is the total number of positive news on day t and Mtneg is the total number of negative news on day t. The range of the sentiment index is between -.05 and +0.5, and the sentiment index below 0 means that the sentiment is negative on t day.

Below is the sample of the sentiment index calculated for a stock.

Sentiment Date	neg	pos	Sentiment_index
2021-12-01	3.0	8.0	0.454545
2021-12-02	4.0	10.0	0.428571
2021-12-03	2.0	9.0	0.636364
2021-12-04	0.0	1.0	0.500000
2021-12-05	1.0	3.0	0.500000

Fig. 5

3.3 Stock Price Prediction

3.3.1 Dataset Compilation

We want to explore whether investor sentiment and technical indicators affect stock price movements. We combine the technical indicators, historical transaction data and sentiment index of each stock calculated for a period of 11 months (1 January 2021 - 31 November 2021). The selected feature sets contain nine-dimensional vectors, which are the open price, close price, high price, low price, volume, sentiment index, stochastic oscillator (%K) and RSI. We combine the dataset of all five stocks and the combined dataset contained 1145 records.

	Date	Open	High	Low	Volume	RSI	%K	Sentiment_index	Close
0	2021-01-04	267.000000	267.579987	260.160004	3572000.0	96.178173	95.960705	0.250000	265.000000
1	2021-01-05	263.880005	273.500000	262.570007	4207100.0	96.178173	95.960705	0.700000	270.929993
2	2021-01-06	276.290009	288.380005	273.100006	6383500.0	96.178173	95.960705	0.090909	285.549988
3	2021-01-07	287.769989	295.890015	286.679993	4009700.0	96.178173	95.960705	0.400000	291.649994
4	2021-01-08	292.000000	292.279999	285.059998	2800800.0	96.178173	95.960705	-0.111111	290.079987

Fig. 6

3.3.2 Stock Price Prediction based on Random Forest Algorithm

We regard the problem of stock price prediction as a regression problem, not a classification problem. We use Random Forests to predict the close price for a stock. Random Forest Regression is a supervised learning algorithm that uses the ensemble learning method for regression. Random Forests are ensembles of decision trees and work by introducing

decorrelation between the trees by randomly selecting a small set of predictors at each split of the tree and outputting the mean of the classes as the prediction of all the trees.

In the training stage, the inputs to this model include: open price, high price, low price, volume, sentiment index, stochastic oscillator (%K), RSI, and the close price is a label. However, in the testing stage, there are only eight dimensions of data and these data are obtained by fusing the previous two modules. The complete dataset is split into train and test datasets, with the training dataset containing 70% of the random records and then trained using the algorithm.

3.3.3 Performance Evaluation Metric

The Model's performance is evaluated by 3 commonly used metrics which include mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE). The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The mean squared error (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line and squaring them. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how to spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. The three metrics are calculated by formulas (5), (6) and (7).

$$MAE = (1/n).\sum |f - y| \quad (5)$$

$$MSE = (1/n).\sum (f - y)^2 \quad (6)$$

$$RMSE = \sqrt{MSE} \quad (7)$$

where f is the predicted value and y is the true value.

We get a mean absolute error of 1.2476 on the test dataset, mean squared error of 3.2022 and root mean squared error of 1.7894.

4. Results and Discussion

To predict and evaluate results distinctively for each of the 5 companies, using sections 3.1 and 3.2, we compile a dataset from 1 December 2021 to 31 December 2021 and then use Random Forest Regression model to predict the close price for the period and then plot them as per the real close price of the stock for that period.

Below are the results for each company distinctively.

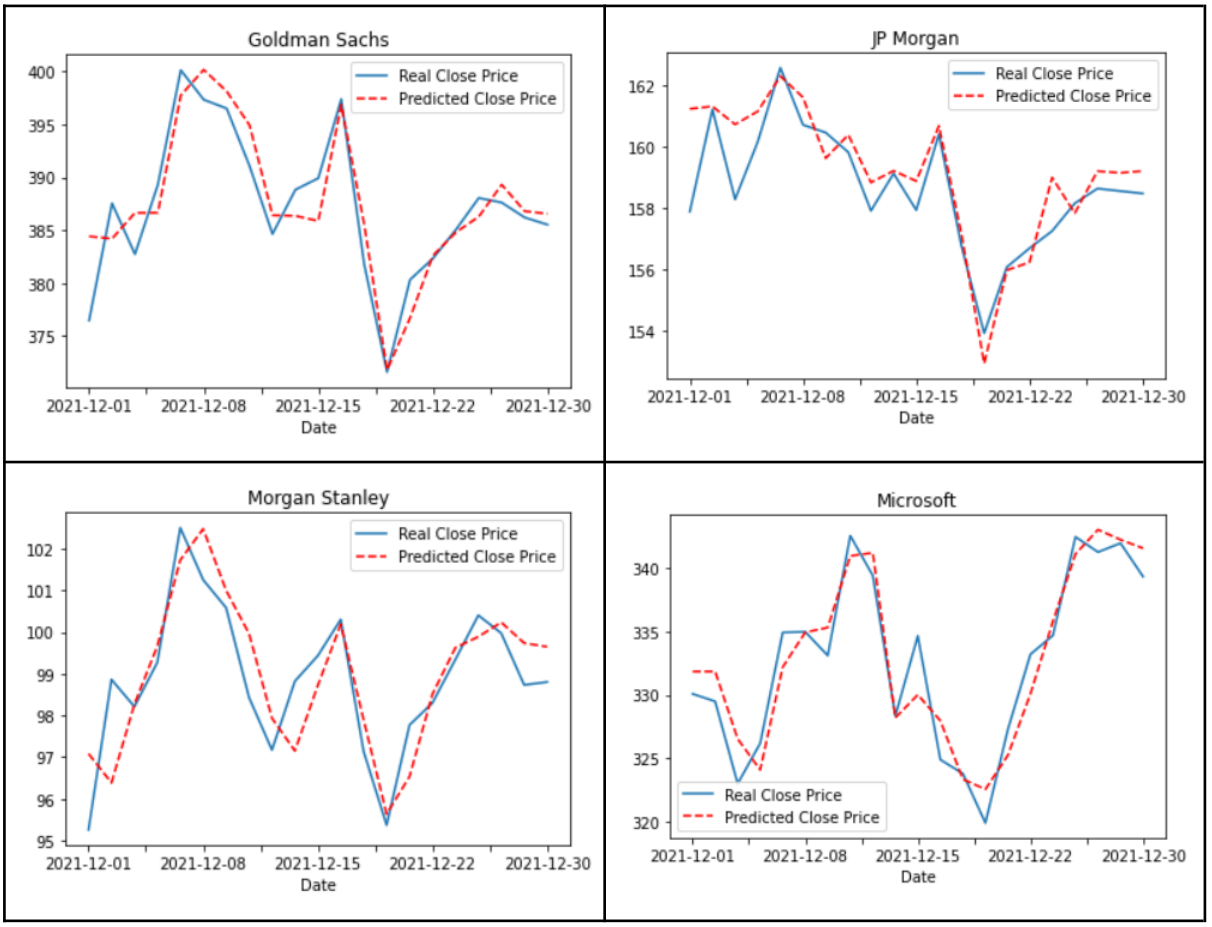

Fig. 7

Fig. 7

Conclusion and Future Work

In this project, we proposed a model to predict stock prices. We discussed the technical and fundamental analysis for stock price prediction and the impact technical indicators, historical transaction data and news data have on stock prices. For fundamental analysis, the sentiments of the news headlines were calculated with BERT which helped the model incorporate the effect of major events affecting the company.

However, due to the lack of time and unavailability of data, this project has limitations and needs to be improved in future research. With more data compiled and research, this model can be used for investment advice by investors and can have practical significance and the addition of other features like more technical indicators could also be beneficial in improving the results.

References

[1] Laszlo Nemes, Attila Kiss. Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, Volume 5, Pages 375-394, Jan 2021

[2] J. Kalyani, P.H.N. Bharathi, and P.R. Jyothi. Stock trend prediction using news sentiment analysis. *arXiv:1607.01958 [cs]*, July 2016

[3] Shengting Wu, Yuling Liu, Ziran Zou, Tien-Hsiung Weng .S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. *Taylor&Francis Online,* Jun 2021

[4] Chetan Gondaliya. Sentiment Analysis and prediction of Indian stock market amid Covid-19 pandemic. et al 2021 *IOP Conf. Ser.: Mater. Sci. Eng. 1020 012023*

[5] Ashwini Pathak, Sakshi Pathak. Study of Machine learning algorithms for stock market prediction. *International Journal of Engineering Research & Technology*, Volume 9, Jun 2020

[6] Sadorsky, Perry. 2021. A Random Forests Approach to Predicting Clean Energy Stock Prices. *Journal of Risk and Financial Management* 14: 48.

[7] Sara Abdali, Ben Hoskins. Twitter Sentiment Analysis for Bitcoin Price Prediction. Stanford, 2020

[8] Nirdesh Bhandari. Stock Market Trend Prediction Using Sentiment Analysis.

[9]https://medium.com/analytics-vidhya/finetuning-bert-using-ktrain-for-disaster-tweets-classification-18f64a50910b

[10] https://mrjbq7.github.io/ta-lib/func_groups/momentum_indicators.html

[11]https://python.plainenglish.io/access-historical-financial-news-headlines-with-python-be1b8faaea9f

Department of Computer Science and Engineering
UIET, Panjab University, Chandigarh

Course Exit Survey

Dear Student

The attainment of course outcome after the completion of the course is required as it would help in the continuous improvement of Course Outcomes (CO). This course exit survey would enable us to know to what extend the subject under consideration and the teaching methodology that has been practised in the institution have contributed towards the attainment of course outcomes. Hence you are asked to provide the attainment level on a scale of Very High (5), High (4), Medium(3), Satisfactory (2), Low (1) for the given course outcomes.

Name of the student: Saksham Agarwal **Roll No:** UE183090	**Year/Semester** 4 year & 7th sem	**Academic Year** 2021-2022
Course Name	Minor Project	
Course Code	CS - 757	
Teacher Name	Dr Ravreet Kaur	

Course Outcomes		**Course Outcome Attainment**				
		Very High (5)	**High (4)**	**Medium (3)**	**Satisfactory (2)**	**Low (1)**
CO1	Identify research and development-oriented project based on problems of practical and theoretical interest	✅				
CO2	Explore, identify and use appropriate methodologies and tools for implementation of identified problem		✅			
CO3	Demonstrate solution at periodic intervals for suggestions and review	✅				
CO4	Write a report outlining the entire problem, including literature survey and various results obtained	✅				

Suggestions for Improvement: None

Saksham Agarwal

Signature of Student

Department of Computer Science and Engineering
UIET, Panjab University, Chandigarh

Course Exit Survey

Dear Student

The attainment of course outcome after the completion of the course is required as it would help in the continuous improvement of Course Outcomes (CO). This course exit survey would enable us to know to what extend the subject under consideration and the teaching methodology that has been practised in the institution have contributed towards the attainment of course outcomes. Hence you are asked to provide the attainment level on a scale of Very High (5), High (4), Medium(3), Satisfactory (2), Low (1) for the given course outcomes.

Name of the student: Saloni Goyal **Roll No:** UE183093	**Year/Semester** 4 year & 7th sem	**Academic Year** 2021-2022
Course Name	Minor Project	
Course Code	CS - 757	
Teacher Name	Dr Ravreet Kaur	

Course Outcomes		**Course Outcome Attainment**				
		Very High (5)	**High (4)**	**Medium (3)**	**Satisfactory (2)**	**Low (1)**
CO1	Identify research and development-oriented project based on problems of practical and theoretical interest	✅				
CO2	Explore, identify and use appropriate methodologies and tools for implementation of identified problem		✅			
CO3	Demonstrate solution at periodic intervals for suggestions and review		✅			
CO4	Write a report outlining the entire problem, including literature survey and various results obtained	✅				

Suggestions for Improvement: None

Saloni Goyal

Signature of Student