CMP6200/DIG6200 Individual Undergraduate Project 2023-2024

Impact of User Sentiment on the Financial Markets

Coursework Assignment Report

Gurpreet Singh
21131818

Supervisor Emmett Cooper

Faculty of Computing, Engineering and the Built Environment
Birmingham City University

July 2024

Contents

List of Tables

List of Figures

Abstract

Summary of the entire report

Acknowledgements

I would like to thank my tutor who has given me insightful feedback and advice over the course of the year. My family, who helped through though times and helped me get through this assignment motivating me constantly.

1 Introduction

The research report will provide an informed evaluation of the impact of user sentiment over the financial markets using machine learning methods; this will be determined with social media and news outlet's data. The report will detail work done previously and examine it whilst providing a new outlook to the research.

1.1 Problem Definition

Investigate the impact of user sentiments on the financial markets by investigating the effect of the different media outlets throughout major geopolitical events.

1.2 Scope

The scope of this research is to investigate the impact of user sentiment onto the stock market utilizing machine learning techniques which outline and analyzing the results using evaluation metrics to investigate effectiveness and efficiency. Different analysis techniques: fundamental analysis, technical analysis and sentimental analysis are to be explored in order to ensure that generally overlooked factors do not limit this study.

1.3 Background rationale

Finance, "a term for matters regarding the management, creation, and study of money and investments" (Hayes, 2023), is a critical sector in the economic growth of a country/ region. The allocation of resources which is done in the financial markets oversees the highest monetary transfers in any industry, all kinds of exchangeable items are traded worldwide.

Financial markets provide the opportunity for individuals and organisations to buy and sell equity along with other valuable items (commodities, forex, crypto) and thereof profit from the fluctuations in price of each item. Although, financial markets are seen unpredictable and highly volatile there are multiple theories which are used to predict market price movements some are proven, some are still contradictory as there is not enough evidence to completely prove the theory or disapprove the theory.

These theories attempt to provide an explanation as to the movements of each share however as the lack of substantial evidence it is not collectively decided which approach would be suitable as there are always some factors are not accounted and trading in such way can be seen as a "gamble".

One of the theories going to be investigated in depth in the literature review is the random walk hypothesis. Random random walk suggests that the short-term movements cannot be predicted as they are not affected by factors which can be calculated beforehand, rendering technical analysis baseless. Technical analysis is the use of previous data and stock market history to predict future movements.

Technical analysis uses a limited amount of data overlooking company reports and assets data which are factored in through fundamental analysis which is overlooked by doing technical analysis individually. Both analysis methodologies would work infinitely well if there were no other external factors which have a substantial impact on the movements, however this is not the case.

Through contradictory beliefs it has been speculated that market sentiments have a huge role in the movements of markets therefore without them any analysis would not be correct. The role of sentimental analysis has been researched various times however the lack of enough evidence cannot fully prove the hypothesis. Additionally, all the studies are focused on single companies or major events therefore the daily effect of market sentiments in the stock market are still to be explored.

1.4 Project Aim and Objectives

1.4.1 Aim

Investigate user sentiment impact on stock market through the analysis of past data from media outlets, traditional and modern.

1.4.2 Objectives

These objectives set milestones to follow throughout the project as shown in Table 1 column "stage".

Obj. No.	Stage	Description
1	Proposal	Constructively examine given feedback and make updated project timeline.
1.1	Proposal	Update timeline (Gantt chart) to ensure it matches with feedback given.
1.2	Proposal	Make a precise literature review plan by week 5 and therefore update literature search methodologies.
2	Lit review	Review potential literature to use in the project to comprehend specific parts of the project and arrange a list.
2.1	Lit review	Analyze and categorize the different sources to align with the Key Themes/Topics.
2.2	Lit review	Extract valuable information into a literature review report.
2.3	Lit review	Criticize each source and findings.
3	Lit review	Evaluate the results and conduct small-scale prototyping to justify results.

Continued on next page

Table 1 – continued from previous page

Obj. No.	Stage	Description
4	Lit review	Assess the solutions and depending on those determine the need of more evidence to determine the best solution.
5	Lit review	Repeat objectives 3 and 4 till a plan with a small-scale prototype is built to justify theory.
6	Design	Analyze the literature review and assemble a solution to the problem.
7	Design	Based on the solution, prepare a plan using the most efficient design methodology to implement the solution.
8	Implementation	Create the first prototype of the machine learning model, by using python and MATLAB to access graphs and visualise the primary results.
8.1	Implementation	Complete pre-processing on the datasets.
8.1.1	Implementation	The dataset is divided appropriately in 80-10-10 split.
8.1.2	Implementation	The dataset has no missing values and is all in numerical format.
8.2	Implementation	Experiment with the prototype and reconstruct it with more methods, getting an accuracy score which is above 90%
8.3	Implementation	Demonstrate finished product, by using graphics and presenting clear results.
9	Implementation	Construct an interface.
10	Implementation	Construct different charts to visualise different finance methodologies, present at least one graph for each hypothesis.
11	Implementation	Construct unit tests.
12	Implementation	Construct automated testing, using testing dataset to get an accuracy on unseen data.
13	Testing	Apply unit tests to the prototype and store results into csv files for later comparison.
14	Testing	Apply automation testing to the prototype and store results into a csv file.
15	Testing	Compare results to determine percentage of discrepancy.
16	Testing	Examine results and record changes to the accuracy score for each model.

Continued on next page

Table 1 – continued from previous page

Obj. No.	Stage	Description
17	Evaluation	Examine solution and record results and discuss aim and actual findings writing a report, suggest improvements.
18	Conclusion	Assess the report and all files and formulate a decision as per the aim and hypothesis.

Table 1: Table of Objectives

1.5 Research questions

Researched based questions allow us to focus the literature review in a specific area and elaborate only on parameters which are needed to implement a solution and gain knowledge in the area. These questions contain exploratory questions and falsifiable hypotheses to accomplish literature review objectives. The main hypothesis being explored are:

1. *Hypothesis number one: "Geopolitical events will not have any impact on the stock market."*

2. *Hypothesis number two: "War or geopolitical events do not create higher volatility in the stock market."*

3. *Hypothesis number three: "The trend seen for one specific event cannot be replicated for others."*

Hypothesis/ Question Number	Hypothesis and Exploratory Questions
1	"Russian-Ukraine war had no impact on the stock market."
2	"The S&P 500 indices are not affected more than 0.05%."
3	"Companies in the military manufacturing will not have positive returns due to war."
4	How are different sectors in the stock market affected?
5	What is the volatility relationship with the uncertainty of war?
6	How does media affect user sentiment and then the stock market?

Table 2: Hypothesis and Exploratory Questions

2 Literature Review

2.1 Introduction

The report will be used to provide evidence to the existing problems/gaps in knowledge found through research of past and contemporary literature, solidifying the need of this research project. Appendix A contains explanations of databases and search engines used along with keyterms.

2.2 Structure

An hybrid approach is undertaken where a narrative- systematic review will be carried out to consider different details in the below themes to better understand and identify the problematic areas and develop a research paper investigating this.

2.3 Themes

Theme	Description
Stock market	The stock market is the main focus for the research paper, this theme will be used to develop knowledge on the main subject.
Geopolitics(news outlets)	This will provide secondary and tertiary data which will be used to analyze the impact of the war on the stock market.
Fundamental Analysis	Exploration of how different approaches are taken to fundamental analysis with its theory.
Technical Analysis	Exploration of how different methods are expressed for technical analysis with its theory.
Sentimental Analysis	Exploration of how different approaches are taken to sentimental analysis with its theory.
AI/ Machine learning in finance	This will present a brief overview of the advancements in the field.
Evaluation	Evaluation techniques will be considered to ensure that results and conclusion is properly evaluated.

Table 3: Themes for Literature Review

2.4 The review

2.4.1 Stock market

The stock market is where investors buy and sell stock based upon maximising profits and minimising risk (Zouaghia, Aouina and Said, 2023). The stock market is very unpredictable and volatile therefore the gains can be substantial for investors which can be affected by a lot of various factors. The use of ANN and SVM along other machine learning models have been used where the evaluation metrics are calculated forming the equation:

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN} * 100 \quad (1)$$

with the accuracy score equation from Zouaghia, Aouina and Said (2023). To present the best result for which model can predict results and it was seen that RF had 61% where the lowest GB had 55% accuracy score explaining that these models by themselves cannot predict the stock market correctly therefore a mixed model is needed (Zouaghia, Aouina and Said, 2023). Moreover, this might suggest that the random walk theory is correct as a convincing prediction cannot be made.

3 Methods

3.1 Introduction

This section includes methodologies for various stages of the project which were undertaken to complete this project. Overall methodologies are outlined along with each section explicit methodology.

3.2 Project methodology

The methodology chosen for this project will present an early hybrid design approach as confining to either one of the strategies will not be beneficial for the project, the Spiral model. The spiral model is a hybrid of the waterfall-agile approaches as it consists mainly of 4 stages iterated in loops. A meta-model which consists of breaking down the project into segments, and undergo for main phases: planning, risk analysis, development, and evaluation (Alshamrani and Bahattab, 2015).

These four stages are iterated a set amount of time by the project lead till the final product is made. In the case of this project this is highly beneficial as we can make changes to the prototype until the final results after all the iterations are as expected. The advantages of this approach are that it has a strong documentation control and monitoring, and changes can always be made at later stages (Alshamrani and Bahattab, 2015).

However, this approach can be time consuming as there is a lot of repetitive tasks and evaluation analysis involved hence not always the best approach. This project being research based, the approach would provide a lot of comparable data therefore it is best suited. Finally, this is the best approach if others researchers want to further develop the project it will be easier to add iterative loops to the design methodology and implementation.

3.3 Literature search methodology

Through this section of the report, a detailed review is given for the final literature search methodology used and how each part is to be carried out. This will be a step-by-step guide of the execution of the search methodology.

The methodology is similar to the initial methodology as it will incorporate all the techniques. This will ensure that there is only relevant literature and credibility is not compromised. All themes will undergo the filtering process defined underneath individually so that research questions are answered thoroughly.

The first technique used will be the keyword search, this will determine all the literature available in said theme. Initially, this will ensure that there is enough literature volume for the literature review to be thorough in the topic area. Then, keywords from the semantic field will be used along Boolean search to refine the literature. After this is done the filtering will begin.

The filtration will be focusing on set parameters, which will determine if the literature is valid for the research these include year, authors, publishers, document type. These factors determine accreditation of the literature and year determines if the literature is still relevant. The year parameter will always above 2019 to establish all information is up to date. The authors will be determined by looking at occurrences whilst not undermining potential risk of not valuable data. Publishers will not be specifically checked this will only be done if there are still too many pieces left. The organisation of the files and data.

The process outlined, is to be carried out for all themes to gain all literature that is needed. This can be done for all databases and search engines using scripts which are to be used alongside the API of each database and search engine. Although, with Scopus this is not needed as it allows to do all the above on its webpage. In Scopus grey literature will be included in the search as to not distinguish it separately and keep all data and knowledge of it the same.

Finally, after the literature review draft is completed, the work will be put through several websites to ensure credibility and create a paper trail for each publication. This will be done through the following websites: Lit review visuals, Litmaps, Research rabbit and Connected papers. This action will ensure that citation search is complete.

4 Design

4.1 Introduction

The design stage of the project is a very important portion as it dictates an effective experimentation method to be carried out in the implementation stage. The clear instructions imposed in this stage will make the implementation run smoothly. The areas of this stage have been amended so that they can meet the type of outcome to be produced, a research-based project.

4.2 Design specification/User requirements

Design specifications allow to develop a clear understanding of what is needed for the project to be successful and set specific outcomes of the project.

1. Data collection needs to be completed before starting the implementation;
2. Model selection must be appropriate therefore all regression models should be used.
3. Data must be pre-processed accordingly to prevent any leakage.
4. Select features of dataset which are most consistent with variables to be examined.
5. Define appropriate evaluation metrics.
6. Fit and train model with specific parameters.
7. Implement cross-validation techniques to produce an accurate set of results.
8. Document each stage and provide feedback on process.
9. Analyse results at each stage (segment/loop).
10. Implement final prototype considering future application.
11. Use user friendly interfaces to visualise results.
12. Provide conclusive results for the research.
13. Evaluate all processes.

4.3 Concept Solution

The solution proposed is a research paper which will thoroughly examine the topic by replicate models specified in the literature review and furthermore developing on said knowledge by experimenting with different regression models and iterate this till the final model is made which is determined by the accuracy of the model.

4.4 Data Collection

Vital to the implementation of machine learning as data is used to fit the model and set parameters beforehand. The data collection needs to be done to ensure quality data is collected to be later used.

4.4.1 Primary data collection - Web Scraper

The web scraper, an integral part of the project, as it allows to fetch article data for the implementation of the models. The web scraper, initially designed to scrape publicly accessible data from various websites in order to minimise bias and increase credibility of data.

The design of the scraper has been modified several times over the course of the project to ensure that the most optimal and efficient design is chosen to strictly adhere to time guidelines.

Original Structure The use of different python scripts was firstly discussed to ensure that each component would work well independently and therefore not become an unsolvable issue where at the last step of the implementation all components would be imported into one file and that would be the executable.

The design revolved around each script being able to do only one component hence the python scripts were first designed as expressed in table 4.

Script Name	Description
scraper	loading the API or bs4.
fetching	accessing different data from the website.
dataFormat	import and export of data to and from the python data types.
main	executing all scripts in order to run the web scraper.

Table 4: Python scripts

Revised Structure The processes were not to be implemented the same way as a decision was made before implementing to use the OOP structure to ensure that a scraper object can be called and used as easily on future projects as well. The differentiation is that for this to be successfully be operational, all components were needed to be further broken down so that each component had subroutines which could be called individually. The subroutines which are listed in Table 2 are carefully outlined so that the implementation would be completed with great regards to time efficiency.

Python Modules It was important to understand that to gain access to data, a secure path was needed to the website, in light of this there were two potential pathways that were explored:

1. API;
2. Selenium;

Subroutine name	Function
init	set self variables
fetchdata	load and manipulate data from json files
loadHtml	get HTML of page
getarticledata	execution of all subroutines to get article data
getlinks	getting all article links from results page
main	executing all scripts in order to run the web scraper.

Table 5: WebScraper class subroutines

3. BeautifulSoup4(bs4).

All three of the options had their advantages and limitation, and predominantly all deemed as not viable options. API's were either specific to the different projects worked in, outdated or simply missing components. Whereas, selenium and bs4, due to lack of prior knowledge, became difficult to comprehend, establish a successfully connection and essentially scrape the website.

To transfer data to files the csv module was initially selected as the most expertise were with that file type, however this was deemed inefficient as the data was not stored to be easily accessible.

Request module was also needed in coherence to the above outlined options 1-3. This module was used to send requests which allowed to gain access to specific data from the HTML.

4.4.2 Secondary data collection - Datasets

Secondary data collection ensures that the data which is extracted does not affect the credibility of the data, but also allows a greater perspective on data which is not otherwise available.

Social Media datasets

Twitter datasets, which can no longer be compiled freely through the website, collection allows to investigate social media. Through the use of financial tickers the dataset can be filtered and used for modelling.

Traditional news datasets Traditional news datasets allow access to otherwise paid news outlets which cannot be easily accessed without paying premiums. Collection through search engines and filtration can be done to ensure only viable data is used.

4.5 Analysis Implementation Design

The following steps will be carried out for each type of analysis: sentimental analysis, fundamental analysis, technical analysis and then be merged into one model to determine the overall outcome.

1. Data collection needs to be completed before starting the implementation;

2. Select the "Ground truth";
3. Split data into training, testing and validation data;
4. Carry out pre-processing of data, consider missing values, non-numerical data and fitting positioning;
5. Select a machine learning model;
6. Define parameters;
7. Define evaluation metrics;
8. Implement machine learning model;
9. Extract primary results;
10. Save in csv file;
11. Iterate step 5 till 10 on different models;
12. Compare results;
13. Decide most effective model;
14. Initialise testing;
15. Test Final Model;
16. Combine all the analysis into one machine learning model using all results provided;
17. Create visual representations;
18. Evaluate results;
19. Provide conclusions to the hypothesis.

4.6 Testing strategies

Testing strategies are needed to determine the best and final outcome for the project. Testing strategies refers to the testing approach undertaken in the implementation cycle (*Test strategy*, 2023). There will be three testing strategies used: manual testing, automated testing, and unit tests.

Manual testing is programmed and executed by the researcher manually whereas automation testing revolves around developing test scripts which can execute specific test cases automatically (Sharma, 2014). On the other hand, unit tests execute a small part of the code in isolation to understand if it works correctly and gives expected results (Olan, 2003).

4.7 Evaluation

Finally, after all stages have been completed performing a set of calculation create a conclusion of the models using already available metrics in the used modules, this will increase efficiency and be in line with the time restrictions.

Furthermore, complete an evaluation of the whole projects outlining further changes or modifications to be implemented at later stages. Also, outline any potential challenges which can be avoided in further examination.

5 Implementation

5.1 Introduction

This section of the report will solely focus on the technical side of this report in which all steps taken will be explained.

5.2 Technical analysis

5.2.1 Exploratory data analysis(EDA)

The begininning of the EDA all needed modules are installed where seaborn and matplotlib is used for data visualisation and pandas for dataframe creation and manipulation. 'os' module is used to access data files from the device.

```
# Terminal command to install required modules
#importing modules
!pip install seaborn matplotlib
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import os
```

There where only two subroutines used as global variables where needed to access files and data manipulation.

```
def get_data():
    #access and load data from the device into readable dataframes

    current_dir = os.getcwd()
    sub_dir = os.path.join(current_dir, 'data/stock_data')
    fileholder = os.listdir(sub_dir)
    df_dir = {}
    for file in fileholder:
        if file.endswith('.csv'):
            src=os.path.join(sub_dir, file)
            df = pd.read_csv(src)
            keyText = os.path.splitext(file)[0]
            keyPfirst = str(keyText.split('_')[0])
            keyPlast = (keyText.split('_')[-1])

            key = (f'{keyPfirst}_{keyPlast}')
            df_dir[key] = df
    return df_dir
```

```
def get_name(key):
    #get company stock name
    key = key.split('_')[0]
    return stock_names[key]
```

get_data access current directory path and searches for csv files in the specified folder path 'sub_dir'. Using conditional statements to find file and create Data frames.

The get_name subroutine on the other hand goes through the dataframes and uses a .split() function on the key to get the first value of the split which is a company indicator.

```
df_dir = get_data()
stock_names = {'AAPL': 'Apple Inc.', 'GOOGL': 'Google (GOOGL)', 'GOOG':
    'Google (GOOG)', 'AMZN':'Amazon', '^GSPC':'S&P 500', 'TSLA':'Tesla',
    '^FTSE': 'FTSE 100', '^FTMC': 'FTSE 250' }
```

These two global variables(df_dir and stock_names) allow the repetitive use of the data without having to execute the subroutine multiple times.

```
    df_dir['AAPL_daily'].columns
```

This command was used to identify columns present in the dataset.

Subroutine for each EDA command is used as there are several companies and it was more efficient to use conditional statements.

```
    #data checking
    def get_shape(interval):
    for key, df in df_dir.items():
        if key and key.endswith(interval):
            company_name = get_name(key)
            print(f"Dataframe size for {company_name}:")
            print(df.shape)
    def get_info(interval):
    for key, df in df_dir.items():
        if key and key.endswith(interval):
            company_name = get_name(key)
            print(f"Datatypes of {company_name}:")
            print(df.info())
            print('      ----------------------------------------      ')

    def mvalue_check(interval):
    for key, df in df_dir.items():
        if key and key.endswith(interval):
            company_name = get_name(key)
```

```
        print(f"Datatypes of {company_name}:")
        print(df.isna().sum())
        print('      ---------------------------------------      ')
```

The print commands where used to make the results more user friendly. The code as seen, uses conditional statements, in all subroutines, to get information for each company data frame individually. The get_shape provides size for each data frame with maximum columns and rows. The get_info subroutine utilises get.info() to access all data properties and mvalue_check ensures there are no missing values using isna().sum().

```
#data visualisation
def line_plot(interval, attr):
for key, df in df_dir.items():
    if key and key.endswith(interval):
        df['Date'] = pd.to_datetime(df['Date'])
        df_sorted = df.sort_values(by='Date')
        company_name = get_name(key)
        plt.figure(figsize=(12, 6))
        plt.plot(df_sorted['Date'], df_sorted[attr],
            label=f'{company_name} {attr} Price', linewidth=2)
        plt.title(f'{company_name} {attr} price')
        plt.xlabel('Date')
        plt.ylabel('Close price')
        plt.legend()
        plt.show()

def company_chart(interval, exclusion):
plt.figure(figsize=(10, 6))
for key, df in df_dir.items():
    if key and key.endswith(interval)and not any(exclusion in key for
        exclusion in exclusions):
        company_name = get_name(key)
        df['Date'] = pd.to_datetime(df['Date'])
        df_sorted = df.sort_values(by='Date')
        sns.lineplot(x='Date', y='Adj Close', data=df_sorted,
            label=company_name)

plt.title('Adj Close stock price comparison')
plt.xlabel('Date')
plt.ylabel('Closing Stock Price')


plt.legend()
```

```
plt.tight_layout()
plt.show()

def heatmap(interval):
for key, df in df_dir.items():
    if key and key.endswith(interval):
        company_name = get_name(key)
        corr = df[['Open', 'High', 'Low', 'Close', 'Adj Close']].corr()
        plt.figure(figsize=(8,8))
        sns.heatmap(corr, annot=True, cmap='coolwarm')
        plt.title(f'Correlation attributes for {company_name}')
        plt.show()
```

Each subroutine is defined with its purpose and creates the aforementioned graphs. Three graph types were used and matplotlib was used as plt to construct the graphs or seaborn as sns. Conditional statements are a constant recurrences throughout as to facilitate all company data.

The results are shown on the EDA_stockdata notebook which is individual to other files. Along with this the charts can also be seen in Appendix B.1.

5.2.2 Pre-processing

This was carried out in the stockdata_modelling notebook where all the technical analysis was done.

```
    import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.svm import SVC
from sklearn.svm import SVR
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import os
```

The modules mentioned where used to create an SVM model as it was deemed to be the most successful between the 4 used in the sentimental analysis.

```
    for key, df in df_dir.items():
        if key and key.endswith(interval):
            df.index = pd.to_datetime(df['Date'])
            df = df.drop(['Date'], axis='columns')
            print(df.head())
```

Indexing date so that it can be used as a uniqure identifier.

```
    X= {}
```

```
y= {}
for key, df in df_dir.items():
        if key and key.endswith(interval):
            company_name = get_name(key)
            df['Pct Change'] = df['Adj Close'].pct_change()
            df['Moving Avg'] = df['Adj Close'].rolling(window=5).mean()

            df.dropna(inplace=True)

            X_var = df[['Pct Change', 'Moving Avg']]
            y_var = df['Adj Close']

            X[company_name]= X_var
            y[company_name]= y_var
```

Added percentage change and moving average columns and defined X and y variables for modelling.

```
    split_percentage = 0.8
X_train = {}
X_test = {}
y_train = {}
y_test = {}

for company_name, X_var in X.items():
    split = int(split_percentage*len(X_var))
    X_train[company_name]= X_var[:split]
    X_test[company_name] = X_var[split:]

for company_name, y_var in y.items():
    split = int(split_percentage*len(y_var))
    y_train[company_name] = y_var[:split]
    y_test[company_name] = y_var[split:]
```

Splitting the dataset 80/20 as per train and test data.

5.2.3 Modelling

```
        classifiers = {}
for company_name in y_test:
    y = y_train[company_name]
    X = X_train[company_name]
    svm = SVR(kernel='rbf')
    cls = svm.fit(X, y)
```

```
    # You might want to store the classifier for each company
    # For example, in a dictionary
     classifiers [company_name] = cls
```

The SVR model was fitted as it is for regression data.

```
        predictions = {}
for company_name in y_test:
    cls = classifiers [company_name]
    X = X_test[company_name]
    y_pred = cls.predict(X)
    print(company_name,':')
    predictions[company_name] = y_pred
```

Predictions where made and stored in a list to be later used.

5.3 Sentimental analysis concept

5.3.1 Exploratory data analysis(EDA

The main python libraries imported are Sklearn, pandas, matplotlib and seaborn. Pandas and Sklearn are both used for executing machine learning techniques whereas matplotlib and seaborn are for result visualisation. NumPy is used for the mathematical operations in between each segment.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as s
import datetime
import numpy as np
import nltk
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, confusion_matrix
```

Figure 1: python libraries used

Figure 2: terminal command- install datasets

Using a terminal command 'pip' the dataset package from "huggingface" website is downloaded to ensure connectivity to dataset and comply with company policies.

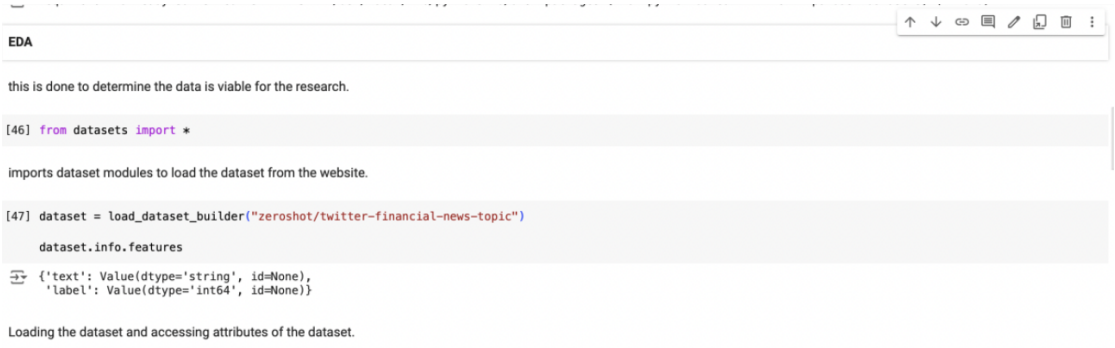

Figure 3: EDA

All attributes are displayed to gain knowledge about the dataset and make sure it is viable with the research.

```
[48] get_dataset_split_names("zeroshot/twitter-financial-news-topic")

→ ['train', 'validation']
```

Understanding the splits executed between the dataset.

```
[49] data_train = load_dataset("zeroshot/twitter-financial-news-topic", split="train")

     data_test = load_dataset("zeroshot/twitter-financial-news-topic", split="validation")
```

Putting both splits into different variables.

```
[50] print (data_train)

→ Dataset({
      features: ['text', 'label'],
      num_rows: 16990
  })
```

Outputting training dataset attribute details.

```
[51] print (data_test)

→ Dataset({
      features: ['text', 'label'],
      num_rows: 4117
  })
```

Outputting validation dataset attribute details.

Figure 4: Splitting dataset

The dataset is split into two parts to be use in the modelling and preprocessing with an 80/20 split.

5.3.2 Pre-processing

Pre processing is done to make sure the computer can understand the data. A bag of words method is used as there were difficulties whilst accessing transformers correctly.

```
[52] x_train = [example["text"] for example in data_train]
     y_train = [example["label"] for example in data_train]

     x_test = [example["text"] for example in data_test]
     y_test = [example["label"] for example in data_test]
```

```
[53] count_vectorizer = CountVectorizer()
```

```
[63] x_train_b = count_vectorizer.fit_transform(x_train)
     x_test_b = count_vectorizer.transform(x_test)
```

```
[64] count_vectorizer_test = CountVectorizer(vocabulary=count_vectorizer.vocabulary_)
```

```
[65] x_test_b = count_vectorizer_test.fit_transform(x_test)
```

Figure 5: Preprocessing code

Simplistic models are used as to provide proof of concept of what can be achieved through ml in sentimental analysis.

5.3.3 Modelling

There are 4 types of traditional models used to provide proof of concept. The models were fitted and trained and then prediction commands were used on all to make a prediction and after displaying an accuracy score the main process was completed. All the models were from the sci-kitlearn library.

logistics regression

```
[66] logistic_regression = LogisticRegression(max_iter = 500)
     logistic_regression.fit(x_train_b, y_train)
```

```
▾ LogisticRegression
LogisticRegression(max_iter=500)
```

```
[68] logistic_regression_predictions = logistic_regression.predict(x_test_b)
```

```
[69] logistic_regression_accuracy = accuracy_score(y_test, logistic_regression_predictions)
     print("Logistic Regression Accuracy:", logistic_regression_accuracy)
```

```
Logistic Regression Accuracy: 0.8350740830701967
```

Figure 6: Logistic regression model fitting

Naive bayes

```
[74] naive_bayes = MultinomialNB()
     naive_bayes.fit(x_train_b, y_train)
     naive_bayes_predictions = naive_bayes.predict(x_test_b)
     naive_bayes_accuracy = accuracy_score(y_test, naive_bayes_predictions)

[75] print("Naive Bayes Accuracy:", naive_bayes_accuracy)

     Naive Bayes Accuracy: 0.6963808598494049
```

Figure 7: Naïve Bayes model fitting

```
random_forest = RandomForestClassifier(n_estimators=100, random_state=42)
random_forest.fit(x_train_b, y_train)

▾ RandomForestClassifier
RandomForestClassifier(random_state=42)

[80] random_forest_predictions = random_forest.predict(x_test_b)

[81] random_forest_accuracy = accuracy_score(y_test, random_forest_predictions)
```

Figure 8: Random Forest model fitting

```
[85] svm = SVC(kernel='linear')
     svm.fit(x_train_b, y_train)

▾ SVC
SVC(kernel='linear')

[87] svm_predictions = svm.predict(x_test_b)

[88] svm_accuracy = accuracy_score(y_test, svm_predictions)

[89] print("SVM Accuracy:", svm_accuracy)
```

Figure 9: SVM model fitting

6 Evaluation

6.1 Introduction

The evaluation ultimately provides with the summary of the project refining it and allowing the findings to be credible and evident. The evaluation that was discussed was time series cross validation as all data was going to be time specific.

6.2 Evaluation metrics

The evaluation ultimately provides with the summary of the project refining it and allowing the findings to be credible and evident. The evaluation that was discussed was time series cross validation as all data was going to be time specific. Along with these for the stock data charts were made to allow data visualisation and correctly understand the different results.

6.3 Datasets

Secondary datasets were used due to time restrictions.For the news data and social media data, a dataset from a trusted website was used during the research (huggingface). The dataset was decided upon after months of futile work trying to get access to real time data as the accessibility to data has been limited in the recent years. The dataset used is a twitter dataset which is used to portray the social media aspect of the research.

For the stock data, historical stock data was downloaded using yfinance python module from Yahoo finance.

6.4 Results

The results section in this report conveys the results of the machine learning models and how they explain the use of machine learning in the financial markets.

6.4.1 Technical analysis Results

The results are shown in the EDA_stockdata file.

6.4.2 Sentimental analysis Results

The results were promising as they outlined the usefulness of the research.

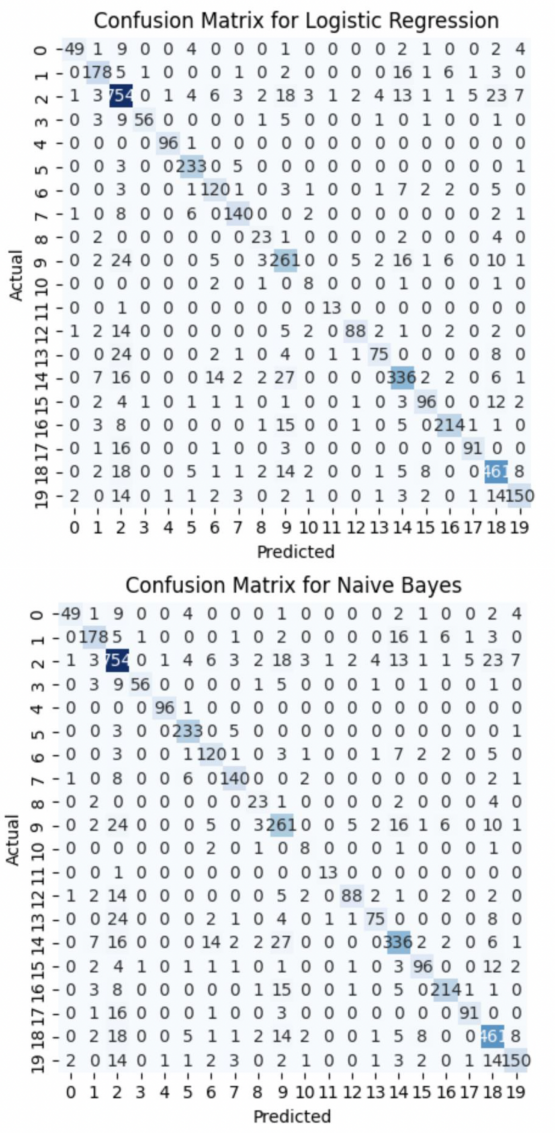

Figure 10: Confusion matrix sentimental analysis

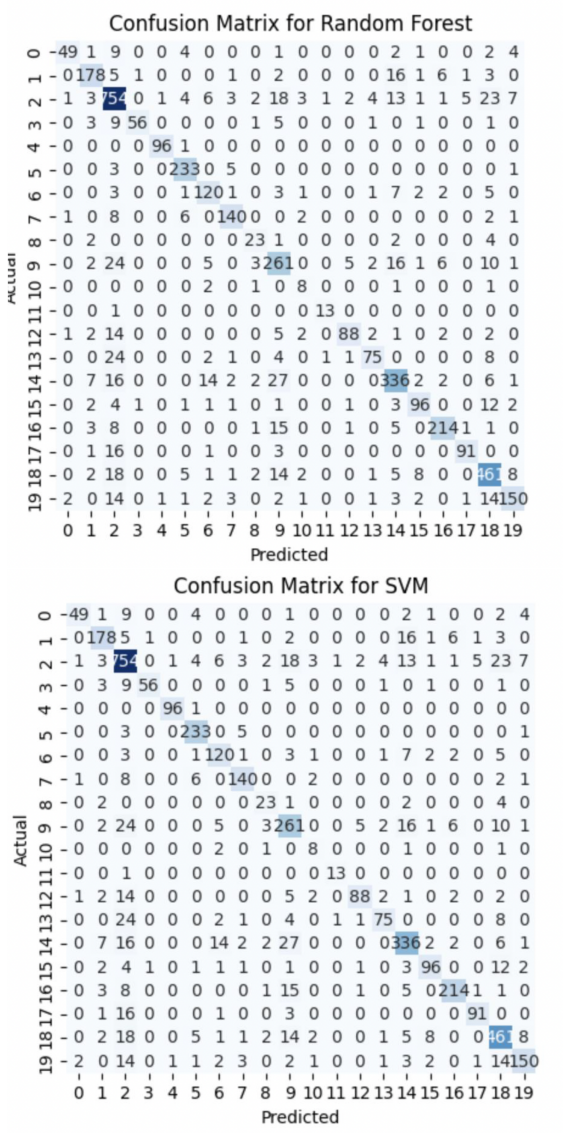

Figure 10: Confusion matrix sentimental analysis

All the confusion matrix, show how well the prediction is for the four different machine learning models. Also, it can be seen the difference between the others and the SVM model how it is and the possibility to further enhance this.

6.5 Discussion

The findings have shown that sentimental analysis can be a useful resource combined with machine learning to investigate financial markets and with 83% accuracy score on the SVM model it can be seen that moving towards non-traditional models will increase the accuracy.

In this report it is seen that social media data can be really valuable in under-

standing user sentiment however due to limitations, this could not be portrayed in the report.

The results from the technical analysis fully support the Random walk theory as there is no correlation between the actual and predicted resuts, eventhough this needs to be further developed and evaluated using techniques such as k-folds.

7 Conclusion

As seen by the evidence SVM modelling for social media provides the best predictions and it is a viable option whilst looking into understanding sentimental analysis and initialising a coding journey into research on financial markets.

Overall, time management needs to be improved as due to unforeseen circumstances the whole project was delayed by several months and indhered project development.

Moreover, objectives need to be set out clearly and understading needs to be clear since the start of the project.

8 Future work

In the future, the report can benefit from modelling on real time data, which can be used in a time series to further approve and compare the predictions to stock market graphs and carry out technical analysis along with fundamental analysis which will allow to refine the predictions even more.

Time series graphs will further enhance the understanding of the problem with visual aid. Moreover, Computer vision can be used to analyse graphs and compare different techniques of analysis with each other efficiently.

Data collection to be completed using webscraper at full completion, the webscraper file is also attached to allow insight into work done on the file. NYT API's were later researched and would be a great oppurtunity for further implementation of traditional news.

Finally, the objective to further develop this study into a publication will be pursued.

A Appendix : Design and methodologies

A.1 Literature search methodologies

A.1.1 Databases and search engines

Source	Description
Google Scholar	one of the search engines with the largest amount of literature; it will have an exhaustive amount of data therefore preliminary literature will be found on here, through a broader overview of the topics.
IEEE Explorer	It will provide technology specific literature so it will be easy to navigate for machine learning topics.
Elsevier	Mainly used for the search as it provides literature in a number of areas relevant to the paper.
Scopus	A product of Elsevier allows quick searching of database.
School libraries	Only used for project design initial ideas to prevent any copyright infringement.

A.1.2 Search terms

Search terms are allocated to each theme (Table 6) relevant to the themes semantics therefore ensuring that the literature is not irrelevant and does not present a leak in credibility. Moreover, all the search terms (Table 7) may or may not be used according to need of the search however they present parameters to finalise the list of competent literature.

Themes
Stock market
Geopolitics(news outlets)
Fundamental Analysis
Technical Analysis
Sentimental Analysis
Evaluation

Table 6: Themes for search

Stock market	Geopolitics (news outlets)	Fundamental Analysis	Technical Analysis	Sentimental Analysis
Volatility	Conflict	Return on investment (ROI)	Moving averages	Opinion mining
Equities	Investor sentiment	Industry analysis	Trading signals	Investor sentiment
SP500	Market trends	Valuation models	Momentum	Social media sentiment
Market trends	Fluctuation	Balance sheet	Volume analysis	Semantic analysis
Indexes	Sanctions	Revenue	Fibonacci	Sentiment detection
Market cap	Trade	Profit margin	Trend	Supervised/ unsupervised learning

Table 7: Key terms

B Appendix: Implementation

B.1 Technical analysis: EDA

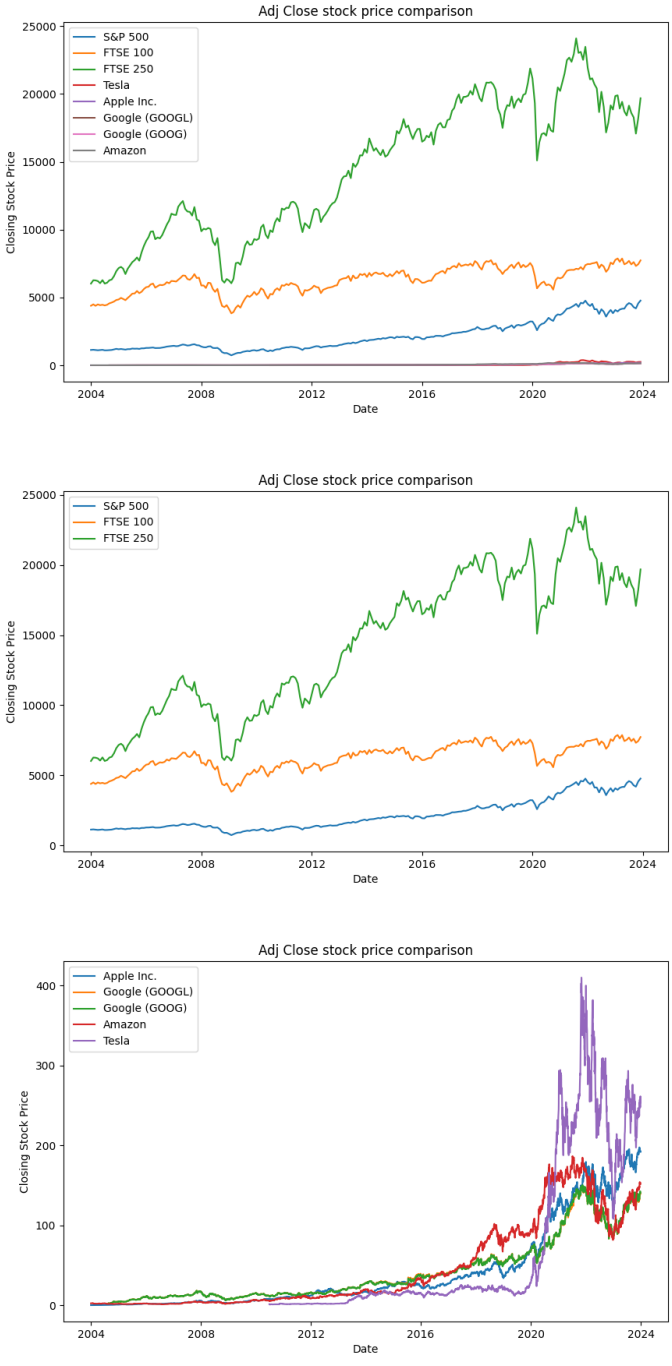

Figure 11: Stock Price Comparison Charts

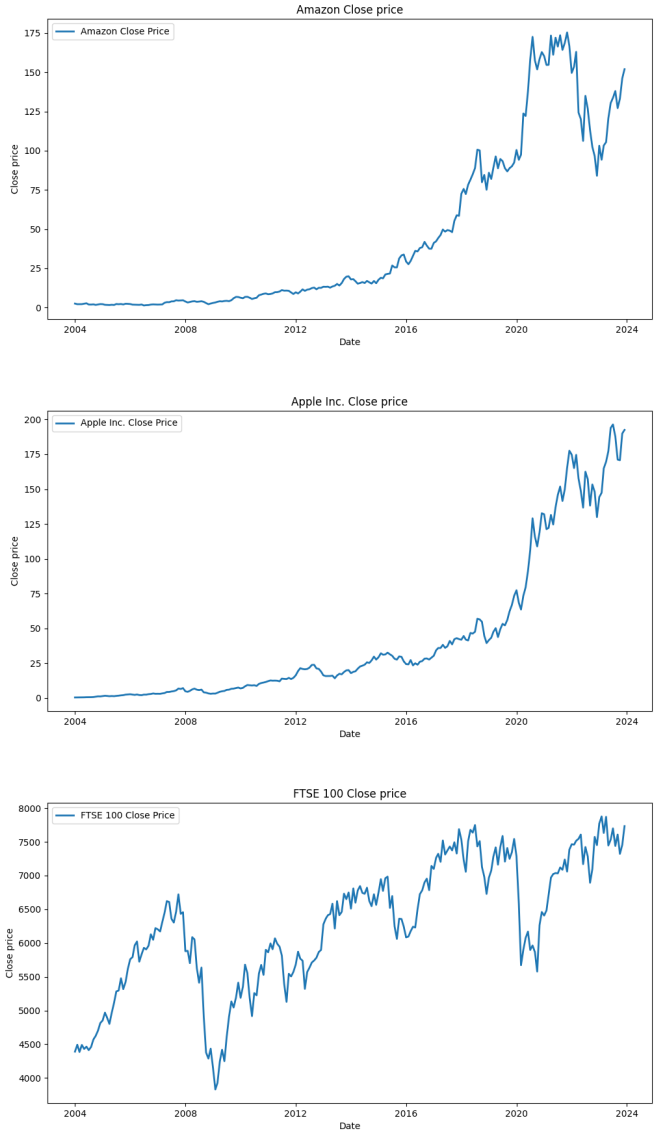

Figure 12: Company line plots

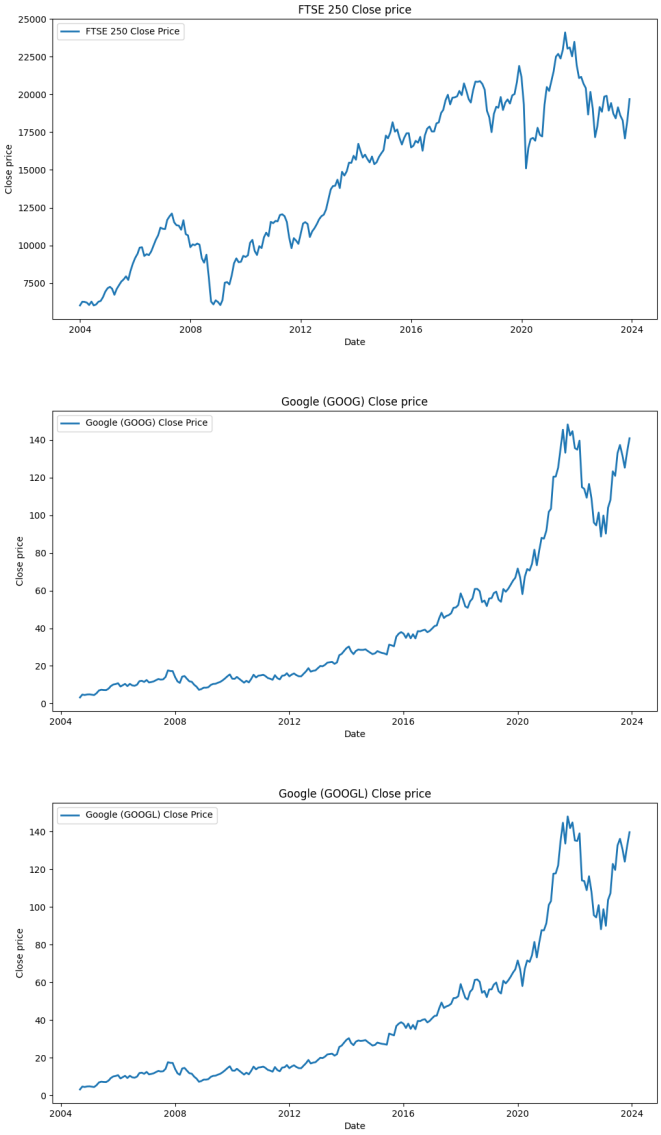

Figure 12: Company line plots

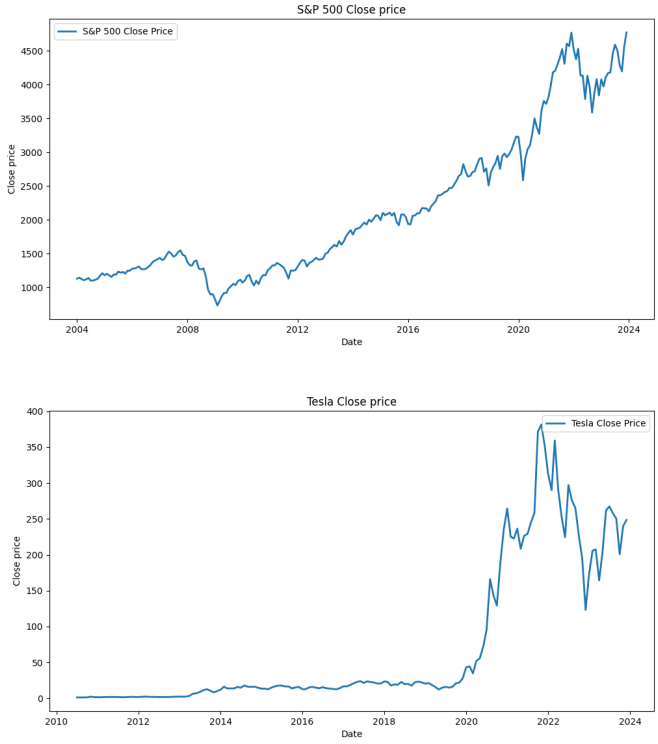

Figure 12: Company line plots

Heatmaps can be seen on the EDA_stockdata notebook along with the values which were checked.

References

Alshamrani, A. and Bahattab, A. (2015), 'A comparison between three sdlc models waterfall model, spiral model, and incremental/iterative model', *International Journal of Computer Science Issues (IJCSI)* **12**(1), 106.

Hayes, A. (2023), 'What does finance mean? its history, types, and importance explained', https://www.investopedia.com/terms/f/finance.asp. Accessed: 09 October 2023.

Olan, M. (2003), 'Unit testing: test early, test often', *Journal of Computing Sciences in Colleges* **19**(2), 319–328.

Sharma, R. (2014), 'Quantitative analysis of automation and manual testing', *International journal of engineering and innovative technology* **4**(1).

Test strategy (2023), https://en.wikipedia.org/wiki/Test_strategy. Accessed: 29 January 2024.

Zouaghia, Z., Aouina, Z. K. and Said, L. B. (2023), Stock movement prediction based on technical indicators applying hybrid machine learning models, *in* '2023 International Symposium on Networks, Computers and Communications (ISNCC)', IEEE, pp. 1–4.