REPORT

STOCK SENTIMENT ANALYSIS

REPORT BY - SAMRUDH I MOON

Enrollment No.- 22115136

INTRODUCTION

What is Stock Sentiment Analysis?

Stock sentiment analysis is the process of analysing textual data, particularly financial news articles, social media posts, or financial reports, to gauge the market sentiment towards a particular stock or the market as a whole. The sentiment can be positive, negative, or neutral and is used to predict the potential impact on stock prices. By understanding the sentiment, investors and analysts aim to make more informed decisions about buying, selling, or holding stocks.

Sentiment analysis works by using natural language processing (NLP) and machine learning techniques to identify and extract subjective information from text.

FLOW OF PROJECT

1. **Data Scraping:**
    - Collecting the dataset containing financial news headlines from different websites in the csv format generally.
    - In this case the dataset is collected from kaggle.com
2. **Data Importing:**
    - Load the dataset containing financial news headlines to Jupyter notebook
    - First save it in csv format in pc and fetch the location in Jupyter notebook.
    - The sentiment column containing negative, positive was replaced by label column in binary digits.
3. **Data Preprocessing:**

- Clean the text data by removing non-alphabetic characters using replace method.
- Converting the text data to lowercase.

4. **Splitting Data:**
   - Divide the dataset into training and testing sets based on dates.
   -
5. **Text Vectorization:**
   - Transform the headlines into numerical data using CountVectorizer and TfidfVectorizer with bigrams and trigrams.
6. **Model Training:**
   - Train machine learning models (RandomForestClassifier and MultinomialNB) on the vectorized training data.
7. **Prediction:**
   - Use the trained models to predict stock market trends on the test dataset.
   - Finding the accuracy of the prediction
8. **Evaluation:**
   - Assess model performance using metrics like confusion matrix, accuracy score, and classification report.

Data collecting

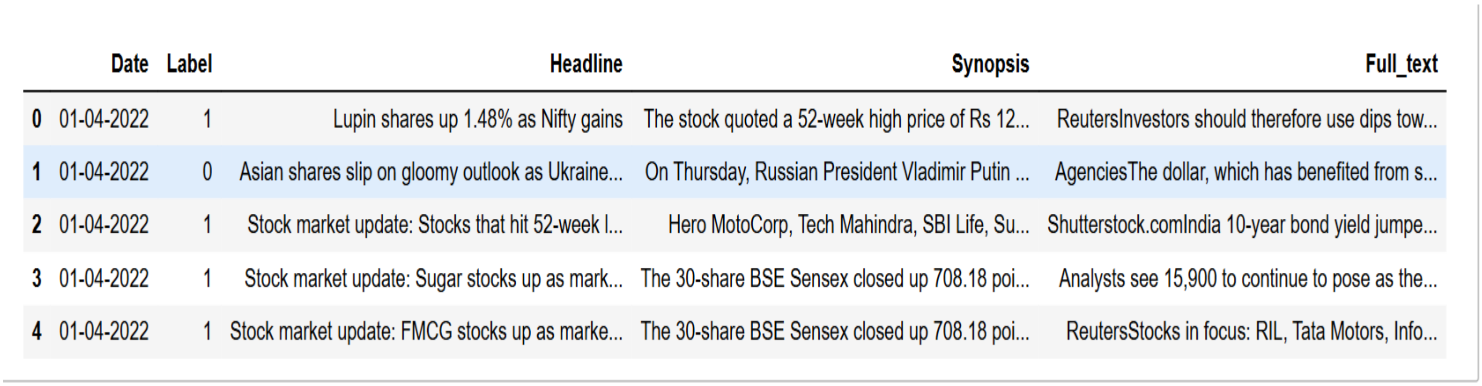
Source - Public text was collected from the source of website kaggle.com. from its datasets
Kaggle/datasets/aravsood7/sentiment-analysis-labelled-financial-news-data

Data Selection

It contains stock news headlines of Indian stock market which contains sentiments of negative, positive attached with news headlines. It contains 5 columns with 400 rows.

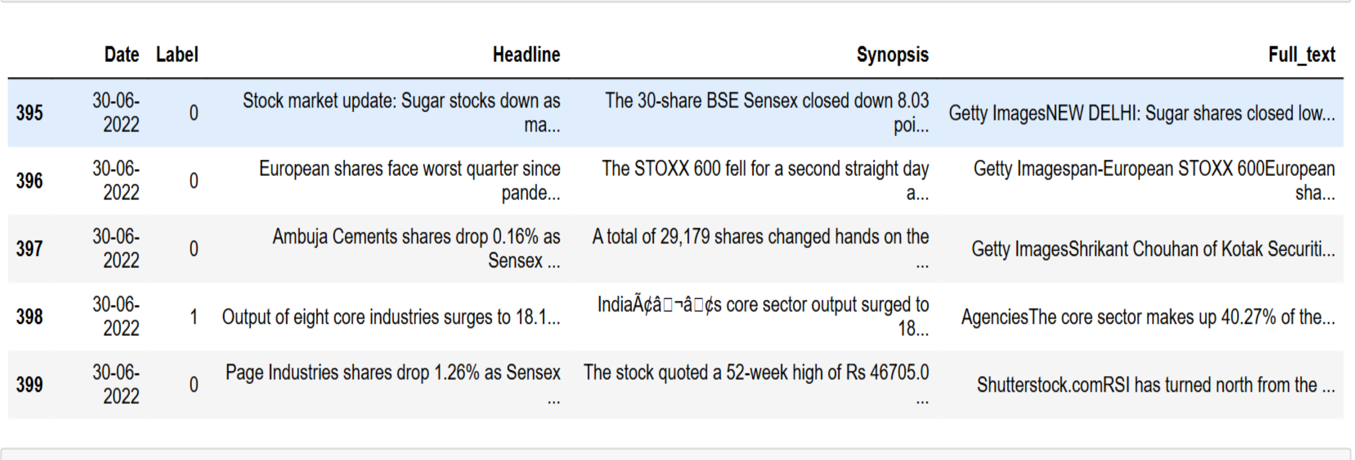
Period of Dataset- 01-04-2022 to 30-06-2022

Here is the top 5 rows of the dataset



	Date	Label	Headline	Synopsis	Full_text
0	01-04-2022	1	Lupin shares up 1.48% as Nifty gains	The stock quoted a 52-week high price of Rs 12...	ReutersInvestors should therefore use dips tow...
1	01-04-2022	0	Asian shares slip on gloomy outlook as Ukraine...	On Thursday, Russian President Vladimir Putin ...	AgenciesThe dollar, which has benefited from s...
2	01-04-2022	1	Stock market update: Stocks that hit 52-week l...	Hero MotoCorp, Tech Mahindra, SBI Life, Su...	Shutterstock.comIndia 10-year bond yield jumpe...
3	01-04-2022	1	Stock market update: Sugar stocks up as mark...	The 30-share BSE Sensex closed up 708.18 poi...	Analysts see 15,900 to continue to pose as the...
4	01-04-2022	1	Stock market update: FMCG stocks up as marke...	The 30-share BSE Sensex closed up 708.18 poi...	ReutersStocks in focus: RIL, Tata Motors, Info...

Here is the bottom 5 rows of dataset



	Date	Label	Headline	Synopsis	Full_text
395	30-06-2022	0	Stock market update: Sugar stocks down as ma...	The 30-share BSE Sensex closed down 8.03 poi...	Getty ImagesNEW DELHI: Sugar shares closed low...
396	30-06-2022	0	European shares face worst quarter since pande...	The STOXX 600 fell for a second straight day a...	Getty Imagespan-European STOXX 600European sha...
397	30-06-2022	0	Ambuja Cements shares drop 0.16% as Sensex ...	A total of 29,179 shares changed hands on the ...	Getty ImagesShrikant Chouhan of Kotak Securiti...
398	30-06-2022	1	Output of eight core industries surges to 18.1...	IndiaÃ¢â□¬â□¢s core sector output surged to 18...	AgenciesThe core sector makes up 40.27% of the...
399	30-06-2022	0	Page Industries shares drop 1.26% as Sensex ...	The stock quoted a 52-week high of Rs 46705.0 ...	Shutterstock.comRSI has turned north from the ...

Data Splitting

- Train dataset- 01/04/2022 - 16/05/2022
- Test dataset – 17/05/2022 - 30/06/2022

(180/220)

N-gram

- An n-gram of size 1 is referred to as a “unigram”, size 2 is a “bigram”, and size 3 is a “trigram”. When *n* is larger than 3, it’s usually referred to by the numerical value (e.g., “4-gram”).
- N-grams are used to model the language for tasks like text prediction, spelling correction, and sentiment analysis. They capture the context of words by considering the sequence in which they appear.
-

TfidfVectorizer:

- The `TfidfVectorizer` converts a collection of raw documents into a matrix of TF-IDF features. It’s equivalent to `CountVectorizer` followed by `TfidfTransformer`.
- TF-IDF stands for Term Frequency-Inverse Document Frequency. It reflects how important a word is to a document in a collection or corpus.
- The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

CountVectorizer:

- The `CountVectorizer` provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words but also to encode new documents using that vocabulary.
- It counts the number of times each word appears in the document.

Usage with N-grams:

- Both `TfidfVectorizer` and `CountVectorizer` can be configured to consider n-grams by setting the `ngram_range` parameter.

`ngram_range=(2,2)`, the vectorizer will consider both bigrams.

Predictions

- <u>With trained RandomForestClassifier model</u>

Entropy:

- Entropy is a measure of impurity or randomness in the dataset.
- In the context of decision trees, which are the building blocks of a random forest, entropy is used to calculate the information gain..
- When building a tree, you can choose ‘entropy’ as the criterion for making a split in order to maximize information gain.

n_estimators:

- The ‘n_estimators’ parameter specifies the number of trees in the forest of the model.
- Generally, a higher number of trees increases performance and makes the predictions more stable, but it also slows down the computation.

Visualization of Model Predictions

The confusion matrix is a powerful tool for understanding the performance of a classification model. It provides a visual representation of the actual versus predicted values, allowing us to quickly assess the number of correct and incorrect predictions.

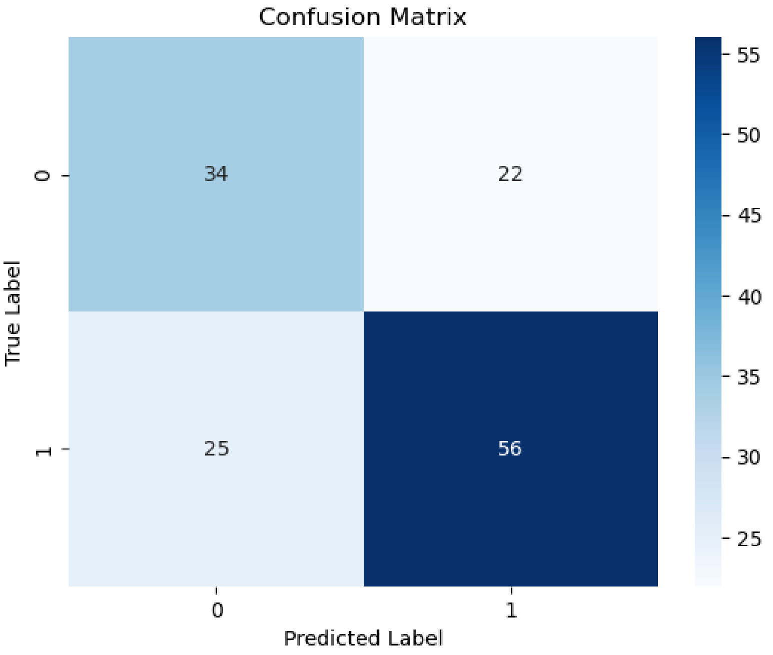
In our heatmap visualization, each cell represents the number of observations for the actual (rows) versus predicted (columns) classes. The diagonal cells (from top left to bottom right) show the number of correct predictions for each class, which are ideally higher than the off-diagonal cells that represent misclassifications.

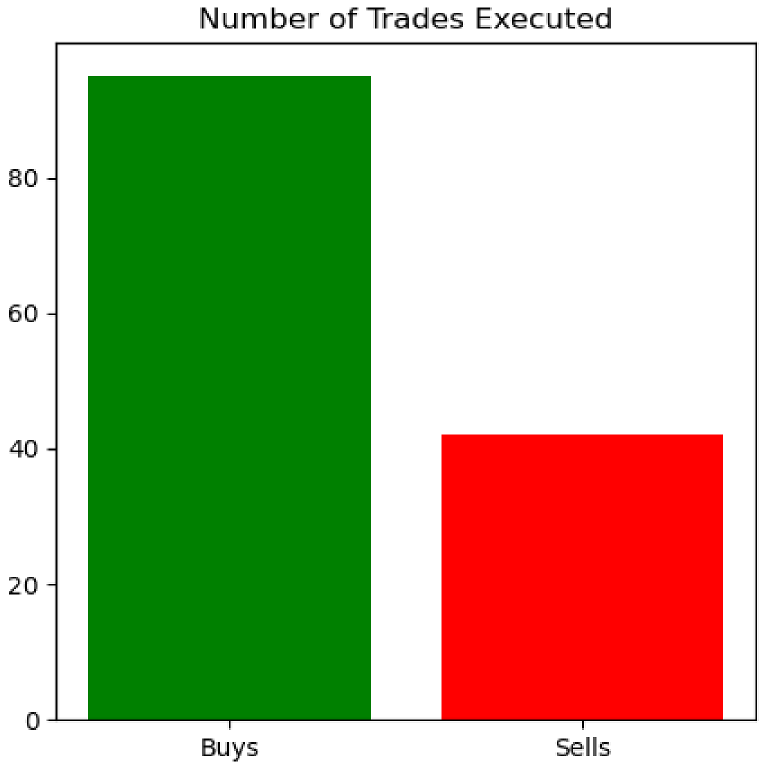
For instance, if we have two classes, 0 and 1, a high value in the top left cell indicates that class 0 is mostly predicted correctly, while a high value in the bottom right cell indicates the same for class 1. Conversely, high values in the top right or bottom left cells would indicate a higher rate of misclassification between the classes.

Using TfidfVectorizer -

	PRECISION	**RECALL**	**F1-SCORE**	**SUPPORT**
CLASS 0	0.58	0.61	0.59	56

CLASS 1	0.72	0.69	0.70	81
ACCURACY			0.66	137
MACRO AVG.	0.65	0.65	0.65	137
WEIGHTED AVG.	0.66	0.65	0.66	137



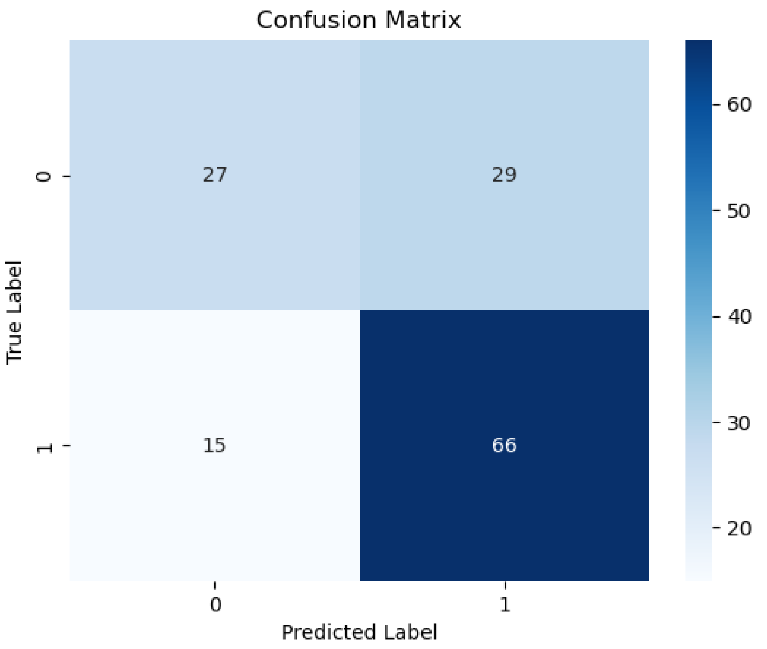


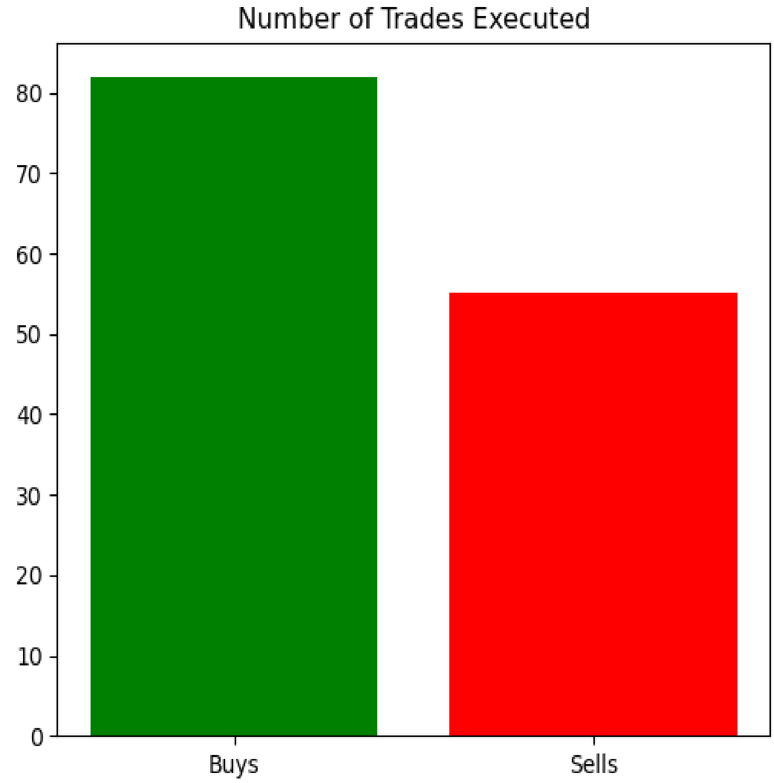
The no. of Trades executed are based on the prediction of binary Labels whether it has value ‘1’ or ‘0’.

Using Countvectorizer

	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.64	0.50	0.56	56

CLASS 1	0.70	0.80	0.69	81
ACCURACY			0.68	137
MACRO AVG.	0.67	0.65	0.65	137
WEIGHTED AVG.	0.67	0.68	0.67	137

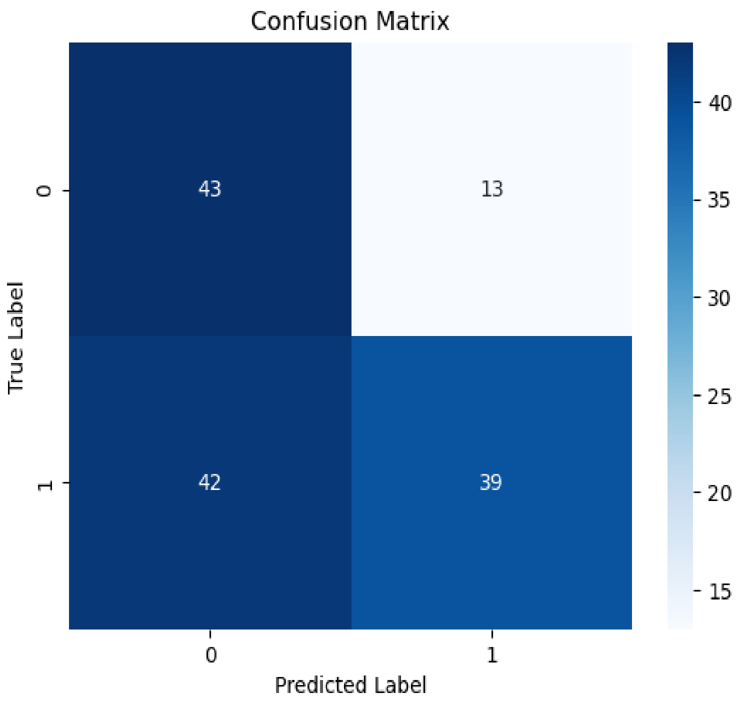


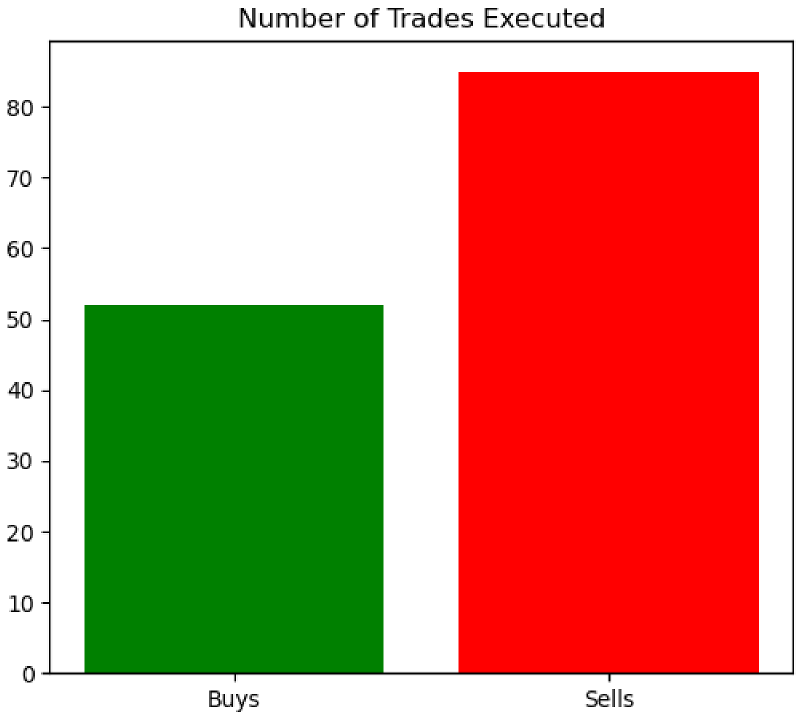


- With Trained MultinomialNB Model

	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.51	0.77	0.61	56

CLASS 1	0.75	0.48	0.59	81
ACCURACY			0.60	137
MACRO AVG.	0.63	0.62	0.60	137
WEIGHTED AVG.	0.65	0.60	0.60	137





RESULTS

- A confusion matrix that shows the number of true positives, true negatives, false positives, and false negatives.
- An accuracy score that reflects the percentage of correct predictions.

- A classification report that includes precision, recall, and F1-score for each class.
- Some classifiers can provide the probability of each class prediction, which can be useful to understand the confidence level of the model's predictions.
- With tree-based models like RandomForest, you can obtain the importance of each feature (in this case, n-grams) used in making predictions.
- These are performance measurement for classification problems at various threshold settings. ROC is a probability curve, and AUC represents the degree or measure of separability.
- This is another useful tool when evaluating binary classifiers, especially when classes are imbalanced.
- You can create visualizations such as:
    - Bar charts showing the most important features (n-grams)
    - Line graphs plotting model accuracy over time or by parameter tuning.
    - Heatmaps of the confusion matrix to make it easier to interpret.

- If your dataset includes timestamps, you can plot sentiment predictions over time to see how sentiment changes.
- You might explore correlations between sentiment predictions and actual stock price movements.

Conclusions:

1. There may be a correlation between the sentiment of news headlines and stock market movements. Positive news could lead to an increase in stock prices, while negative news could lead to a decrease.
2. The model's accuracy in predicting stock market movements based on sentiment analysis can indicate the potential predictive power of sentiment data.
3. Different machine learning models and feature extraction techniques (like CountVectorizer and TfidfVectorizer) may have varying levels of effectiveness in analyzing sentiment and predicting stock prices.

4. Significant events (like product launches, earnings reports, or economic changes) can have a noticeable impact on sentiment and subsequently on stock prices.
5. The market may already reflect the information contained in news headlines, suggesting a level of efficiency in how quickly information is incorporated into stock prices.
6. Sentiment analysis could be used as part of an investment strategy, but it should be combined with other forms of analysis to make informed decisions.
7. There may be opportunities for future research to improve the model's accuracy, such as

Recommendations:

- Be aware of legal implications when scraping data from websites and ensure compliance with terms of service and copyright laws.
- Experiment with different n-gram ranges and classifiers to improve model accuracy.
- Consider using additional features such as sentiment scores or named entity recognition for more nuanced analysis.
- Regularly update the dataset with recent news articles for timely predictions.
- Ensure your dataset is clean, relevant, and has a good mix of positive and negative samples. The quality of your data is crucial for training accurate models.
- Incorporate historical stock price data to correlate sentiment analysis results with actual market performance.
- Incorporate historical stock price data to correlate sentiment analysis results with actual market performance.
- Consider setting up a system for continuous learning where the model can learn from new data over time to stay current with market sentiments.
- Explore advanced text preprocessing techniques like stemming, lemmatization, and stop word removal to refine your features.

References

1. **Books:**

- "Text Mining: Applications and Theory" by Michael W. Berry and Jacob Kogan.

- “Mining the Web: Discovering Knowledge from Hypertext Data” by Soumen Chakrabarti.

2. **Online Courses and Tutorials:**

- Sentiment Analysis courses on platforms like Coursera, Udemy, or edX.
- Tutorials on Natural Language Processing (NLP) with Python on websites like Real Python or DataCamp.

Online videos:

https://youtu.be/g3jC8SDjRLA?si=VkoU30guuZJNCypK

https://youtu.be/h-LGjJ_oANs?si=vgv-TpAc6mOImqQQ

1. **Financial Data Sources:**

- Yahoo Finance or Google Finance for historical stock price data.
- Financial news websites for sentiment analysis.