QuantForecast: Final Assignment
Forecasting Stock Prices Based On News Headlines

Objective: The objective of this assignment is to implement a sentiment analysis model using Natural Language Processing (NLP) techniques. You will build a model that analyzes news headlines to predict whether the stock price of a company will go up or down.

Context: Sentiment analysis, or opinion mining, is a subfield of NLP that focuses on identifying and categorizing opinions expressed in text. It is widely used in applications such as market analysis, customer service, and social media monitoring. For instance, businesses analyze customer reviews to gauge public sentiment about their products, while financial analysts study news headlines to predict stock market movements.

In this assignment, you will work with a dataset containing the top 25 news headlines of a company and a binary variable indicating stock price movement (0 for a price decrease and 1 for a price increase). You will preprocess the text data, divide the dataset into training and testing sets, build a sentiment analysis model, and use it to predict stock price movements based on the headlines.

Deliverables:

1. **Code**: Implement the NLP models and provide a well-documented script or Jupyter notebook.
2. **Report**: A detailed report discussing the methodology, analysis, results, and answers to the provided questions.

Models to Use:

1. **Naive Bayes Classifier with TF-IDF Vectorizer**
2. **Random Forest Classifier**
3. **Naive Bayes Classifier for Classification**

Text Vectorization Methods:

1. **Bag of Words Model**
2. **TF-IDF Vectorizer**

Questions to Answer:

1. **What preprocessing techniques did you apply to the text data, and why?**
    - Discuss in detail the various preprocessing steps you took, such as tokenization, stemming, lemmatization, removal of stop words, handling punctuation, and any others. Explain the rationale behind each step and its impact on the data.
2. **Which model did you choose for sentiment analysis, and what are its advantages?**
    - Justify your choice of model(s) (Naive Bayes Classifier with TF-IDF Vectorizer, Random Forest Classifier, or Naive Bayes Classifier for

    Classification). Discuss the advantages and limitations of each model in the context of this problem.

3. **What metrics did you use to evaluate the model's performance, and why?**
    - List the performance metrics used (e.g., accuracy, precision, recall, F1-score, AUC-ROC) and provide a detailed explanation of why each metric is relevant for this classification problem.
4. **Based on the model's predictions, what insights can you draw about the relationship between news headlines and stock price movements?**
    - Analyze the model's predictions and provide insights into how news headlines correlate with stock price movements. Discuss any patterns or trends observed and their potential implications for stock market prediction.

Constraints:

- Use only the provided dataset for model training and testing.
- Implement both Bag of Words Model and TF-IDF Vectorizer for text vectorization.
- Use at least two different models from the specified list for classification.
- Document all code thoroughly and provide clear explanations in the report.

Submission Guidelines:

- Submit the code as a Python script or Jupyter notebook.
- The report should be in PDF format, not exceeding 10 pages, including figures and tables.
- Ensure that your code runs without errors and produces the expected results.

Dataset: You are provided with a CSV file containing the following columns:

- `Headlines`: Headlines.csv

Deadline: 25-07-2024

Submit your code and report by the due date. Make sure to follow best practices in coding and documentation.