Project Report

Stock Price Prediction Using Sentiment Analysis and NLP in Python

---

1. OBJECTIVE

This project proposes a predictive model for stock price movement, leveraging NLP and sentiment analysis. The project's objectives are to analyze the sentiments found in news articles, as well as other textual sources, and their relationship to the trend of stock prices. Two machine learning models, which are Logistic Regression and Random Forest, have been used to compare the accuracy and precision of the two models in predicting stock sentiment. It has been designed to help investors with data-driven insights that will reduce their financial risk.

2. SYSTEM ARCHITECTURE AND DESIGN

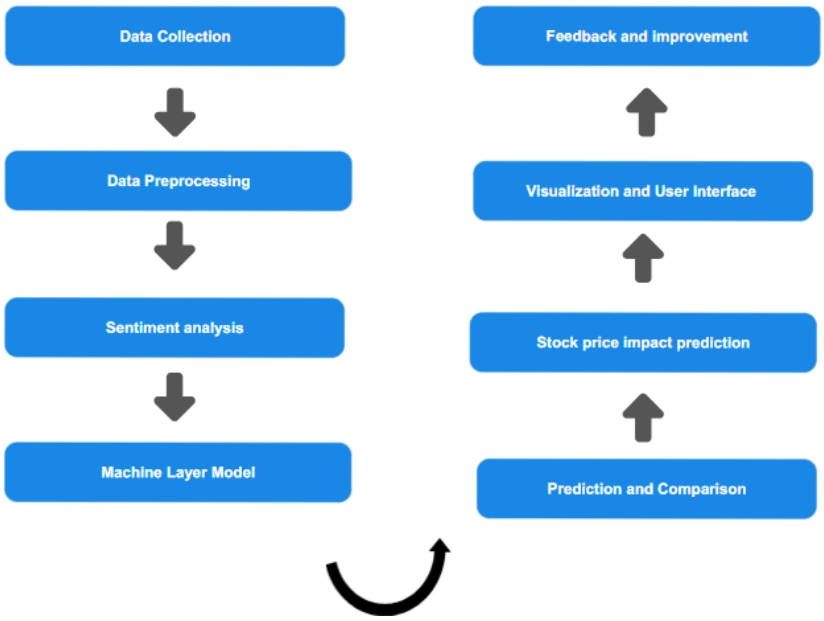
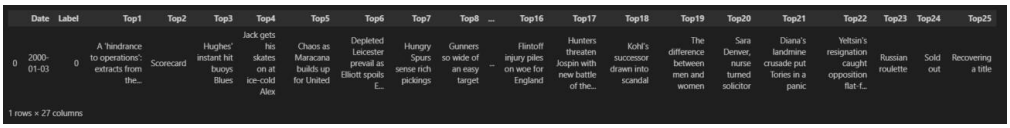


Fig 1: System Flow

The dataset employed has over 6,086 entries where each entry comprises a date, top 25 news headlines of the day and a label that denotes the impact on the market (0-negative and 1-positive).

Dataset Visualization:



	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...	Top16	Top17	Top18	Top19	Top20	Top21	Top22	Top23	Top24	Top25
0	2000-01-03	0	A 'hindrance to operations': extracts from the...	Scorecard	Hughes' instant hit buoys Blues	Jack gets his skates on at ice-cold Alex	Chaos as Maracana builds up for United	Depleted Leicester prevail as Elliott spoils E...	Hungry Spurs sense rich pickings	Gunners so wide of an easy target	...	Flintoff injury piles on woe for England	Hunters threaten Jospin with new battle of the...	Kohl's successor drawn into scandal	The difference between men and women	Sara Denver, nurse turned solicitor	Diana's landmine crusade put Tories in a panic	Yeltsin's resignation caught opposition flat-f...	Russian roulette	Sold out	Recovering a title

1 rows × 27 columns

Fig 2: Dataset rows

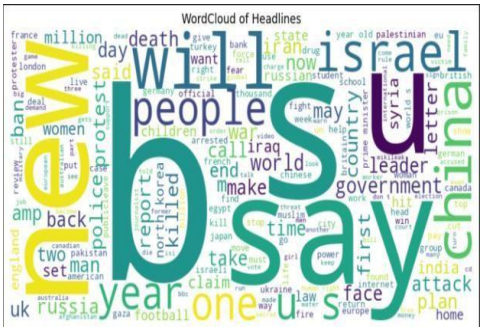


Fig 3: Word cloud

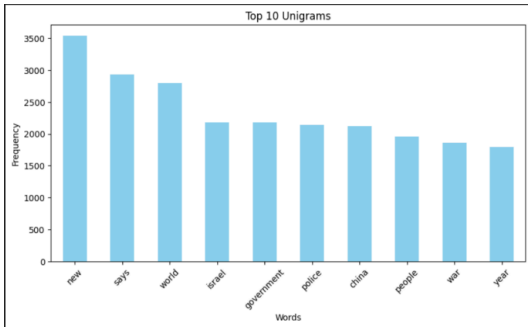
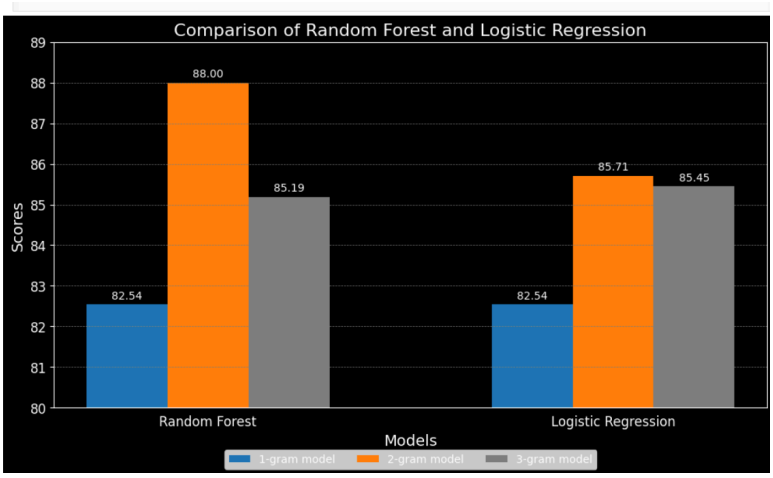


Fig 4: Histogram of most frequent words

3.Key Takeaway

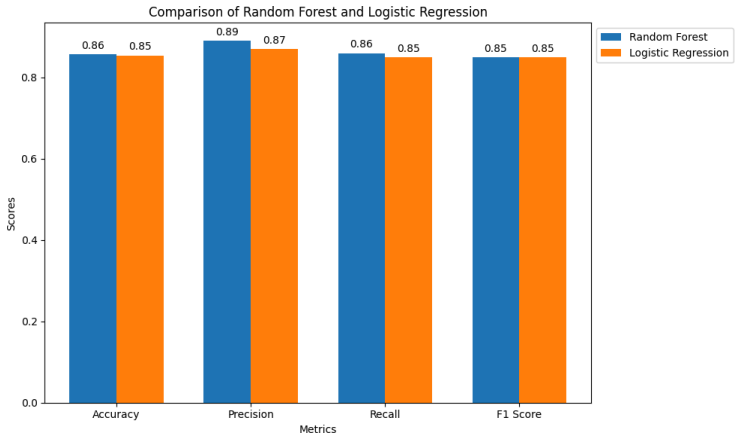
1. Best Performing n-gram:

   For both models, the 2-gram configuration yielded the highest performance, with Random Forest achieving the highest accuracy at 88% and Logistic Regression following closely at 85.71%. This suggests that the 2-gram model effectively captures sentiment-related context, making it a suitable choice for sentiment analysis in stock price prediction.



2. Best Performing Model:

   Among the two models used (Random Forest and Logistic Regression) Random Forest has better Results with the accuracy of 88% while Logistic Regression had an accuracy of 85.71%.



3.Class-Specific Performance:

   Both models had a high precision and recall values across classes in the 2-gram configuration, with Random Forest showing better recall for Class 1 (Positive sentiment), indicating its suitability for applications focused on identifying positive sentiment trends.