Title: Woodstock: Using Twitter Sentiments to Predict Stock Price Change

Team Composition : Joel Paul(G23AI1017), Anmol Raj(G23AI1002), Soumalya Mondal(G23AI1042)

Introduction

The Woodstock project aims to leverage sentiment analysis of Twitter tweets to predict changes in stock prices. By analyzing the sentiment of tweets related to specific companies or the stock market in general, we can identify patterns and correlations between public sentiment and stock price movements. This application will utilize big data technologies to process and analyze large volumes of tweets in real-time, providing investors and financial analysts with valuable insights to inform their trading strategies.

Key Technology Challenges

Data Collection: Gathering large volumes of tweets in past and in real-time (using the Twitter API and Tweepy) and filtering relevant data.

Sentiment Analysis: Accurately analyzing the sentiment of tweets using natural language processing (NLP) techniques.

Data Integration: Combining tweet sentiment data with historical stock price data.

Predictive Modeling: Developing and fine-tuning machine learning models to predict stock price changes based on sentiment analysis.

Scalability: Ensuring the system can handle the high volume of tweets and stock data.

Technology Stacks

Data Collection: Twitter API, Tweepy

Data Processing: Python (Pandas, NumPy)

Sentiment Analysis: Natural Language Toolkit (NLTK), TextBlob,

Hugging Face Transformers

Database: Amazon RDS or Google BigQuery

Machine Learning: Scikit-learn, TensorFlow, Keras

Visualization: Matplotlib, Seaborn

Deployment: MLFlow on AWS

Version Control: Git, GitHub

Deliverables

Application Demonstration: A live demonstration of the sentiment analysis application, showcasing how tweet sentiments correlate with stock price changes and providing real-time predictions.

Benchmarking Study and Comparative Analysis: A comprehensive report comparing the performance of different sentiment analysis and predictive modeling techniques, highlighting their accuracy and effectiveness.

Codebase: Comprehensive code repository with documentation for reproducibility and further development.