MACHINE LEARNING PROJECT WRITE UP
TOPIC 59 : Predicting market volatility and building short term trading strategies using data from Reddit's WallStreetBets

Team members:
Harshini Dharniraj PES1UG23CS238
Hemanth P PES1UG23CS244

Problem Statement

This project investigates whether sentiment from Reddit's *WallStreetBets* subreddit can be used to predict **next-day stock price direction**, using **opening price movement** as the target. Instead of financial indicators, the goal was to test whether **social sentiment alone** contains predictive power for **high-volatility meme stocks**.

Approach

Over **53,000 Reddit posts** were collected, with **18,042 containing stock tickers**. For each stock and each day, posts were aggregated to compute:

- **Average sentiment score**
- **Average post score (upvotes)**
- **Average comment count**
- **Average post length**

Although data existed for multiple tickers, **only GME, AMC, and TSLA** were modeled to match the referenced research paper.

Implementation

Each trading day was labeled **Up (1)** if the **next day's opening price** was higher than the current day's, otherwise **Down (0)**. Three models were compared:

- **Random Baseline** (50/50 up/down guess)
- **Logistic Regression**
- **Neural Network (Dense Layers)**

Logistic Regression generally outperformed both the Neural Network and Random Baseline, except for GME where random guessing performed unusually well likely due to extreme volatility.

Conclusions & Challenges

Reddit sentiment shows **predictive value for short-term stock direction**, especially for AMC and TSLA. However:

- **Sarcasm and slang** in Reddit posts reduce sentiment accuracy
- **Neural Networks overfit easily due to limited data**
- **Binary price movement is a weak proxy for true volatility**