r/WallStreetBets: Predicting market volatility and building short term trading strategies using data from Reddit's WallStreetBets

Using NLP and Machine Learning to Analyze Retail Investor Behavior

Presented by: Harshini Dharniraj PES1UG23CS238 & Hemanth P PES1UG23CS244**Course / Instructor:** *Surabhi Narayan*

r/Objective: To determine whether **social media sentiment** from Reddit's *WallStreetBets* can be used to **predict next-day stock price direction**.

Why this is important:

Traditional indicators often **fail to capture retail-driven volatility**

Meme stocks like **GME, AMC, TSLA** move strongly based on **public sentiment rather than fundamentals**

Can sentiment be used as a **trading signal**?

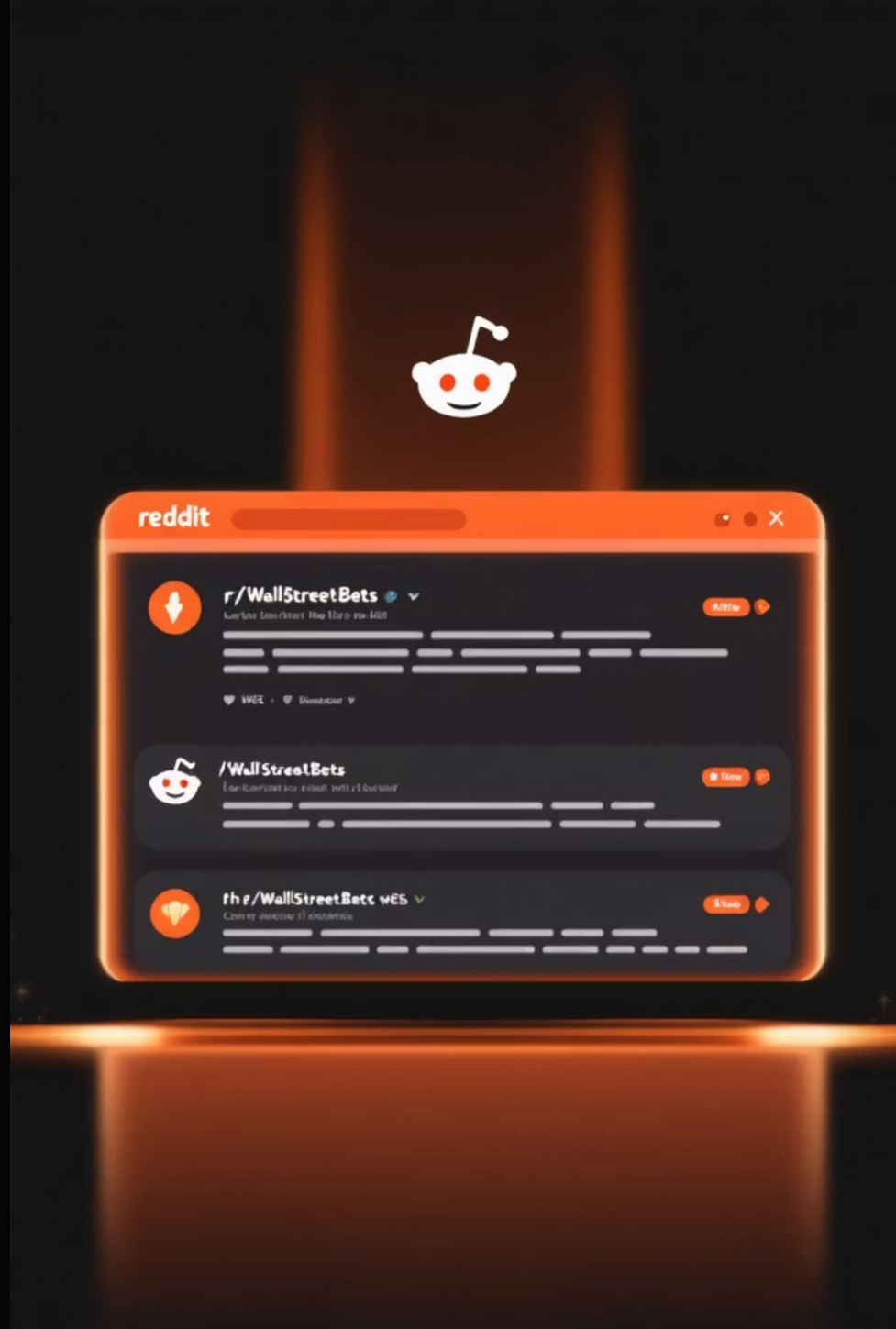
Prediction Target:Up (1) if **next-day opening price > current day's opening price**, otherwise **Down (0)**

r/Data Collection _Feature Engineering

Dataset Source:• 53,187 posts scraped from Reddit's *WallStreetBets*• 18,042 posts contained valid stock tickers

Tickers Detected (Top Mentions):• GME — 13,584 posts• AMC — 5,336 posts• TSLA — 602 posts*(Others parsed but excluded from final training due to limited consistency)*

Target Variable:• **1 = Price Up** → Next-day opening price > current opening price• **0 = Price Down**



r/Dataset_Overview

Total Reddit posts scraped: 53,187

Posts containing valid stock tickers: 18,042

Top Mentioned Stocks:

Stock	No. of Posts
GME	13,584
AMC	5,336
TSLA	602

(Additional tickers were collected but excluded from modeling due to inconsistency.)

r/Methodology

Scrape Reddit posts

Gathering raw Reddit posts from the platform to form the initial dataset for analysis.

Filter for stock tickers

Identifying and extracting only those posts that contain valid stock tickers, ensuring data relevance for financial analysis.

Perform sentiment analysis

Applying natural language processing to gauge the sentiment (positive, negative, neutral) expressed in the filtered Reddit posts.

Aggregate daily metrics

Calculating daily mean sentiment, total post scores, comment counts, and post lengths for each stock ticker.

Merge with stock price data

Combining the aggregated Reddit data with historical stock price data for comprehensive market context.

Label Up/Down movement

Categorizing stock movements as 'Up' or 'Down' based on the next-day's opening price change, for model training.

Train models & evaluate

Developing and training predictive models using the prepared data, followed by rigorous evaluation of their performance and accuracy.

r/Modeling_Approach

Model	Description
Random Baseline	Predicts Up/Down randomly (50% chance)
Logistic Regression	Linear classifier for sentiment-based movement
Neural Network	Dense layers for nonlinear feature learning

Training Strategy:

Models trained **independently for GME, AMC, and TSLA**

Evaluated using **accuracy**

r/Results

Stock	Random	Logistic Regression	Neural Network
GME	**0.765**	0.706	0.471
AMC	0.471	**0.588**	0.471
TSLA	0.429	**0.571**	**0.571**

Key Observation:

Logistic Regression generally outperformed the Random Baseline and Neural Network

Exception: GME performed best with random guessing → indicates extremely chaotic movement

r/Challenges_&_Insights

Challenges:

Sarcasm and slang made sentiment scoring imperfect

Small dataset per stock → caused **Neural Network overfitting**

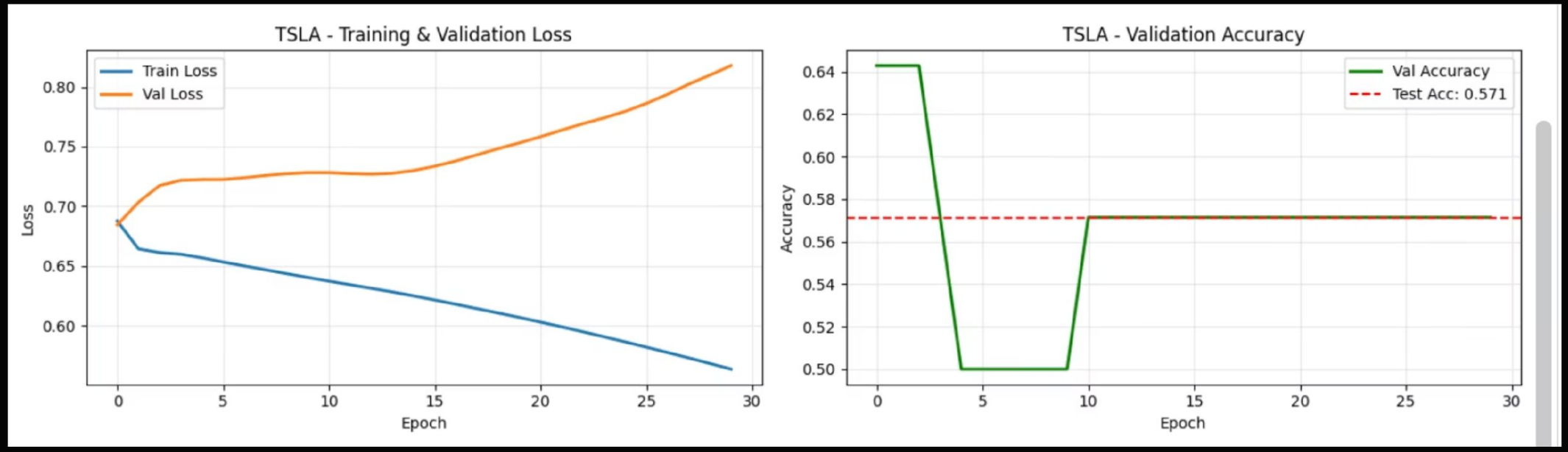
Binary direction is a **simplified version of volatility**

Insights:

Reddit sentiment contains predictive power, especially for AMC and TSLA

- **Simple models outperform deep models when data is limited**

TSLA Model Performance:



GME Model Performance:

r/Conclusion

There **is a measurable correlation** between Reddit sentiment and short-term stock movement — **WallStreetBets posts do influence price direction.**

Logistic Regression performed the most consistently, while the **Neural Network showed slight potential but struggled with limited data.**

The project confirms the **promise of sentiment-based prediction**, but also shows that **unlocking its full power requires more data and better language modeling.**

Future Directions

Expand dataset with **more Reddit sources, news articles, and financial media**

Replace generic sentiment scores with **finance-specific language models (e.g., BERT-based)**

Explore **shorter or longer prediction windows** and **real-time streaming prediction for trading bots**

Thank you

Any questions?