THE MARKET'S VIEW ON AI - ANALYSIS AND PREDICTION SYSTEM

DS598 TEAM 7

TEAM 7

CASLOW CHIEN

TEAM MEMBER
MS DATA SCIENCE ‘25

ZACH PAO

TEAM MEMBER
DATA SCIENCE ‘25

YIXIN LYU

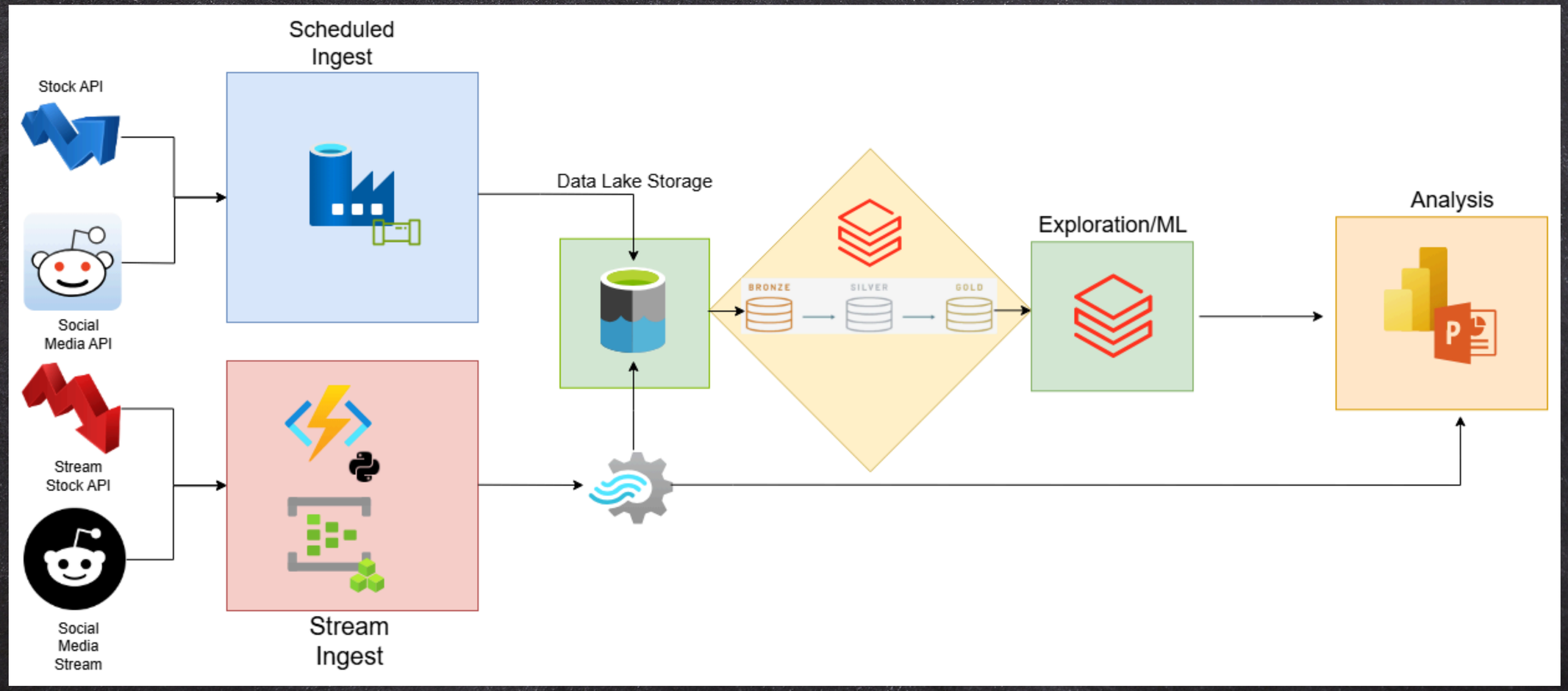
TEAM MEMBER
MS DATA SCIENCE ‘25

INTRODUCTION

Using Stock Market and Social Media data to predict how trendy “AI” is

- AI's growth reshapes markets but is hard to predict.
- Automated pipeline analyzes Stock API, Reddit API, and real-time data.
- Links stock performance, sentiment, and innovation to forecast trends.
- Offers predictive models and actionable insights.

METHODOLOGY



DATA INGESTION

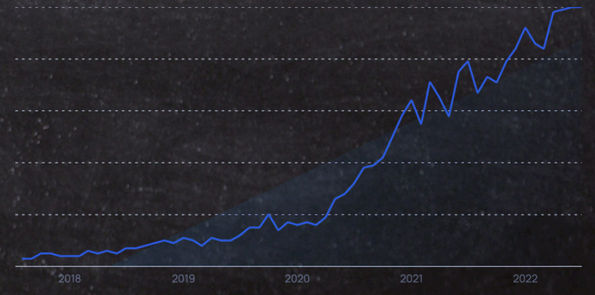
Stock Data: Alpha Vantage

- MSFT, GOOGL, NVDA, META, and AMD
- Everyday from 2021-Now

ALPHA VANTAGE

Social Media Data: Reddit

- Twitter API - Too Expensive/Little Data
- Reddit - Free!!!
  - Limits API to 50 posts/min
  - r/artificialinteligence
  - r/stocks



DATA CLEANING

For data cleaning, we used Databricks to preprocess datasets from the Stock API and Reddit API stored in our team's container. These were combined into two unified datasets. To ensure accuracy, we implemented key preprocessing steps:

Missing values were addressed through imputation (e.g., mean/mode) or removal, ensuring data integrity. Timestamps were standardized to "YYYY-MM-DD HH:MM:SS" for consistent sorting and filtering. A "week_range" column grouped data into seven-day intervals (e.g., "2021-01-04 to 2021-01-10") to simplify trend analysis.

week_range
2021-01-04 -- 2021-01-10
2021-01-04 -- 2021-01-10
2021-01-04 -- 2021-01-10
2021-01-04 -- 2021-01-10
2021-01-04 -- 2021-01-10
2021-01-11 -- 2021-01-17
2021-01-11 -- 2021-01-17
2021-01-11 -- 2021-01-17
2021-01-11 -- 2021-01-17
2021-01-11 -- 2021-01-17

MODEL - FEATURE SELECTION

Aggregate data by average value for each week (Monday to Sunday).

- Diff_percent: Percentage change in stock's closing value compared to the previous row.

$$\text{Diff\_percent} = \frac{\text{Close}_{\text{current}} - \text{Close}_{\text{previous}}}{\text{Close}_{\text{previous}}} \times 100$$

- Volume: Average stock trading volume for the week.
- AI_previous_x: AI trend index from x week ago (x from 1 to 3)

MODEL - STRUCTURE

LSTM + Fully Connected layer

Why?
Sequential Memory, Non-Linear Relationship

Additional Features
- Optimizer
- Early Stopping
- Learning Rate Scheduler

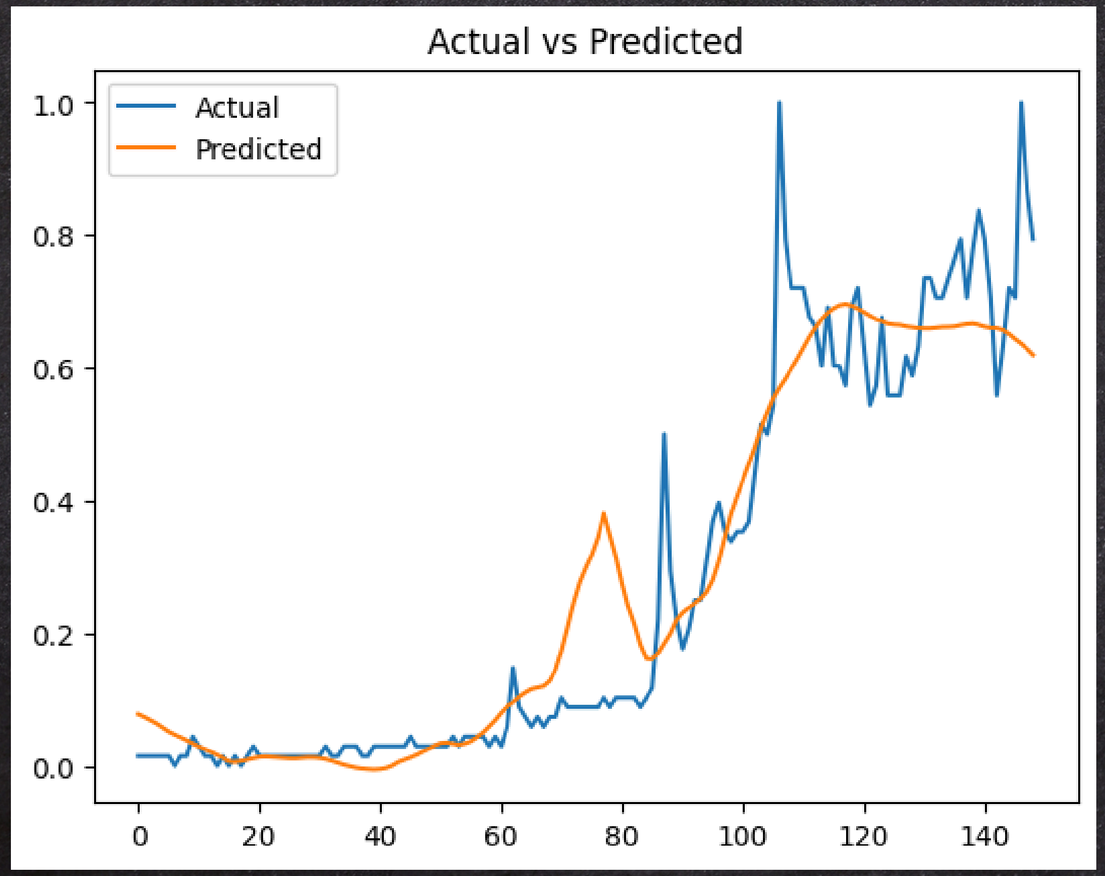
```
window_size = 10
batch_size = 2
epochs = 100
early_stopping_patience = 20
learning_rate = 0.0001
dropout = 0.4
```

Params

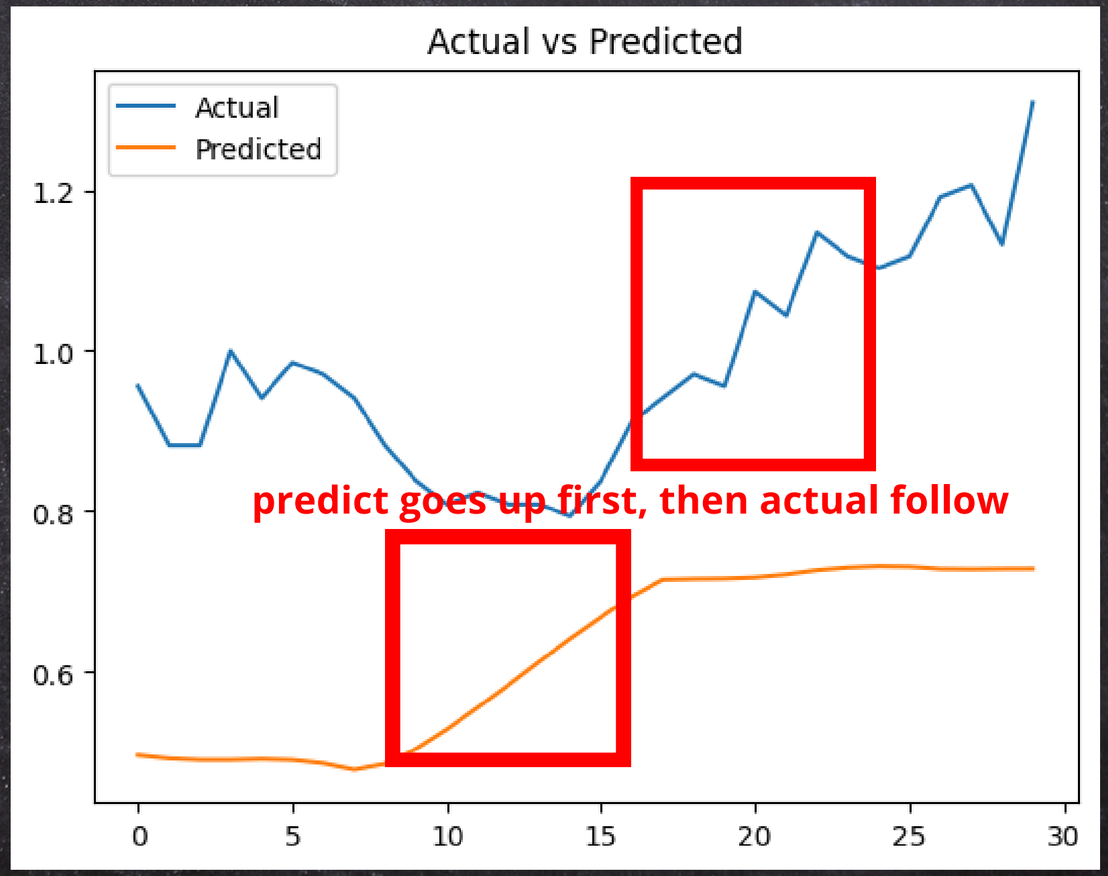
```
class LSTMNet(nn.Module):
    def __init__(self, input_dim, hidden_dim, num_layers, output_dim, dropout=0.3):
        super(LSTMNet, self).__init__()
        self.lstm = nn.LSTM(input_dim, hidden_dim, num_layers, batch_first=True,
        dropout=dropout)
        self.fc = nn.Sequential(
            nn.Linear(hidden_dim, hidden_dim // 2),
            nn.ReLU(),
            nn.BatchNorm1d(hidden_dim // 2),
            nn.Linear(hidden_dim // 2, output_dim)
        )
```

MODEL - RESULTS

Actual vs Predict (Training)



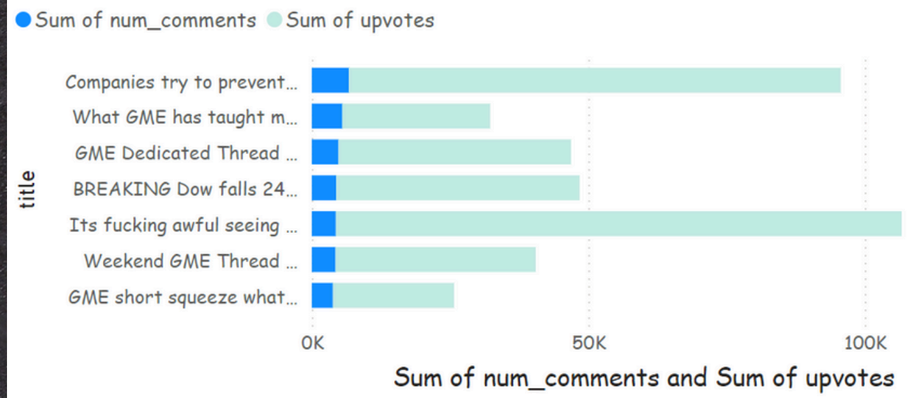
Actual vs Predict (Testing)



ANALYSIS

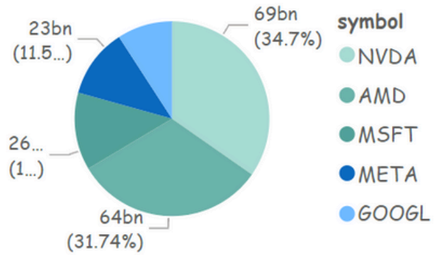
Insights into Stock Performance and Public Sentiment: Exploring Volume, Comments, and Upvotes Trends

Distribution of Comments and Upvotes by Title

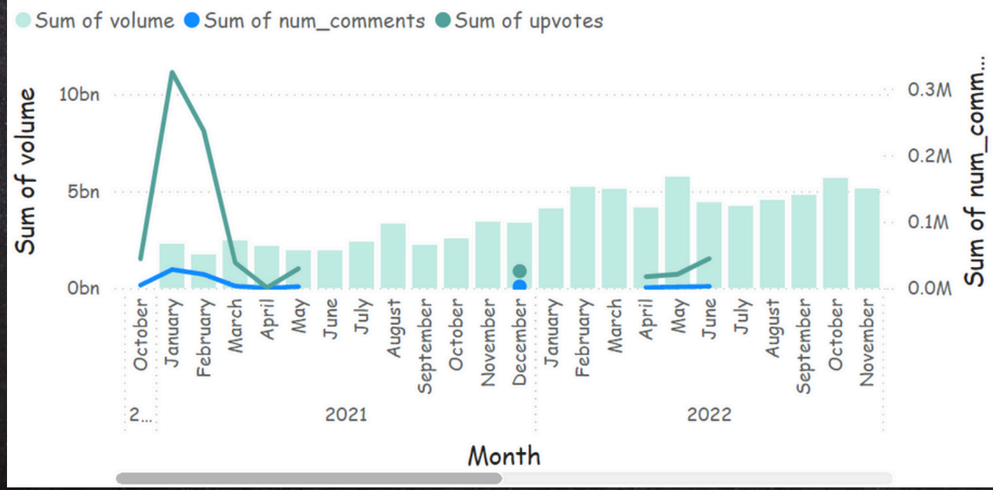


post_id	Sum of num_comments	Sum of upvotes
l6uva1	6707	88950
lazcak	5531	26723
l64xvw	4824	42054
j6b8k8	4412	44016
la34bh	4339	102335
l8rhr3	4289	36197
lbuhp0	3805	21913
l8b4pp	2769	27609
v8kp8d	2654	25719
ku2ymw	2568	18175
Total	77241	804572

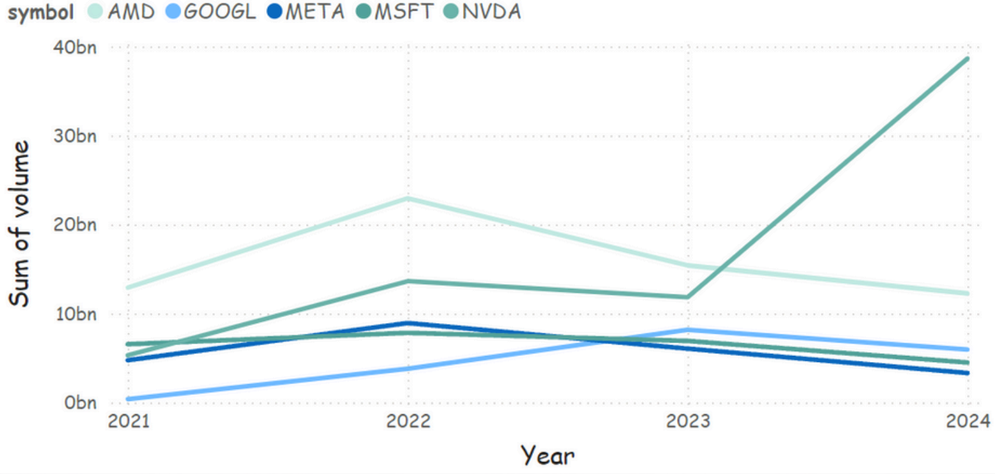
The Volume Percentage of Stocks



Distribution of Stock Volume, The Number of Comments, and The Upvotes



Different Stock Volume by Years



CONCLUSION

Our project successfully demonstrates the potential of an automated pipeline to analyze and predict the popularity of AI by linking public sentiment, stock performance, and innovation trends. Using advanced tools like Azure, Databricks, and Power BI, we created a robust system to process diverse datasets, apply predictive models, and visualize actionable insights for stakeholders.

The results highlight the importance of understanding the relationship between public discussions and market trends, showcasing how sentiment and trading activity align to influence AI-related stocks.

RECOMMENDATION & FUTURE WORK

1. **Automated scheduling system**: Periodically retrains the model when sufficient new data becomes available.
2. **Implementing Synapse**: Prepares for scalability issues if the dataset grows significantly larger.
3. **Expanding social media dataset**: Enhances model performance; explore alternative platforms due to high data acquisition costs.
4. **Enriching dataset diversity**: Improves predictive accuracy by including updates and news from leading AI companies.


THANK
YOU!