PROJECT PROPOSAL

Data Wrangling & Visualization

Nursultan Abdullaev, Elisei Smirnov, Iakov Saparov
Data Science Track, B22
Innopolis University, Spring 2025

1 Project Title

LLM Pulse: Real-Time Tracking and Sentiment Analysis of LLM Trends on Twitter.

2 Project Goal and Vision

The goal of LLM Pulse is to provide an interactive dashboard that captures, analyzes, and visualizes both real-time and historical Twitter data focused on discussions around large language models such as GPT-4, Gemini, and Claude. Through a combination of sentiment analysis, trend tracking, and keyword extraction, the application tells the story of how public discourse about these models evolves over time. Users—ranging from researchers and developers to entrepreneurs and investors—will gain key insights into the popularity, sentiment, and emerging topics associated with each LLM. The visualization aims to answer questions like: Which LLM is currently trending? How is public sentiment changing? What are the key keywords driving these discussions? The tool empowers users to make informed decisions by providing dynamic, interactive visual representations of complex social media trends.

3 Dataset Description

3.1 Data Source and Collection Plan

The primary data source for LLM Pulse will be Twitter, accessed via the Twitter API. We will utilize Tweepy to scrape tweets that mention key LLMs such as GPT-4, Gemini, and Claude in real time, as well as fetch historical tweet data for trend analysis. The scraping process will include capturing tweet text, timestamps, user metadata, and engagement metrics. To manage potential issues like website changes and API rate limits, we will implement robust error handling, automatic retries, and adhere strictly to Twitter's API usage policies. Additionally, interactive features in the visualization will include filtering options and tool-tips, particularly within the D3.js panels, to enable users to drill down into specific timeframes or data subsets.

3.2 Data Content

The dataset includes tweet text, timestamps, user metadata, engagement metrics (e.g., retweets and likes), and sentiment scores derived from tools like VADER or TextBlob. It covers a rolling one-month historical window along with continuous real-time streams, aggregating thousands of tweets daily for rich insights into geographic trends and public sentiment.

4 Visualization App Architecture

The application follows a modular pipeline. The architecture (Figure 1) ensures seamless data flow from collection to interactive visualization.

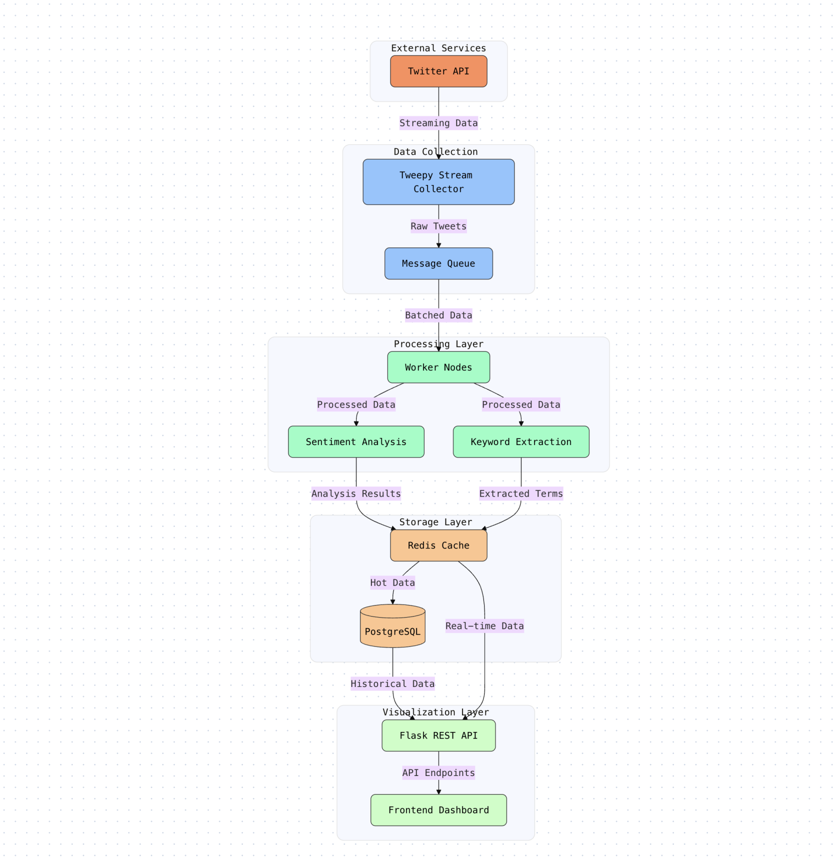

Figure 1: LLM Pulse Application Architecture

5 Proposed Visualizations and Features

The final dashboard integrates multiple interactive visualizations for comprehensive analysis of LLM discussions on Twitter. Key components include:

- **Comparison Bar Chart:** Compares mention frequency and average sentiment scores for LLMs over time intervals. Interactive features include date filtering, LLM selection, and tooltips (Figures 2 – 3).
- **Trending Models Leaderboard:** Ranks most-discussed LLMs based on real-time metrics, with sorting, expandable details, and linked filtering (Figure 4).
- **Word Cloud:** Dynamic word cloud of trending keywords with highlights and click-to-filter functionality (Figure 5).

Preliminary Python implementations (using Plotly, Matplotlib, etc.) produced the following prototypes:

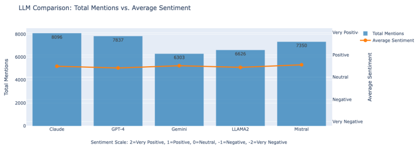

Figure 2: Comparison Bar Chart: Mentions vs. Sentiment.

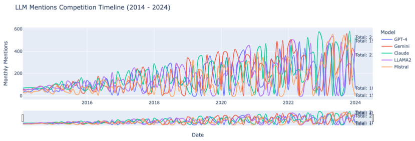

Figure 3: Mentions Timeline (2014–2024).

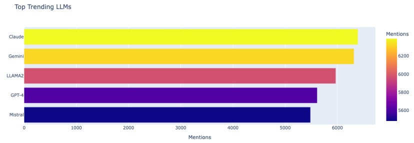

Figure 4: Trending Models Leaderboard.

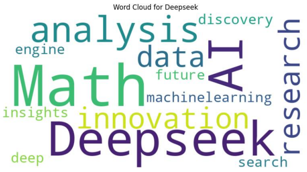

Figure 5: Dynamic Word Cloud.

6 Project Timeline and Milestones

- Week 1: Setup and Twitter API integration.
- Week 2: Develop and test the data collection module; begin data cleaning with Pandas.
- Week 3: Conduct exploratory data analysis, implement sentiment analysis, and start the Flask API.
- Weeks 4–5: Create main visualizations using D3.js with interactive filtering and tooltips.
- Weeks 6–7: Build additional visualizations, integrate cross-panel interactions, and enable real-time updates.
- Week 8: Final testing, performance optimization, and presentation preparation.