Team members:

Cesar Flores
Fernando Batarse
Raul Guillen

1. Problem Statement & Impact Analysis

The cryptocurrency market is highly volatile and largely driven by public sentiment, speculation, and news cycles. Unlike traditional financial assets, the value of cryptocurrencies such as Bitcoin and Ethereum is not tied to company fundamentals or government regulation but rather to mass perception and rapidly shifting narratives.

The core problem this project addresses is the **lack of structured and real-time insight into how public sentiment on platforms like Twitter (X), Reddit, and Yahoo Finance affects cryptocurrency prices**. Investors and analysts often rely on intuition or anecdotal trends from social media, which leads to delayed reactions or poor decision-making. By quantifying sentiment in real-time, our solution aims to enhance market transparency and support data-driven investment strategies.

Impact Potential:

- Provide actionable insights to retail and institutional investors.
- Detect early signs of market shifts due to hype, fear, or breaking news.
- Help financial analysts correlate sentiment with price volatility.
- Support regulatory and risk-analysis bodies in monitoring potential bubbles or crises triggered by mass hysteria.

---

2. Data Strategy

Data Sources:

- **Twitter (X):** Real-time tweets using hashtags and keywords like #Bitcoin, #BTC, #Ethereum, $ETH, #crypto. Filter by language (English), and geotags where available.

- **Reddit:** Posts and comments from cryptocurrency-focused subreddits like r/CryptoCurrency, r/Bitcoin, r/Ethereum.

- **Yahoo Finance News:** Article headlines and summaries related to BTC/ETH published by trusted financial outlets.

Data Collection Methods:

- Twitter API (Academic Research or v2 Standard) for tweet streams and metadata.

- Reddit API via PRAW or Pushshift for structured post/comment retrieval.

- Web scraping or RSS feeds for Yahoo Finance news headlines and summaries.

Preprocessing:

- Language filtering, spam removal, and noise reduction (e.g., URLs, emojis, boilerplate).

- Named entity recognition and coin normalization ($ETH = Ethereum).

- Time alignment of sentiment and price data to enable correlation analysis.

Sentiment Analysis Approach:

- Fine-tuned transformer-based models (e.g., RoBERTa, BERTweet) for domain-specific sentiment classification.

- Optional lexicon-based backup (e.g., VADER) for cross-validation and fallback cases.

Storage & Pipeline:

- Raw and preprocessed data stored in a cloud-hosted PostgreSQL or BigQuery database.

- Scheduled ETL jobs (Airflow or Prefect) to update sentiment scores and maintain freshness.

- Dashboard integration for visualizing sentiment trends and anomalies.