**Coding Challenge - Machine Learning Sentiment Classifier**

**Objective**

Build a minimal sentiment classifier for cryptocurrency-related Reddit comments, accessible via a basic REST API.

**Background**

We are developing an advanced sentiment analysis pipeline to analyze cryptocurrency-related discussions from social media platforms, especially Reddit. Accurate, comprehensive, and scalable sentiment analysis is critical to various downstream analytical applications.

This challenge tests your ability to build a practical NLP proof of concept, including a basic but production-ready REST API.

**Challenge Details & Requirements**

Create a minimal pipeline to train, evaluate, and deploy a sentiment classifier designed specifically for crypto-related Reddit comments. Assume the classifier will be deployed to handle approximately 10 million comments daily within a computational budget of roughly $100/month.

Sentiment analysis will support aggregate analysis and provide reliable insights even when working with smaller data segments (≤1,000 samples).

You will use a dataset containing 562 records, equally distributed between two sentiment classes (Positive and Negative). The key columns to use are:

* **Comment**: The content of Reddit comments.
* **Sentiment**: The sentiment label (Positive or Negative).

The dataset can be accessed [here](https://drive.google.com/file/d/1D5XjGR3F9QsVf6jCvAE6Ck8jWFup-qGI/view?usp=sharing).

An excerpt from the dataset is shown below:

user\_id,Comment,Sentiment,Reddit URL

XYNN2Y4VCF3G,"I bought 2200 at the ico, at 0.50$ per coin. Hold everything and sold it 3 months ago and it helped me to buy a bigger house.",Positive,https://www.reddit.com/r/Avax/comments/uzggar/comment/iabc390/?utm\_source=reddit&utm\_medium=web2x&context=3

DR6XNZMT9KRH,"Harmony one, algorand, Cardano, solana, vechain gonna fly if there is ever a next bull market. Otherwise just buy and stack satoshis",Positive,https://www.reddit.com/r/CryptoCurrency/comments/v09a1p/comment/iag2c78/?utm\_source=reddit&utm\_medium=web2x&context=3

**Key Expectations**

* Target a high accuracy level (ideally >90%).
* Include exploratory data analysis (EDA) to address any data-specific challenges (e.g., text length, variety, or imbalance). Your repository should reflect your data preparation process.
* Choose any sensible training/evaluation split or other evaluation methodology. Document and justify your evaluation approach clearly.
* Consider scalability, ensuring your chosen method can handle the target comment volume efficiently within the specified cost constraints.
* Clearly outline your methodology, model selection, and the rationale behind each decision.
* Include evaluation code (scripts or notebooks) demonstrating your performance and accuracy metrics.
* Develop a minimal REST API to serve model predictions. Feel free to omit features like authentication, rate limiting, or production-level scalability.

**Languages & Libraries**

Python is recommended, though you can use any preferred language, framework, or library that effectively demonstrates your capability to create robust machine learning solutions (e.g., PyTorch, TensorFlow, Scikit-learn, spaCy, Transformers, etc.).

**Timeframe**

We value your time and expect you to complete this challenge within **2 days**.

**Submission**

Please submit your solution via a Git repository. Include clear documentation in the repository outlining your approach, decisions, evaluation results, and clear instructions for running the solution.