Real-Time Reddit Sentiment Analysis Pipeline

DS5110: Essentials of Data Science
Iteration#02

Shaikh Basim
Mohammad Aasim Shaikh

1 Project Overview

The project launch for the ”Real-Time Reddit Sentiment Analysis Pipeline” is described in this report. Building a scalable app using Reddit Api that fetches real time data from various subreddits and displays the sentiment analysis on the dashboard. This project focus on data engineering concepts and technologies.

1.1 Project Goals

The following are the project's precise objectives and anticipated results:

- Contrust a dashboard using real time data pipeline to include sentiment visualization
- Acquire and learn how to use Apache kafka (by confluent) and MongoDB Atlas
- Develop a top-notch portfolio project that showcases a scalable architecture for streaming.
- Get practical experience with cloud infrastructure, credential management, and enterprise-grade data processing platforms.

1.2 Scope of Project

The scope has been properly established to ensure that the project can be completed within the allotted time.

- **In Scope:** Streamlit-based dashboard, Kafka streaming pipeline, sentiment analysis (using VADER and/or transformers), Reddit API interface, and MongoDB storage.

2 Deliverables and Milestones

Here below are the six main stages the project is divided into.

2.1 Major Deliverables by Phase

1. Kafka Producer script of extract Reddit Data from Reddit API Praw..
2. Setting up Confluent Cloud and configuring topics.
3. Kafka Consumer script which handles sentiment analysis and saving data to MongoDB.
4. Implementation of a MongoDB Atlas connection and data storage.
5. Data visualization dashboard most likely using steamlit
6. Final deployment, documentation, and testing.

3 Team Roles and Capabilities

Two people are working on the project as a team.

3.1 Team Capabilities Assessment

- **Current Strengths:** The team has a great grasp of data processing fundamentals, strong core capabilities in Python programming, and an ability to integrate APIs.
- **Gaps Identified:** The team members were unaware of Kafka, Confluent Cloud and MongoDB Atlas initially.
- **Acquired Skills:**Through the use of the internet the team gained knowledge of Kafka, Confluent Cloud and MongoDB Atlas

3.2 Roles and Responsibilities

The roles are divided amongst the teammates in a manner that addresses each ones capabilities.

3.2.1 Shaikh Basim

- Kafka Producer script Implementation
- User Interface Design
- Database management
- Security management

3.2.2 Mohammad Aasim Shaikh

- Integration of sentiment analysis
- Kafka Consumer script Implementation
- MongoDB Integration into dashboard
- Streamlit Dashboard Development and Data Visualization

4 Tools, Technologies, and Platforms

The project utilizes a modern, cloud-based data stack selected for its scalability and industry relevance.

4.1 Key Libraries and Frameworks

A curated list of libraries is used to build the pipeline, summarized in Table 1.

Table 1: Key Software Libraries and Frameworks

Category	Library / Framework
Data Extraction	`praw`
Data Streaming	`confluent-kafka-python`
Database	`MongoDB Atlas`
Sentiment Analysis	`vaderSentiment`, `transformers`
Dashboard	`Streamlit`
Visualization	`Plotly`, `Matplotlib`

5 Initial Setup and Configuration

5.1 Environment for Development

A virtual environment is created to avoid dependency issues that can occur between various libraries

5.2 Version Control

Version control has been successfully set up in a Git repository. A thorough.To ensure security, `gitignorefile` is used to exclude all credentials, configuration files `config.py`, and virtual environment files. The repository is accessible to every team member.

6 Data Management

6.1 Data Source

For this project, no static or pre-existing datasets are employed. By consuming real-time posts and comments from the Reddit API, the system creates its own data in real-time. The Kafka producer script allows for the customization of the source subreddits.

6.2 Data Format

MongoDB processes and stores data as JSON documents. The original postmetadata, the processed text, a timestamp, and the calculated sentiment score are all included in each document.

7 Conclusion and Next Steps

The project has finished its initial phase. The team is working on Phase 5—the creation of the Streamlit dashboard—while also aggressively constructing the basic backend pipeline (Phases 1-4). This report and the Excel progress tracker are among the necessary documents that are ready for stakeholder review.