Real-Time Reddit Sentiment Analysis Pipeline

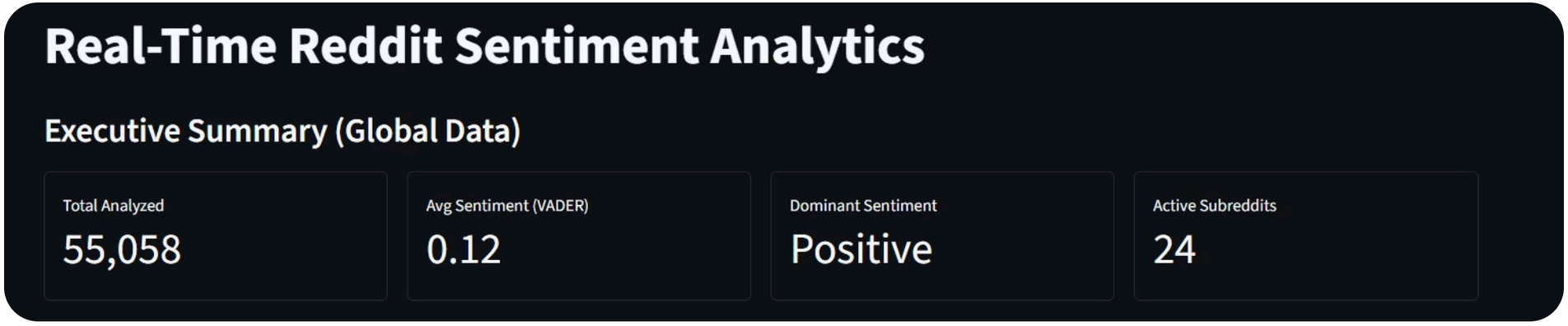
DS5110 - Essentials of Data Science

Fall 2025

Mohammad Aasim Shaikh **Shaikh Basim**

reddit-trend-analysis.streamlit.app

Project Overview: Autonomous Real-Time Social Listening



This automated pipeline ingests and analyzes Reddit discourse in real-time, transforming unstructured social data into actionable sentiment intelligence without manual intervention.

55K+

Records Processed

Posts and comments analyzed

24

Subreddits

Active communities monitored

11

Days

Continuous autonomous operation

Brand Monitoring

Track public perception instantly

Trend Detection

Identify viral topics early

Autonomous Data Collection

Zero-touch continuous ingestion

Data Flow Strategy

The architecture utilizes an event-driven model to ensure low latency and high availability. Data flows seamlessly from ingestion to visualization.

- **Source:** Reddit API
- **Broker:** Kafka (Confluent Cloud)
- **Storage:** MongoDB Atlas
- **View:** Streamlit

Automation Logic

Orchestrated fully via GitHub Actions to maintain uptime without dedicated server infrastructure.

Producer Cycle

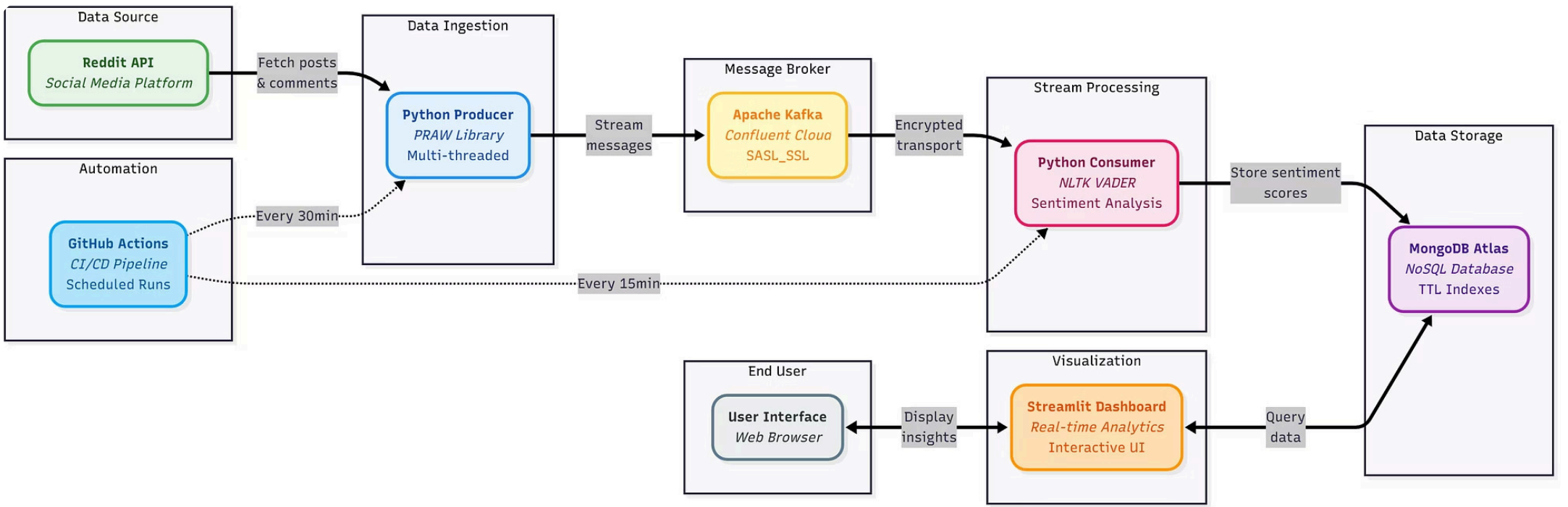
Runs every 15 minutes to fetch new posts.

Consumer Cycle

Runs every 30 minutes to process and store data.

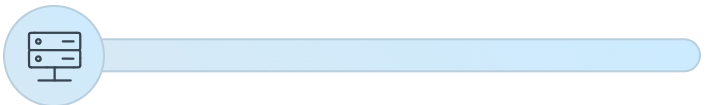
System Architecture

Event-Driven Pipeline Design



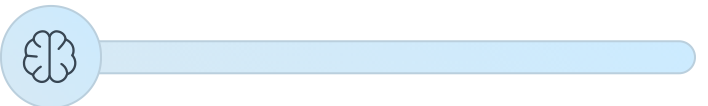
Data Pipeline Workflow

Three-Stage Processing Logic



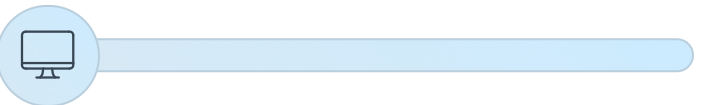
1. Producer Stage

The ingestion layer uses multi-threaded scraping to target 6 subreddits in parallel. It includes robust rate limit handling and fetches a mix of 5 posts plus 5 top comments per post to ensure depth of conversation.



2. Consumer Stage

The processing layer applies VADER sentiment analysis to incoming payloads. It ensures data integrity through duplicate detection (using unique indexes) and utilizes a MongoDB upsert strategy to prevent redundancy.



3. Dashboard Stage

The presentation layer delivers real-time metrics with an auto-refresh cycle every 5 minutes. Users can explore data via interactive filters for specific timeframes or subreddits.

Technology Stack

Modern Data Engineering Tools

Data Source

Reddit API (PRAW) for accessing high-volume social discourse.

Message Broker

Confluent Cloud Kafka using SASL_SSL for secure, decoupled messaging.

Stream Processing

Python + NLTK VADER for efficient natural language processing.

Storage

MongoDB Atlas with a 7-day TTL policy for efficient storage management.

Visualization

Streamlit for rapid deployment of interactive data dashboards.

Automation

GitHub Actions CI/CD for scheduled execution and continuous integration.

Sentiment Classification with VADER

The system utilizes NLTK VADER (Valence Aware Dictionary and sEntiment Reasoner), a model specifically attuned to sentiments expressed in social media contexts.

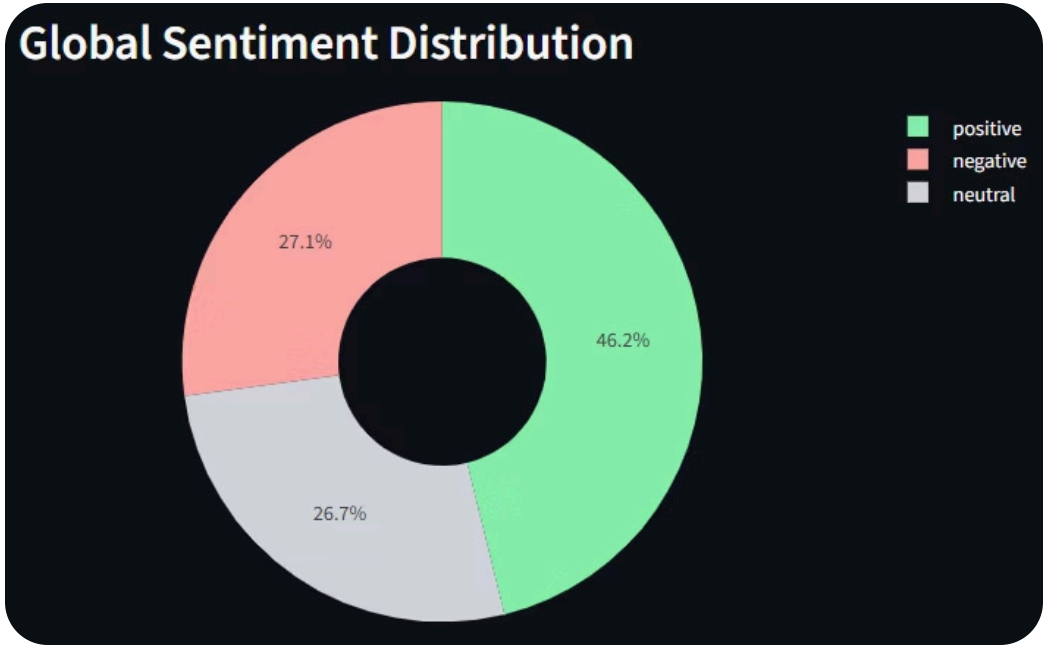
Positive (≥ 0.05)

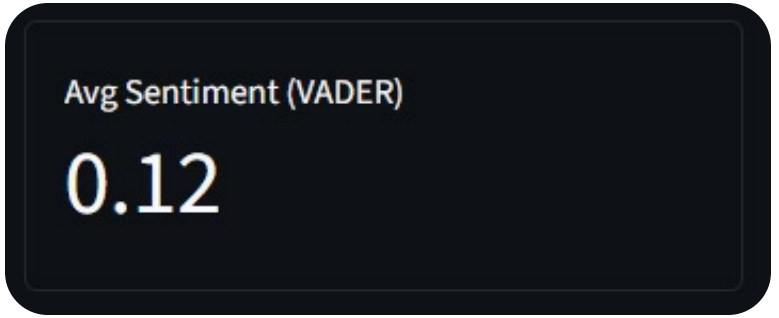
Neutral (-0.05 to 0.05)

Negative (≤ -0.05)

Dashboard Insights

Executive Summary View





Snapshot Metrics

The dashboard processes over 55,058 documents to provide a high-level view of community health.

0.12

Average Compound Score

Positive

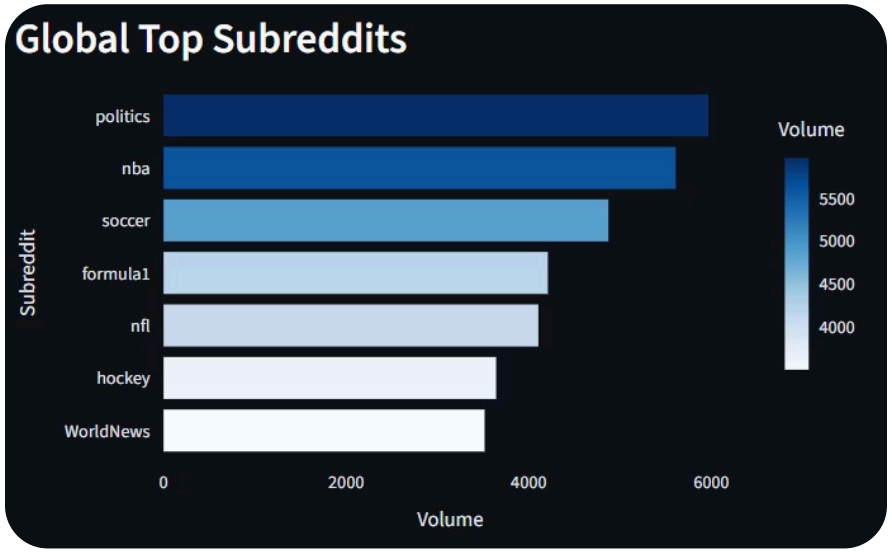
Dominant Sentiment

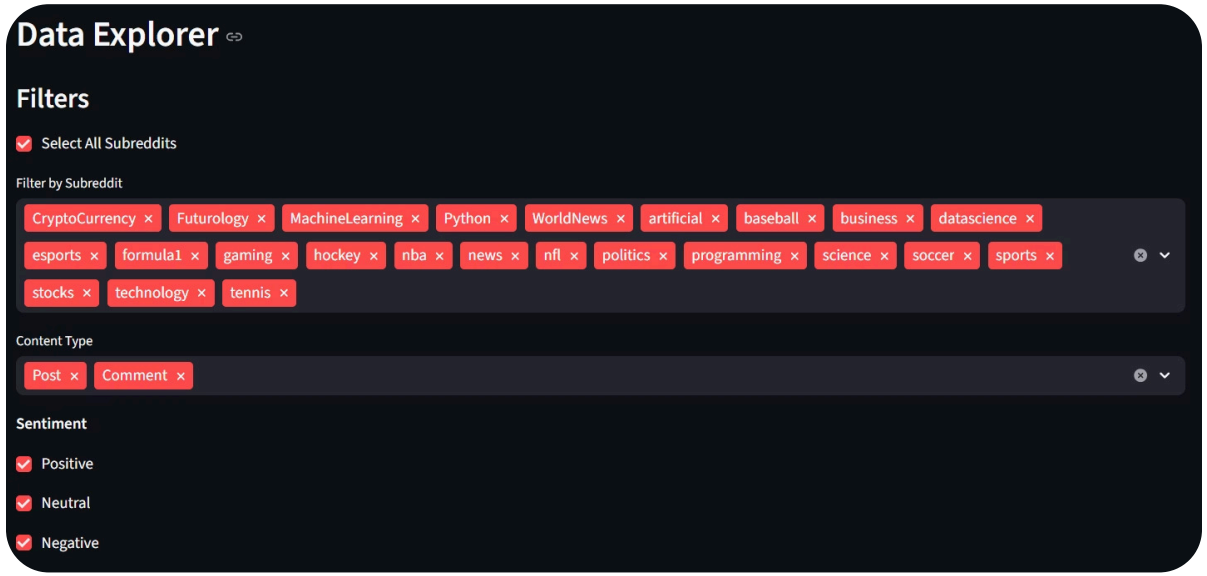
The system tracks 24-hour sentiment trends with granular half-hourly updates, allowing for immediate detection of mood shifts.

Dashboard Insights

Deep Dive Analytics

Granular controls allow users to filter by specific subreddits, content types, and sentiment thresholds.





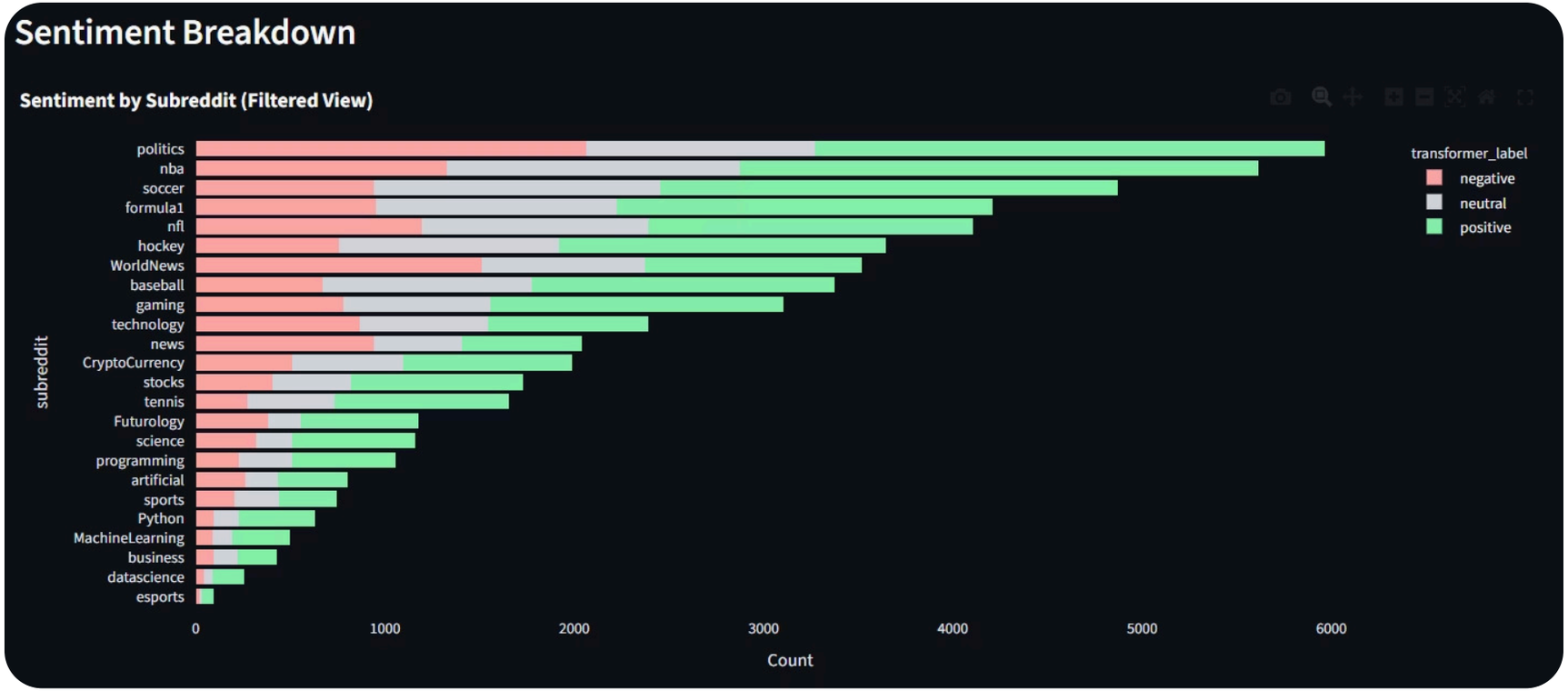
Volume Leaders

Top activity detected in large-scale communities like **r/politics**, **r/nba**, and **r/soccer**.

Distribution Profile

Sentiment is generally balanced, with a plurality (46.2%) leaning positive, while negative and neutral sentiments maintain roughly equal shares.

Key Findings & Patterns



News

Negative Bias: Communities like r/WorldNews trend negative due to crisis and conflict coverage.

Tech

Positive Bias: r/Python and r/MachineLearning focus on innovation, driving positive scores.

Finance

Volatile: Crypto and stock subreddits fluctuate heavily based on market performance.

Sports

Enthusiastic: r/nba and r/soccer show high positivity driven by fan engagement.

Achievements & Future Work

Technical Achievements

- **Zero-Touch ETL**

  Fully automated pipeline requiring no manual intervention.

- **Data Integrity**

  100% unique records via robust duplicate detection.

- **Production Ready**

  Comprehensive error handling within a scalable event-driven architecture.

Future Enhancements

→ **Transformers**

Replace VADER with DistilBERT or RoBERTa for nuance.

→ **Alerting**

Implement anomaly detection for sudden sentiment spikes.

→ **Advanced NLP**

Add LDA/BERTopic modeling and multilingual support.

Conclusion

Our real-time Reddit sentiment analysis pipeline autonomously transforms raw social discourse into actionable intelligence, providing immediate insights into public perception and emerging trends.

Instant Insights

Unlock rapid understanding of community sentiment and reactions.

Continuous Monitoring

Automated, zero-touch data ingestion and analysis for always-on intelligence.

Strategic Value

Inform business and communication strategies with real-time social intelligence.