Real-Time Sentiment Analysis Pipeline for Reddit Data Using Kafka

Maxime APPERT

INSTITUT POLYTECHNIQUE DE PARIS

Table Of Contents

Table Of Contents

Why Sentiment Analysis is Interesting

- Sentiment analysis is important because it helps us understand public opinion and detect trends in textual data. It is particularly useful for businesses to gauge customer satisfaction across their different products.
- It can be used to analyze public opinion on social trends and how these evolve over time, in particular if we have access to targeted communities.
- For this project, analyzing Reddit data allows us to have a wide range of topics while still being able to target specific subjects.

Why Reddit was Chosen Over Twitter

- Twitter's free API tier imposes strict rate limits, making it difficult to collect sufficient data for meaningful analysis.
- Reddit offers flexibility with its API, allowing easier access to historical data and unlimited post collection without heavy restrictions.
- Subreddits enable sentiment analysis to be targeted to specific communities, making it easier to gain insights into more specific topics or demographics.

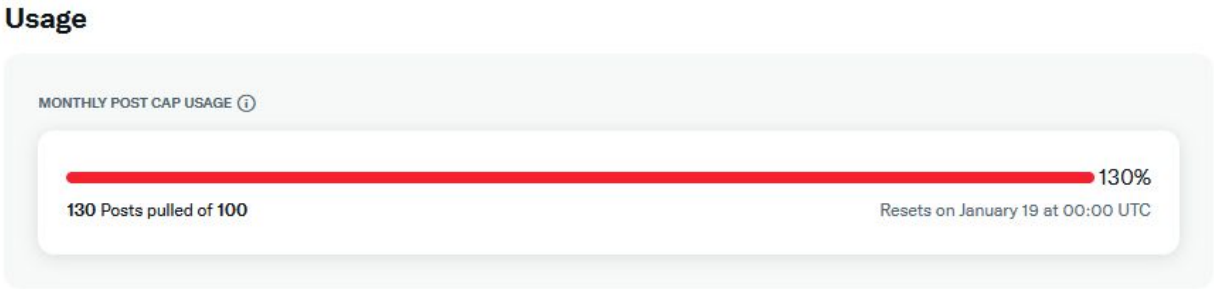


Table Of Contents

Pipeline Overview

The pipeline is built around two Kafka topics: **raw_reddit_posts** and **processed_reddit_posts**. Each script in the pipeline plays a specific role in producing, processing, and consuming data through these topics:

- **Producer Script**:
  The producer script is the starting point of the pipeline. It fetches Reddit posts in real-time using the Reddit API and sends them to the Kafka topic **raw_reddit_posts**. These posts are sent as JSON objects, containing metadata like the title, text, subreddit, and post ID.
- **Consumer Script**:
  The consumer script consumes posts from the **raw_reddit_posts** topic. It uses a pre-trained machine learning model to classify the sentiment of each post as positive or negative. If the model is uncertain about a prediction, it collects user feedback and adjusts the sentiment accordingly by adding the feedback in a file used for retraining. The processed posts, along with their classified sentiment, are then sent to the **processed_reddit_posts** topic for further use.
- **Archiver Script**:
  The archiver script reads data from the **processed_reddit_posts** topic. It separates posts by their sentiment and saves them incrementally to local JSON files, positive_posts.json and negative_posts.json. This archival system ensures that all classified posts are stored for long-term analysis and retrieval.

Pipeline Overview

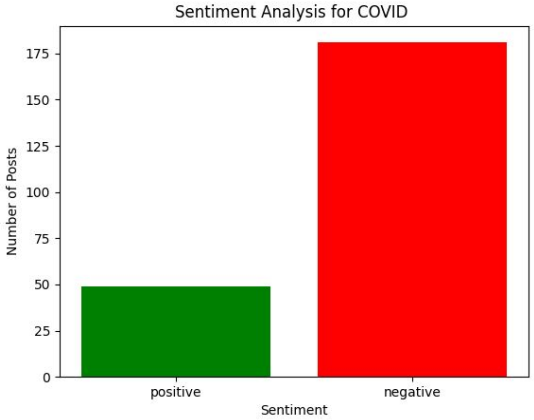
- **Graph Script**:
  The graph script also consumes data from the processed_reddit_posts topic. It maintains a running count of positive and negative sentiments and periodically generates bar graphs to visualize the sentiment distribution. These graphs help identify trends and provide quick insights into the overall mood of the analyzed posts.
- **WordCloud Script**:
  Similar to the graph script, the word cloud script consumes data from the processed_reddit_posts topic. It analyzes the text of posts grouped by sentiment, creating word clouds that highlight the most frequently used words in positive and negative posts. These visualizations are useful for uncovering recurring themes and keywords.
- **Training Script**:
  While not directly connected to the Kafka topics, the training script plays a crucial role in the pipeline by retraining the sentiment classification model. It combines the Sentiment140 dataset with user feedback saved in userfeedback.csv to improve the model's accuracy over time. The updated model is then used in the consumer script for sentiment analysis. There is also a very simple testing script to ensure that model functions correctly, it inputs very simple positive or negative texts and tests to see whether the model answers correctly.

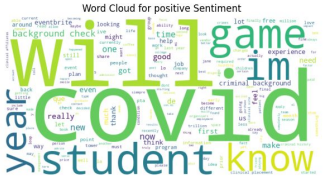
Table Of Contents

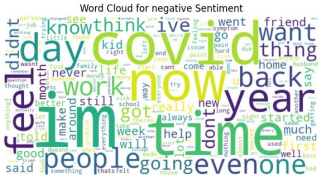
Section 3:Experiments and Results

COVID Keyword Analysis:

- Observed significantly more negative sentiments, as expected.
- Positive words: “student” (perhaps students that enjoyed skipping school).
- Negative words: “year”, “friend”, “house” (indicating stress during lockdown, forced to stay inside the house and unable to see friends).



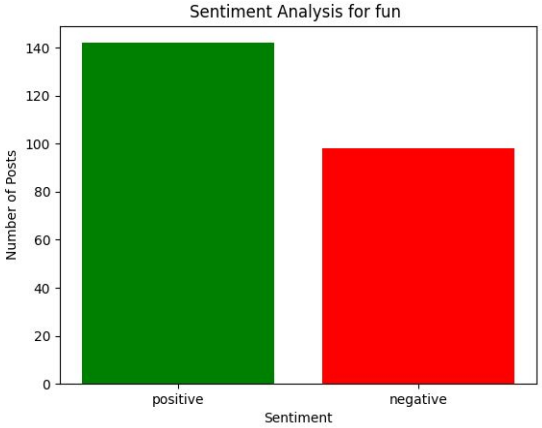


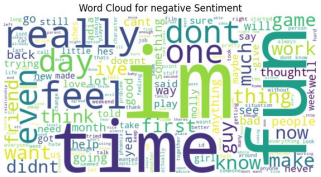


Section 3:Experiments and Results

Fun Keyword Analysis:

- Observed more positive sentiments, as expected.
- Positive words: "game", "play", and "new" (new activities that can easily be linked to feelings of joy).
- Negative words: "don't", "feel", and "even" (possibly people ranting about troubled periods in their lives).

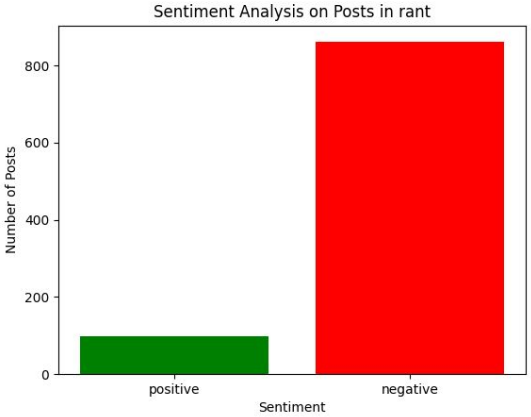


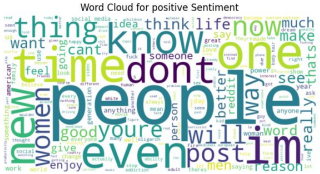


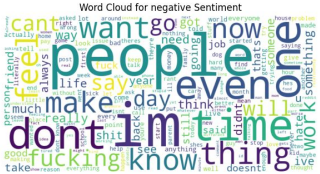
Section 3:Experiments and Results

Rant Subreddit Analysis:

- Significantly more negative sentiments, as expected.
- Words like “people”, “don’t”, “time” dominated both positive and negative word clouds, which can be explained by the fact that even texts that can have a positive note are frustrated and venting.



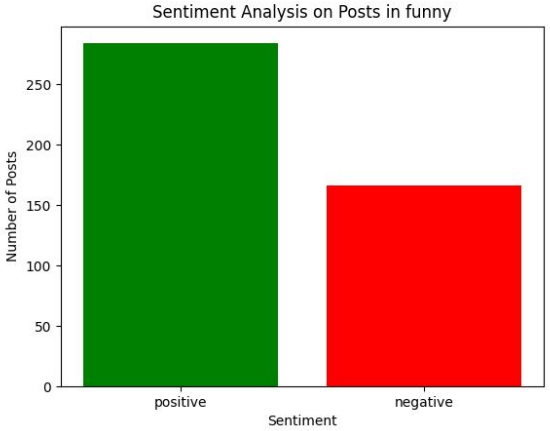


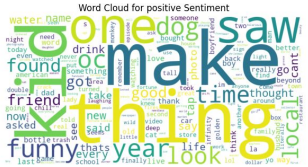


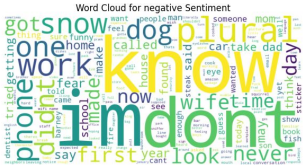
Section 3:Experiments and Results

Funny Subreddit Analysis:

- More positive sentiments, as expected.
- Positive words: “dog”, “kid”, “friend” (expected results as family, friends and pets are common causes of joy).
- Negative words: “work”, “never” (expected results as well, common complaints people have include their work or impossibilities in their lives).



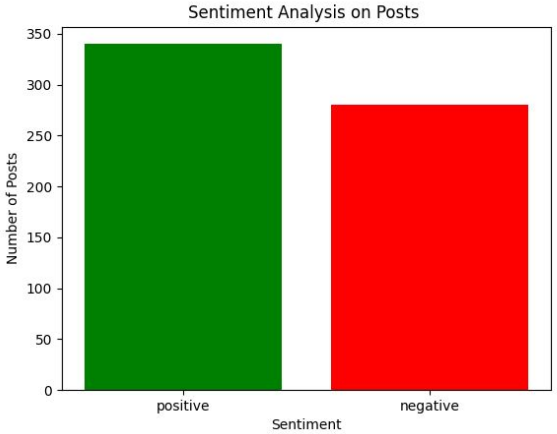


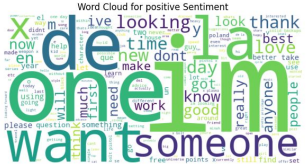


Section 3:Experiments and Results

All Subreddits Analysis:

- Balanced positive and negative sentiments, with slightly more positive sentiments.
- Positive words: “new”, “love”, “want” (This is an expected result as these are common themes of gratitude and optimism, not centered around one particular topic).
- Negative words: “don’t”, “work”, “year” (This is an expected as well as these are common themes of stress and dissatisfaction).





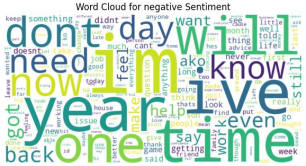


Table Of Contents

Section 4: Conclusion and Future Improvements

The sentiment analysis pipeline built in this project has shown solid performance across various scenarios, it effectively processes Reddit posts in real-time and generates useful visualizations like graphs and word clouds, as well as allowing the incorporation of user feedback to refine the model. The experiments also helped to show that the model behaves as expected across a wide variety of topics.

While the pipeline is functional and delivers valuable results, there are opportunities for improvement. Enhancing the feedback mechanism with a more user-friendly interface could improve user interactions, and adopting more advanced machine learning models than logistic regression could improve sentiment classification, particularly for nuanced or ambiguous posts. Overall, despite these areas for growth, the current implementation is a robust framework that performs well on large-scale textual data.

Table Of Contents

Section 5: Questions and Answers