

DATA ENGINEERING-2 REPORT: YOUTUBE COMMENT SENTIMENT ANALYSIS WITH GCP

Group Names: Rohan Mahaveer, Manoj Saligrama Harisha, Namratha Prakash, Valeska Joshna Dsouza

Course: MSc in Data Science and Analytics

➤ Introduction

Real-Time YouTube Sentiment Analysis Pipeline

Unlocking Audience Insights at Scale

Why This Matters:

YouTube generates millions of comments daily—raw, unstructured feedback representing real audience emotions and reactions.

Content creators and brands need to understand:

- What viewers love or hate about videos
- When sentiment shifts occur (viral spikes, controversies)
- How to respond proactively to build engagement and loyalty

This Pipeline Delivers:

Automated, serverless GCP architecture that transforms chaotic comment streams into actionable sentiment metrics—hourly, at scale, without ops overhead.

From Raw Comments → Strategic Insights in minutes, not weeks.

➤ Problem Statement

Monitoring YouTube Sentiment at Scale

The Challenge:

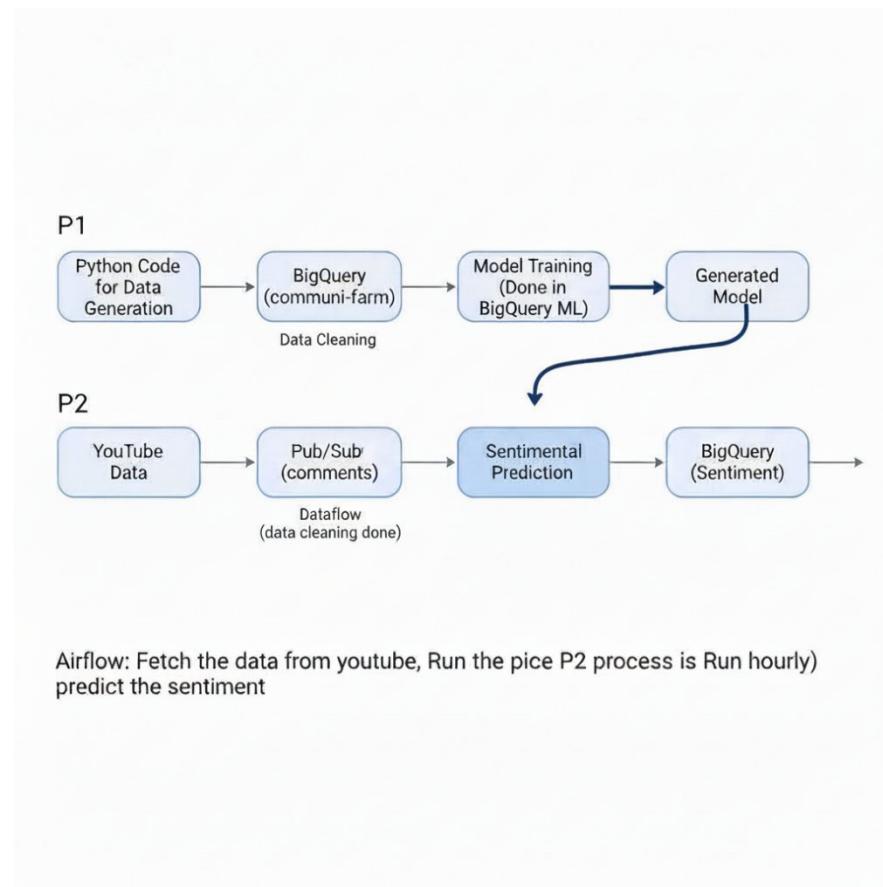
- Thousands of YouTube comments arrive hourly across channels
- Manual sentiment analysis is unscalable and time-delayed
- Content creators need real-time audience reaction insights
- Traditional approaches require expensive servers and DevOps overhead **The Challenge.**

- High-volume YouTube channels receive thousands of comments hourly
- Manual sentiment analysis is unscalable and time-consuming
- Need to understand audience sentiment trends in near real-time
- Cannot respond quickly to negative sentiment spikes.

➤ Architecture OverView

P1: Training Pipeline (Batch)

P2: Prediction Pipeline (Streaming) Orchestration: Airflow hourly schedule Core Pattern:
Pub/Sub → Dataflow → BigQuery ML → Sentiment Table



➤ TWO PHASE ARCHITECTURE

Phase 1: Model Training Pipeline - Build & train the sentiment model once

Step 1: Python generates labeled training data

Step 2: BigQuery SQL cleans & preprocesses data.

Step 3: BigQuery ML trains sentiment model (SQL).

Output: Reusable ML model ready for predictions.

Phase 2: Real-Time Prediction Flow - Apply model to new comments hourly

YouTube API: Fetch new comments hourly

Pub/Sub: Stream comments as messages

Dataflow: Clean & transform text (Apache Beam)

BigQuery ML: Apply model, store results with sentiment labels.

➤ YouTube Data API Concepts

- `commentThreads.list()` - Fetch top-level comments
- Pagination with `nextPageToken`
- Rate limits: 10,000 units/day (100 comments = 1 unit).
- Authentication: API key or OAuth 2.0.

➤ Pub/Sub Messaging

Topic: youtube-comments (publish JSON payloads) Subscription: comments-sub (Dataflow subscriber) Message format: `{"video_id": "...", "text": "...", "timestamp": "..."}` At-least-once delivery guarantees.

The screenshot shows the Google Cloud Pub/Sub Subscriptions interface. On the left, there's a sidebar with navigation links like 'Topics', 'Subscriptions', 'Schemas', and 'Pub/Sub Lite'. The main area displays a subscription named 'temp' with the following details:

- Subscription name:** projects/ambient-elf-487017-d6/subscriptions/temp
- Subscription state:** active
- Topic name:** projects/ambient-elf-487017-d6/topics/y-comments-topic
- Tags:** Insufficient permission to list tags

Below this, there are tabs for 'Metrics', 'Details', and 'Messages'. The 'Messages' tab is selected, showing a list of 15 messages. Each message includes fields for 'Ordering key', 'body.comment', 'bodyId', and 'Ack'. Most messages have 'Deadline exceeded' in the 'Ack' column. The messages content varies, including comments like 'You're freaknnnnn amazing mannn...', 'You will be in jail for making joke on turba...', and 'Persian not Iraq 😂 I prefer Persian'.

➤ Dataflow (Apache Beam)

- **Runner:** DataflowRunner (managed).
- **Transforms:** ParDo for cleaning, ML.PREDICT calls.
- **Windowing:** Fixed windows for hourly batches.
- **Auto-scaling:** 1-100 workers based on load.

The screenshot shows the Google Cloud Dataflow Job details interface. The left sidebar has links for 'Overview', 'Monitoring', 'Jobs' (which is selected), 'Pipelines', 'Workbench', and 'Schemas'. The main area shows a 'Job Graph' with three stages:

- ReadFromPubSub**: Running, Data Lag: 37.18 sec, Total Walltime: 1 min, Max Op Latency: < 1 sec, 1 stage.
- CleanText**: Running, Data Lag: 37.18 sec, Total Walltime: 1 min, Max Op Latency: < 1 sec, 1 stage.
- WriteToBigQuery**: Running, Data Lag: 37.18 sec, Total Walltime: 1 min, Max Op Latency: < 1 sec, 2 stages.

On the right, there are sections for 'Step info', 'Input collections', and 'Output Data Freshness'. The 'Input collections' section shows a chart for 'Throughput (elements/sec)' with no data available. The 'Output Data Freshness' section shows a chart for 'Data freshness by stages' over time (UTC+1: 9:15 PM, 9:20 PM, 9:25 PM).

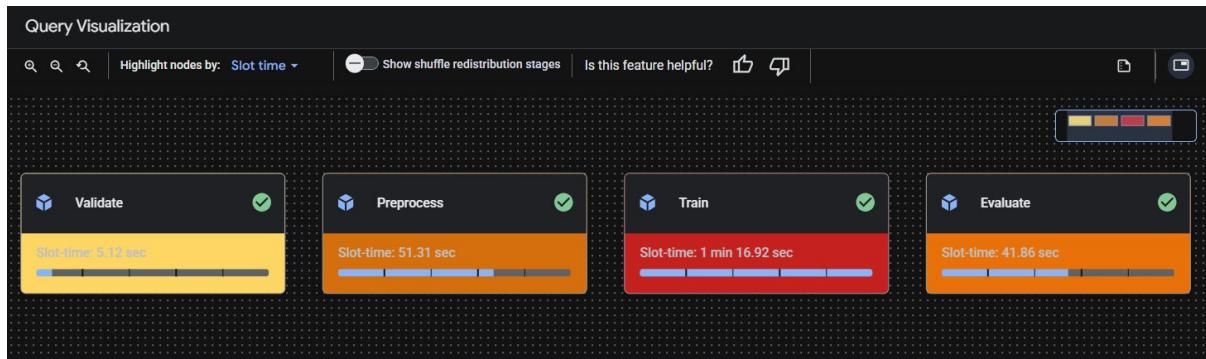
➤ BigQuery ML Model Lifecycle

CREATE MODEL → ML.TRAINING_INFO → ML.PREDICT → ML.EVALUATE

Model type: logistic_reg (binary classification)

Features: TF-IDF vectors from text

Hyperparams: max_iterations=20, L2_reg=0.1

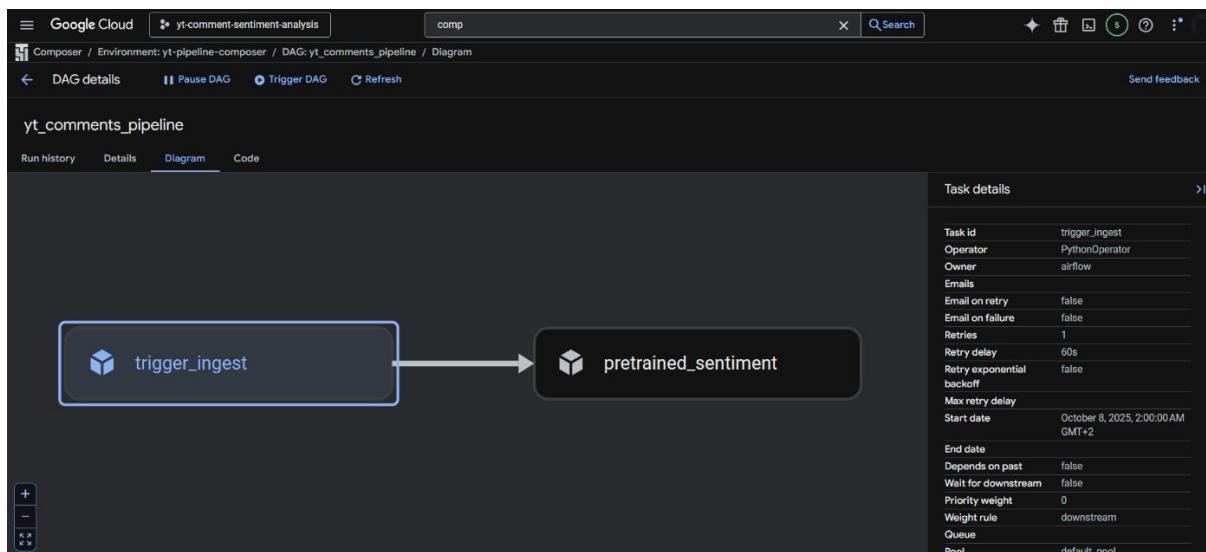


➤ Airflow DAG Patterns

Operators:

- CloudFunctionOperator (fetch comments)
- DataflowTemplatedJobStartOperator (prediction)
- BigQueryOperator (model refresh)

Schedule: @hourlyDependencies: fetch >> predict >> validate.



➤ Data Cleaning Pipeline

Steps:

1. Lowercase + trim
2. Remove URLs (REGEXP 'http')
3. Remove @mentions, #hashtags
4. Min length 10 chars

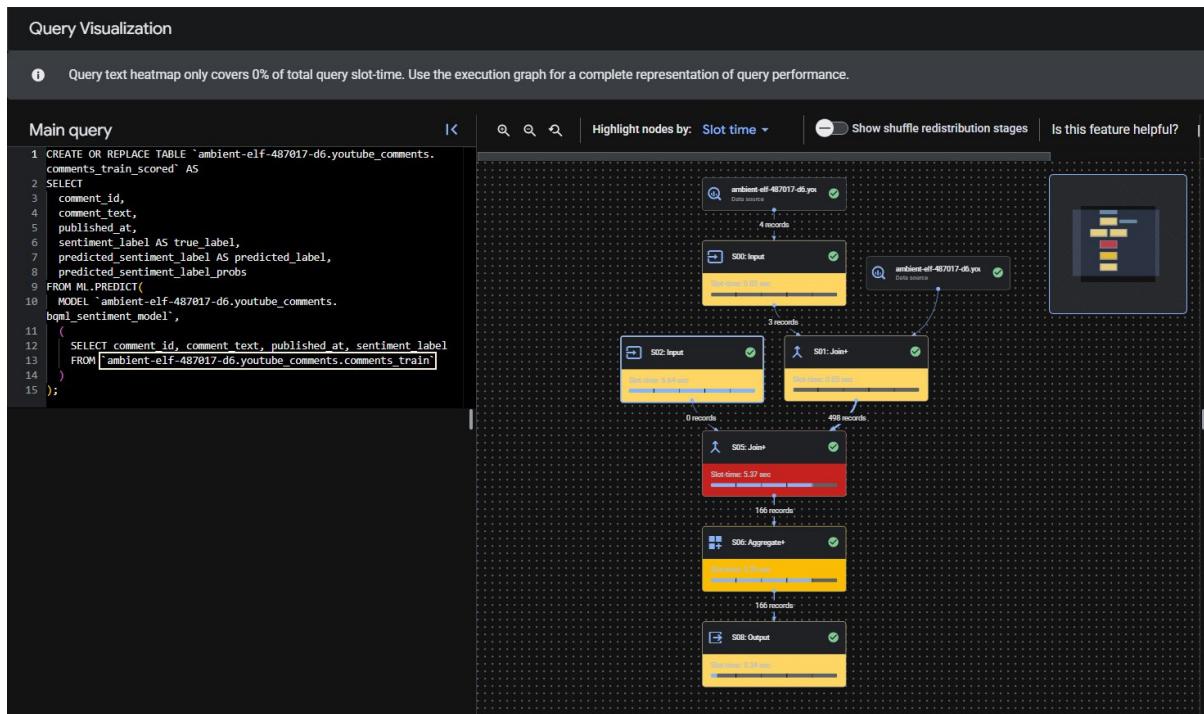
5. Remove duplicates (ROW_NUMBER)

➤ Error Handling

1. Dataflow dead-letter queue
2. Airflow retries (3x)
3. BigQuery query validation
4. Pub/Sub message ACK timeouts.

➤ CI/CD Pipeline

1. GitHub Actions
2. Terraform for infra
3. Cloud Build for Dataflow templates
4. Cloud Composer environments.



➤ Outputs

yt-comments-487009 / Datasets / youtube_comments / Tables / comments_train_scored

comment_id comment_text published_at true_label predicted_label predicted_sentiment_label_prob_... predicted_se...

Row	comment_id	comment_text	published_at	true_label	predicted_label	predicted_sentiment_label_prob_...	predicted_se...
1	1cb1904b-5f07-4c7a-9b54-c0d4...	Build election shake trial imagine.	2026-02-11 08:55:47.771869 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
2	a92b5e5a-7c88-4a77-b141-0e5...	Of among total general thus.	2026-02-11 08:41:47.771203 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
3	501a9ea9-00b9-4ab2-b9cc-28a...	Deep whole development move...	2026-02-11 08:20:47.771549 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
4	fbc97752-0fc3-4453-a25a-e52c...	Bad fill long debate suggest ma...	2026-02-11 09:23:47.771726 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
5	1210ec12-47b4-4fd3-b7a2-b19...	Create live number well charge common play quickly determine physical matter happen reason clearly finish subject cell candidate.	2026-02-11 07:50:47.771549 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
6	a435a911-e55a-48ce-ac30-41d...	Most course pass goal program check establish save several big reflect weight.	2026-02-11 07:01:47.771869 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
7	a9a1678a-8d55-404e-bc0f-a83f...	Well point similar listen than ma...	2026-02-11 06:57:47.771726 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
8	cb2f742f-d991-469f-a2c5-335...	White spring exactly treat build ...	2026-02-11 07:25:47.771869 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
9	f3e1d8f0-c2f4-49fe-8b64-dcb6c...	We player i especially produce b...	2026-02-11 06:54:47.771869 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
10	cc47f8ee-cb0d-4f45-8b86-9027...	Better few course heart much woman house catch surface model stuff.	2026-02-11 07:28:47.771869 UTC	negative	negative	0.377380920...	
					positive	0.312947996...	
					neutral	0.309671083...	
11	8b75d147-2d5b-4411-9ba6-fdb...	Style choice matter still them surface thought figure officer	2026-02-11 07:39:47.771869 UTC	negative	negative	0.377380920...	

➤ Model Performance

The screenshot shows the Google BigQuery web interface. On the left, the sidebar includes sections for Overview, Studio (selected), Pipelines & Integration, Dataform, Scheduled queries, Scheduling, Governance, Sharing (Analytics Hub), Policy tags, Metadata curation, Administration, Partner Center, and Settings. The main area displays a query titled "Untitled query" with the following SQL code:

```

WITH eval AS (
  SELECT *
  FROM ML.EVALUATE(
    MODEL 'ambient-elf-487017-d6.youtube_comments.bqml_sentiment_model'
  )
)
SELECT
  accuracy,
  precision,
  recall,
  2 * (precision * recall) / (precision + recall) AS f1_score
FROM eval;
  
```

The status bar indicates "Query completed". Below the code, the "Query results" section shows a single row of data:

Job information	Results	Visualization	JSON	Execution details	Execution graph
Row	accuracy	precision	recall	f1_score	
1	0.243243243243...	0.081081081081...	0.333333333333...	0.130454782608...	

➤ Future Enhancements

- Vertex AI (Gemini 1.5) integration
- Multi-language support
- Video title/description analysis
- Competitor channel benchmarking

