YouTube Comment Sentiment:
Cloud-Scale Pipeline for Video Impact Insights

CS-131 SJSU Big Data Lab Final Report
Team 10 – *The Tag Team (TTT)*

Team Members

- Karmehr Arora
- Jacob Atanacio
- Trae Carr
- Hien Ly
- Richa Vakharia

Table of Contents

Executive Summary

This project builds a cloud-based big data pipeline that analyzes YouTube comment sections to answer a simple question:

"Is this video worth my time?"

Using YouTube's Data API and PySpark/Spark, we collect, clean, and process ~190K+ YouTube comments to generate summaries, engagement metrics, and positivity signals for videos. Our goal is to help students, professionals, and the general public quickly decide which videos are most relevant, high-quality, and worth watching—especially when time is limited.

Over the semester, our team:

- Assembled and cleaned a large-scale comments dataset
    - Started with ~300K rows from YouTube; after cleaning, retained 189,843 valid comments across 2,013 videos and 154,059 distinct authors.
- Moved from Unix-based scripts to a PySpark pipeline, reconstructing prior frequency and Top-N analyses in a distributed environment.
- Deployed the pipeline to a cloud-based Spark cluster, reading and writing data from cloud object storage and demonstrating distributed execution.
- Deepened analysis of engagement patterns, including: Top channels, videos, and authors by activity, high-activity days and posting windows, network-style patterns of hubs (viral videos and power users), and simple positivity and keyword-based analysis (e.g., "great") as a first step toward sentiment scoring.

Top outcomes of our big data work:

1. We can now successfully compute core engagement metrics (by video, channel, author, and date) on Youtube comments in minutes in the cloud rather than relying on manual scripts on a single machine.
2. We show that attention and engagement are highly centralized: a small number of videos, channels, and repeat commenters dominate the conversation.
3. We provide a foundation for video-worthiness signals with engagement patterns, positivity, and topic keywords that can double as summaries, tags, and future recommendation tools.

Key recommendations:

- Invest in converting this pipeline as a real-time cloud job with downstream dashboards so product teams can experiment with "comment summaries" as a feature.
- Prioritize high-engagement and highly-positive videos for pilots
  (a "cram-friendly" video finder for students based on comment-derived insights.)

1. Project Overview and Context

1.1 Problem and Vision

YouTube is a primary learning and information platform. But when users search for a topic, they face too many videos and too little time.

Our vision:

> **Use big data on YouTube comments to quickly tell users which videos are actually useful, engaging, and positively received.**

We focus on the comment section, treating it as a crowdsourced review system. By analyzing large volumes of comments at scale, we aim to:

- Highlight **engagement intensity** (how much discussion a video generates).
- Surface **positivity and enthusiasm** in comments.
- Identify **keywords and recurring topics** from comment text.

1.2 Target Audience

- **Students** trying to cram before exams or interviews and choose the "one best" video.
- **Professionals, academics, and general viewers** selecting efficient, high-value videos.
- **Platform or product teams** at media or ed-tech organizations who want to build features like "comment summaries," "community mood," or "top discussion topics."

1.3 Example Scenario

A student has **one hour** before an exam and finds five possible YouTube crash-course videos. Instead of guessing, they could:

- View a summary of the comment section for each video.
- Check tags derived from top keywords (e.g., "good for beginners," "too fast," "great visual examples").
- See a positivity indicator based on comment patterns.

They then choose the video with the strongest signal of "all-in-one crash course" plus high positive engagement—powered by our pipeline.

2. Dataset Introduction

2.1 Data Source and Scope

- **Primary Source**: YouTube Data API (comments resource).
  - We created our own dataset as a sample for processing.
- **Scale (project dataset)**:
  - Initial raw rows: ~301,410 comments
  - Cleaned rows: **189,843** comments
  - Distinct authors: **154,059**
  - Distinct videos: **2,013**
- **Scope:**
  - All data was collected from specified youtube channels including up to 15 videos per channel
  - All data was captured from the span of several months this year (April - September)

For each comment, we capture:

- **Video-level fields**:
  - video_id, video_title, channel_id, channel_title, video_published_at
- **Comment-level fields**:
  - comment_id, author_display_name, author_channel_id, comment_text
  - like_count, reply_count, is_reply, parent_comment_id
  - published_at, updated_at
  - language_detected

Data arrives from the API as **JSON** and is then normalized into **CSV/Parquet** for further processing.

2.2 Data Quality Issues

Key issues we had to handle:

- Null and invalid IDs for videos, channels, or authors.
- Duplicate rows in early harvests.
- Inconsistent text and language detection, including non-English or mixed-language comments.
- Missing video metadata in early sprints (initially had only IDs, not titles or like counts).

* Issues were addressed via cleaning steps before PySpark aggregation and cloud deployment.

---

3. Why This Problem Matters

3.1 Organizational Value

For a platform like YouTube, comment analysis can:

- Improve recommendation quality by adding community-driven signals.
- Help educators and content teams quickly identify high-impact videos.
- Support product features like “Top Community Themes” or “Most Appreciated Crash Courses.”

3.2 Why This Dataset?

We chose YouTube comment data because it:

- Reflects real-world social and learning behavior at scale.
- Is naturally suited to big data tools due to volume and complexity.
- Connects directly to relatable use cases (students, learners, creators).

3.3 Potential Impact

A scalable comment analytics pipeline can:

- Allow continuous monitoring of video engagement and sentiment.
- Reveal attention hubs and long-tail patterns in online communities.
- Serve as a foundation for future sentiment models and recommendation systems.

4. Work Across the Semester – Building the Pipeline

4.1 Sprint 1 – Project Setup and Dataset Definition

Focus: Vision, dataset choice, and basic tooling.

- Defined the vision: summaries, tags, and positivity ratings for YouTube videos based on comments.
- Identified initial data sources:
    - YouTube Data API (primary)
    - Kaggle YouTube datasets (backup and reference).
- Designed the initial schema, including comment and video fields listed above.
- Set up the GitHub repository (team-10--Social-media-behavior), directory structure (data/, src/, out/), and project management board (Google Sheets).

We also identified core challenges:

- How to filter out "filler" or low-information comments.
- How to compute pros/cons and sentiment.
- How to connect videos, channels, and comments into a coherent view.

4.2 Sprint 2 – Unix-Based Frequency Analysis

Focus: Scaling early analysis with shell tools on larger CSV samples.

- Expanded dataset to another ~100K comments, pushing total comment volume upward.
- Used Unix tools (shuf, cut, grep, sort, uniq) to:
    - Sample rows (shuf 1000) for spot-checking.
    - Count comment frequency by channel and by date.
    - Identify top videos by comment volume.

Key insights:

- Top channels: A small set of channel IDs appeared much more frequently than others, indicating highly active creators and communities.
- High-activity dates: Comment frequencies revealed "peak comment days", which can inform optimal upload times or reveal viral spikes.
- Top videos: A short list of video IDs received tens of thousands of comments, highlighting viral or deeply engaging content.

Limitations at this stage:

- Most insights were tied to IDs rather than human-readable titles.
- Analysis was still limited to single-machine Unix tools.

4.3 Sprint 3 – Network and Cluster Analysis

Focus: Treating the dataset as a social graph.

We modeled relationships as edges in a graph

After filtering for entities with at least 10 occurrences, we:

- Built cluster-size histograms.
- Generated Top-30 lists of videos, authors, and comments by connectivity.
- Visualized clusters in Gephi, especially the “great” sentiment cluster.

Findings:

- Engagement follows a power-law / long-tail distribution:
    - A few videos and authors act as major hubs.
    - Most users and videos have only small clusters of activity.

Limitations:

- Using a frequency cutoff of 10 excluded small but potentially meaningful groups.
- Lack of full relational joins limited interpretability.

4.4 Sprint 5 – Transition to PySpark

Focus: Porting prior logic into a distributed PySpark environment.

- Imported the YouTube comments dataset into a PySpark DataFrame.
- Cleaned the data:
    - Dropped null and invalid identifiers.
    - Standardized types and removed malformed rows.
- Resulting clean dataset:
    - 189,843 comments, 154,059 authors, 2,013 videos.

With Spark SQL and DataFrame operations, we rebuilt and extended earlier analyses:

- Comment frequency by date.
- Comment activity by author.
- Engagement metrics per video and channel

This sprint scaled our earlier Unix analyses into a big data framework, setting the stage for cloud deployment and advanced analyses.

---

5. Final Sprint – Cloud Distributed Run and New Analysis

5.1 Cloud-Based Distributed Execution

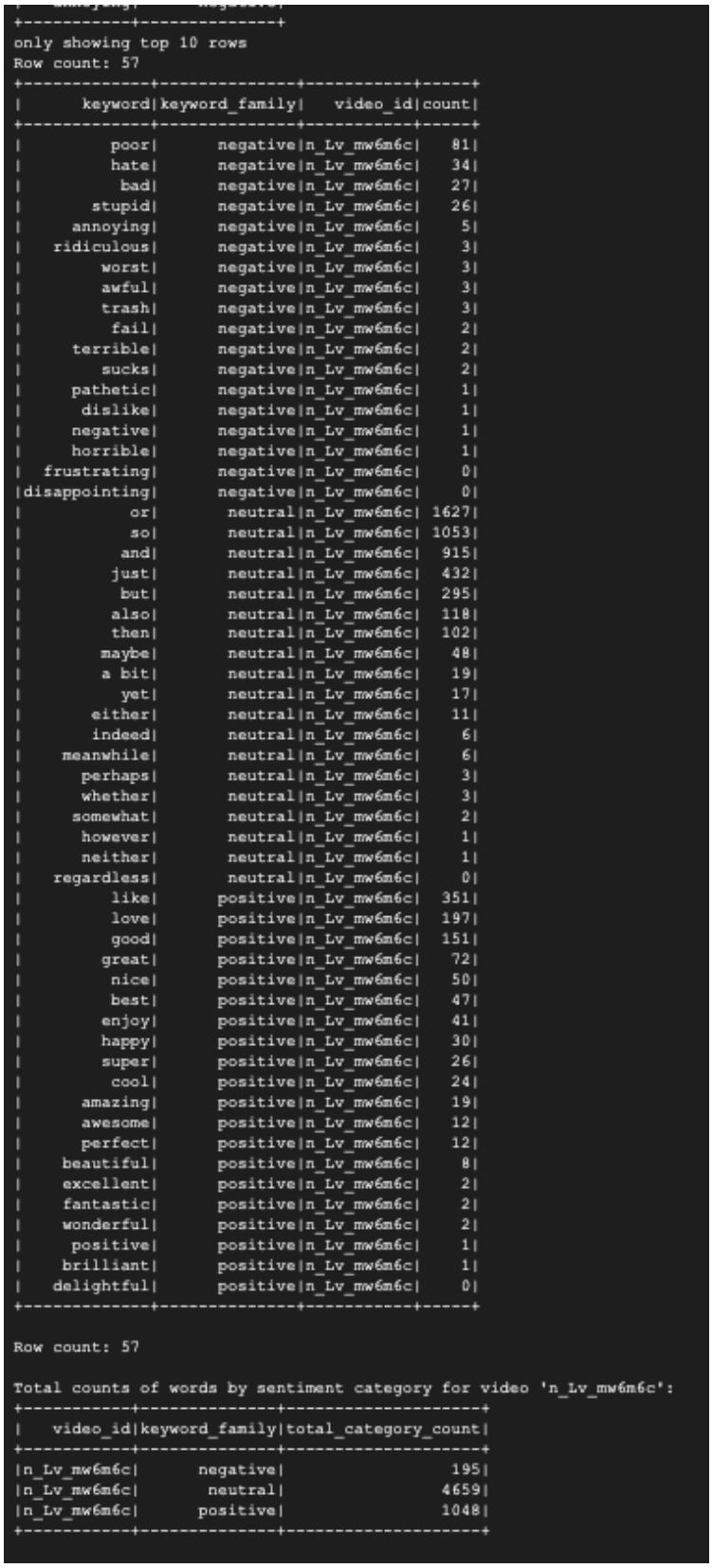
In the final sprint, we moved from local PySpark to a cloud-managed Spark environment

Setup:

- Compute: Spark cluster in the cloud (driver + multiple workers).
- Storage: Cloud object storage for:
    - Raw ingested comments.
    - Cleaned/aggregated tables written as partitioned Parquet.

End-to-end pipeline in the cloud:

1. Read raw YouTube comment data from cloud storage.
2. Clean & transform:
    - Remove null IDs, normalize timestamps, handle language metadata.
    - Join with video and channel metadata where available.
    - Create derived features like: is_reply, day-of-week, hour-of-day, basic positivity flag based on keywords).
    - Used Spark's explode to turn each comment's list of extracted keywords into separate rows, which reduces data skew and makes it easier to group by month, sentiment category, and keyword frequency.



```
+-----------+-------------+
only showing top 10 rows
Row count: 57
+------------+-------------+-----------+-----+
|     keyword|keyword_family|   video_id|count|
+------------+-------------+-----------+-----+
|        poor|     negative|n_Lv_mw6m6c|   81|
|        hate|     negative|n_Lv_mw6m6c|   34|
|         bad|     negative|n_Lv_mw6m6c|   27|
|      stupid|     negative|n_Lv_mw6m6c|   26|
|    annoying|     negative|n_Lv_mw6m6c|    5|
|   ridiculous|     negative|n_Lv_mw6m6c|    3|
|       worst|     negative|n_Lv_mw6m6c|    3|
|       awful|     negative|n_Lv_mw6m6c|    3|
|       trash|     negative|n_Lv_mw6m6c|    3|
|        fail|     negative|n_Lv_mw6m6c|    2|
|    terrible|     negative|n_Lv_mw6m6c|    2|
|       sucks|     negative|n_Lv_mw6m6c|    2|
|    pathetic|     negative|n_Lv_mw6m6c|    1|
|     dislike|     negative|n_Lv_mw6m6c|    1|
|    negative|     negative|n_Lv_mw6m6c|    1|
|    horrible|     negative|n_Lv_mw6m6c|    1|
|  frustrating|     negative|n_Lv_mw6m6c|    0|
|disappointing|     negative|n_Lv_mw6m6c|    0|
|          or|      neutral|n_Lv_mw6m6c| 1627|
|          so|      neutral|n_Lv_mw6m6c| 1053|
|         and|      neutral|n_Lv_mw6m6c|  915|
|        just|      neutral|n_Lv_mw6m6c|  432|
|         but|      neutral|n_Lv_mw6m6c|  295|
|        also|      neutral|n_Lv_mw6m6c|  118|
|        then|      neutral|n_Lv_mw6m6c|  102|
|       maybe|      neutral|n_Lv_mw6m6c|   48|
|       a bit|      neutral|n_Lv_mw6m6c|   19|
|         yet|      neutral|n_Lv_mw6m6c|   17|
|      either|      neutral|n_Lv_mw6m6c|   11|
|      indeed|      neutral|n_Lv_mw6m6c|    6|
|   meanwhile|      neutral|n_Lv_mw6m6c|    6|
|     perhaps|      neutral|n_Lv_mw6m6c|    3|
|     whether|      neutral|n_Lv_mw6m6c|    3|
|    somewhat|      neutral|n_Lv_mw6m6c|    2|
|     however|      neutral|n_Lv_mw6m6c|    1|
|     neither|      neutral|n_Lv_mw6m6c|    1|
|  regardless|      neutral|n_Lv_mw6m6c|    0|
|        like|     positive|n_Lv_mw6m6c|  351|
|        love|     positive|n_Lv_mw6m6c|  197|
|        good|     positive|n_Lv_mw6m6c|  151|
|       great|     positive|n_Lv_mw6m6c|   72|
|        nice|     positive|n_Lv_mw6m6c|   50|
|        best|     positive|n_Lv_mw6m6c|   47|
|       enjoy|     positive|n_Lv_mw6m6c|   41|
|       happy|     positive|n_Lv_mw6m6c|   30|
|       super|     positive|n_Lv_mw6m6c|   26|
|        cool|     positive|n_Lv_mw6m6c|   24|
|     amazing|     positive|n_Lv_mw6m6c|   19|
|     awesome|     positive|n_Lv_mw6m6c|   12|
|     perfect|     positive|n_Lv_mw6m6c|   12|
|   beautiful|     positive|n_Lv_mw6m6c|    8|
|   excellent|     positive|n_Lv_mw6m6c|    2|
|   fantastic|     positive|n_Lv_mw6m6c|    2|
|   wonderful|     positive|n_Lv_mw6m6c|    2|
|    positive|     positive|n_Lv_mw6m6c|    1|
|   brilliant|     positive|n_Lv_mw6m6c|    1|
|  delightful|     positive|n_Lv_mw6m6c|    0|
+------------+-------------+-----------+-----+

Row count: 57

Total counts of words by sentiment category for video 'n_Lv_mw6m6c':
+-----------+-------------+--------------------+
|   video_id|keyword_family|total_category_count|
+-----------+-------------+--------------------+
|n_Lv_mw6m6c|     negative|                 195|
|n_Lv_mw6m6c|      neutral|                4659|
|n_Lv_mw6m6c|     positive|                1048|
+-----------+-------------+--------------------+
```

Figure 5.1: Keyword Count Separated by Family + Final Keyword Family Counts

3. Aggregate & analyze:
    - Comments per video/channel/author/date.
    - Engagement ratios (comments and likes per video).
    - Keyword-based and sentiment-lite summaries.
4. Write results back to cloud storage as partitioned Parquet tables

Evidence of distributed execution:

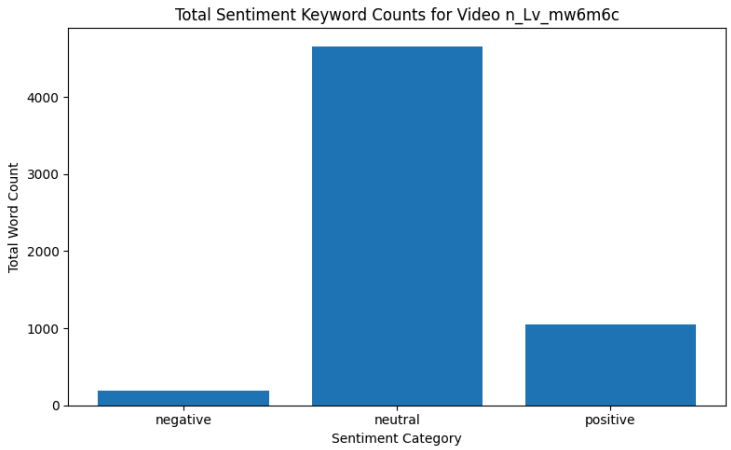


Figure 5.2: Sentiment Keyword Count

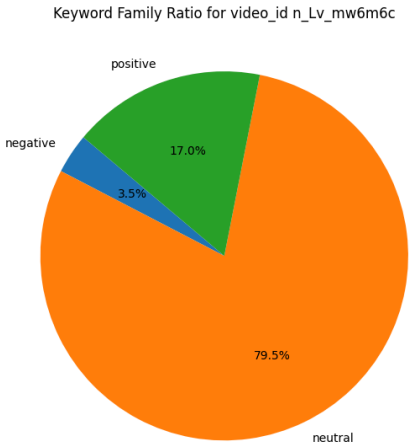


Figure 5.3: Keyword Family Ratio

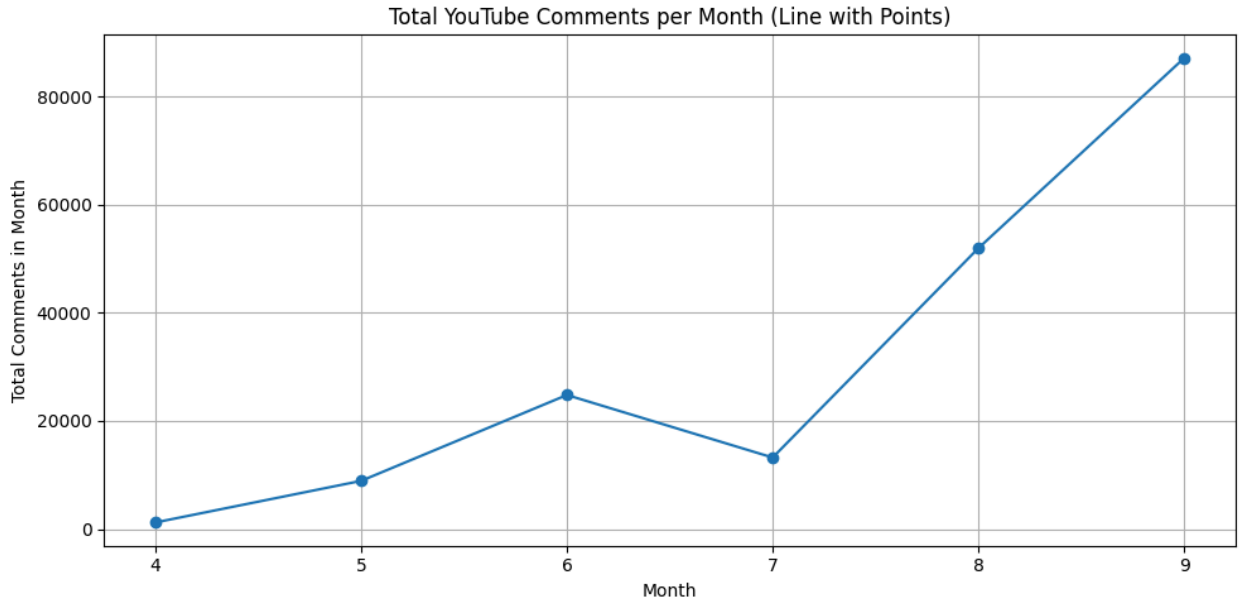


Figure 5.4 Total Youtube Comments per Month

Scalability considerations:

- With current design, a 10× increase in data volume can be handled by:
    - Scaling the cluster size.
    - Preserving partitioning by date and/or channel, enabling parallelism.
- Potential bottlenecks:
    - Large joins between comments and richer video metadata.
    - Skewed channels (very popular creators) concentrating data in a few partitions.
- Mitigation strategies include broadcast joins, salting keys, and more granular partitioning for high-volume channels.

5.2 New Analysis – Understanding Engagement and Positivity

In the final sprint, we added analysis layers that deepen our understanding of how engagement and sentiment distribute across the platform.

5.2.1 Enhanced Aggregations

Using Spark, we created:

- Top-k tables:
    - Top channels by comment count, average engagement per video.
    - Top videos by total comments and replies.
    - Top authors by number of comments.
- Time-based activity profiles:
    - Comment counts per day and per hour-of-day.
    - Identification of high-activity posting days and peak commenting windows.

These build on earlier Unix work but now scale more naturally with larger datasets.

5.2.2 Keyword and “Lightweight Sentiment” Views

Building on the “great” cluster from Sprint 3:

- Generated frequency tables for positive keywords
- Flagged comments containing specific positive keywords as a basic positivity indicator.
- Created per-video positivity metrics, such as:
    - Number and share of comments containing positive keywords.

While not a full sentiment model, this provides a first, explainable signal of how enthusiastic the comments are for each video.

5.2.3 Interpreting the Results for Leaders

For each new metric, we asked:

- What question does this answer?
    - Example: “Which videos generate both high volume and high positivity?”
- What did we learn?
    - High-engagement hubs are often channels with educational or trending content.
    - Positive keyword clusters tend to appear in these hub videos, reinforcing their value.
- Why does this matter to decision-makers?
    - Leaders can prioritize these videos for recommendations, spot new emergent topics, or replicate successful patterns.

6. Key Findings and Recommendations

Below are the core findings across the project, with implications for decision-makers.

Finding 1 – Engagement Is Highly Centralized

Observation: A small number of videos and channels accumulate disproportionate shares of comments, while most have very few.

Evidence:

- Top video IDs with tens of thousands of comments.
- Long-tail distribution of cluster sizes

Implication:

- The platform is driven by attention hubs.
- Targeting improvements or features at these hubs yields outsized impact.

Recommendation:
Focus early product experiments (comment summaries, sentiment badges, topic tags) on top-engagement videos and channels to maximize user impact.

Finding 2 – Comment Activity Follows a Long-Tailed User Distribution

Observation: Most users post once or twice, while a small set of users post many times.

Evidence:

- Cluster-size histogram

Implication:

- A small group of power commenters act as anchors in discussions.

Recommendation:
Consider features or recognition mechanisms for high-engagement commenters this may sustain discussion quality and help surface more useful comment content.

Finding 3 – Positive Engagement Clusters Around Popular Content

Observation: Positive keyword “great” appeared thousands of times, forming dense clusters around certain videos.

Evidence:

- Frequency tables and visualizations of “great” comment clusters.

Implication:

- Positivity clusters around high-value videos and channels, indicating strong community appreciation.

Recommendation:
Use simple positivity indicators as a first-level signal in recommending videos for learners who need quick, trustworthy content.

---

7. Limitations

7.1 Data Limitations

- Some early analyses were based mainly on IDs (video_id, channel_id) without rich metadata, reducing interpretability.
- The keyword-based positivity measure is simplistic and does not capture sarcasm, mixed sentiment, or non-English comments.
- Comments are self-selected and may not represent the full viewer population.

7.2 Pipeline Limitations

- While our pipeline runs in the cloud, it is not yet a fully automated production system:
    - Manual triggering rather than scheduled jobs + No continuous monitoring/alerting
- Performance tuning is basic but not fully optimized for much larger (10×–100×) datasets.

7.3 Analysis Limitations

- Network analysis focused on comments and replies, not richer interaction types (likes, shares, watch-time logs).

---

8. Future Work

To expand this into a production-quality system, future steps include:

- **Operationalizing the pipeline**:
    - Real-time runs in the cloud with job orchestration and alerts.
    - Clear data quality checks and logging.
- **Richer sentiment and topic modeling**:
    - Use NLP models to detect positive/negative/neutral sentiment and topic clusters beyond simple keywords.
- **User-facing features**:
    - Integrate outputs into dashboards