THE NAS DAILY ANALYSIS JOURNEY: FROM RAW DATA TO STRATEGIC INSIGHTS

A Comprehensive Data Analysis Project

Team Name: Decision Dynamics

Group Members:

1. Abdullah Zafar [2022329]
2. Nabil Orya [2022475]
3. Ijlal Ahmed [2022229]
4. Aqib Muhammad [2022103]

Course Instructor: Sir Hafiz Muhammad Muslim.

Course: Digital Business Analytics.

Institution: Ghulam Ishaq khan Institute of Science and Technology.

Date: 27/12/2025

"Every number tells a story. Every graph answers a question. Every insight drives a decision."

Faculty of Computer Science and Engineering.
GIK Institute of Engineering Sciences and Technology.

Table Of Content

Welcome to the complete journey of transforming YouTube playlist data into actionable business intelligence. This is not just a technical walkthrough—it's a narrative of discovery, where each step reveals something new about content performance, audience behavior, and strategic opportunities.

Chapter 0: The Setup

Before our journey begins, we need our tools and our map.

The Workspace

```
nas_daily_analysis/
├── data/
│   ├── raw/              # Where the story begins—raw data from YouTube
│   │   ├── videos/       # Video metadata treasure trove
│   │   ├── comments/     # The voice of the audience
│   │   └── transcripts/  # The actual words spoken
│   └── processed/        # Where raw data transforms into insights
├── notebooks/            # Our seven chapters of discovery
├── src/                  # The tools that make it all possible
└── Process.md            # This story you're reading now
```

Preparing for the Journey
Activate your environment:

```
cd E:\nas_daily_analysis
.\.venv\Scripts\Activate.ps1
```

Your essential tools (install if needed):

- The data collectors: `pandas`, `google-api-python-client`, `youtube-transcript-api`
- The timekeepers: `isodate`, `tqdm`
- The pattern finders: `scikit-learn`, `transformers`, `spacy`, `nltk`
- The storytellers: `matplotlib`, `seaborn`

Your secret key:

Make sure `src/config.py` holds your YouTube Data API v3 key—this is your passport to the data.

Chapter 1: Gathering the Foundation Stones

Notebook: `01_data_collection.ipynb`

The Story: Our first step into the unknown—collecting the raw materials that will become our insights.

The Quest Begins
Imagine you're an explorer, about to map an entire playlist. The YouTube Data API is your telescope, and each video is a distant star you need to observe.

What happens here:

1. **Connecting to the Source**

   - We establish a connection to YouTube's vast database

   - Using our API key, we authenticate ourselves as authorized data collectors

2. **The Great Collection**

   - We point our tools at the playlist: `PLwem0A53mZRr9QzyydE_6Mtk2uTEcc1UX`

   - Like gathering stars, we collect every video ID, one constellation at a time

   - For each video, we capture its essence:

     - **Identity:** `video_id`, `title`, `published_at`

     - **Impact:** `views`, `likes`, `comments`

     - **Temporal fingerprint:** `duration` (that magical ISO 8601 format)

3. **Securing the Treasures**

   - Each piece of data is carefully stored in `data/raw/videos/only_1_minute_videos.csv`

   - This becomes our foundational dataset—everything else builds upon this

The Output:

```
data/raw/videos/only_1_minute_videos.csv
```

Key Columns Captured:

- `video_id` - The unique identifier
- `title` - What the video promises
- `published_at` - When it entered the world
- `views` - How many eyes saw it
- `likes` - The approval signals
- `comments` - The conversation starters
- `duration` - The temporal footprint

What We've Learned: We now have the basic facts—who, what, when, and the initial metrics. But this is just the surface. The real story lies deeper.

Chapter 2: Listening to the Voices

Notebook: `02_comments_collection.ipynb`

The Story: Numbers tell us what happened, but comments tell us how people felt about it.

Entering the Conversation

With our video catalog complete, we turn to the most human element: what did people actually say? Comments are where passive viewers become active participants. This is where we discover not just what worked, but why.

The Process:

4. **Opening the Dialogue**

- We load our video collection from Chapter 1

- Each video ID becomes a door we knock on

- Behind each door: conversations, reactions, emotions

5. **The Comment Harvest**

- Using our `fetch_comments()` function (a trusted tool from `src/comments.py`)

- We visit each video, collecting up to 500 comments per video

- For each comment, we capture:

  - **Who spoke:** `author`, `author_channel_id`

  - **What they said:** `comment_text`

  - **How it resonated:** `like_count`

  - **When they spoke:** `published_at`

6. **The Graceful Failures**

- Some videos have disabled comments—we note this and move on

- Not every door opens, but we respect those boundaries

- Our code handles errors with dignity, logging what couldn't be collected

The Output:

`data/raw/comments/only_1_minute_comments.csv`

What We've Gathered:

- `video_id` - Which video sparked the conversation
- `comment_id` - Each unique voice
- `author` - The person behind the words
- `author_channel_id` - Their digital identity
- `comment_text` - The actual sentiment
- `like_count` - Community agreement level
- `published_at` - The moment they spoke

What We've Learned: Numbers are cold. Comments are warm. We now have both—the quantitative foundation and the qualitative texture. The real magic happens when we combine them.

Chapter 3: Crafting the Metrics

Notebook: `03_feature_engineering.ipynb`

The Story: Raw data is like raw ore—valuable, but you need to refine it to see its true worth. This is where we transform facts into insights.

The Transformation
Imagine you have pieces of a puzzle. Each piece (raw data point) is important, but they don't show the full picture until you arrange them meaningfully. Feature engineering is our way of arranging these pieces to reveal patterns.

Act 1: Video Metrics Evolution
From Duration to Efficiency:

The journey begins with time itself. That cryptic ISO 8601 duration string (`PT1M5S`) becomes something we can actually calculate with: seconds. This simple transformation unlocks everything.

The Metrics We Forge:

7. `duration_sec` - Time translated into numbers

 - `PT1M5S` becomes `65.0` seconds

 - Now we can measure everything per second

8. `engagement_rate` - The heart of interaction

 - Formula: `(likes + comments) / views`

 - This answers: "Of everyone who watched, how many actually engaged?"

 - High rate = strong connection; Low rate = passive consumption

9. `like_rate` - The approval percentage

 - Formula: `likes / views`

- Measures positive sentiment at scale

10. `comment_rate`- The conversation indicator

- Formula: `comments / views`

- Shows which videos spark discussion

11. `views_per_second` - The attention efficiency metric

- Formula: `views / duration_sec`

- Reveals: Does the 1-minute format maximize attention per second?

12. `comments_per_second` - The engagement velocity

- Formula: `comments / duration_sec`

- Shows conversation intensity

Feature	Description
`engagement_rate`	`(likes + comments)/views`
`like_rate`	`likes/views`
`comment_rate`	`comments/views`
`duration_sec`	Convert ISO 8601 `duration` to seconds
`views_per_second`	`views/duration_sec`
`comments_per_second`	`comments/duration_sec`

Graph 3.1: Video Features Summary Table

- Location: After video feature engineering
- Shows: Sample rows with all new metrics
- Purpose: Verify calculations and see transformed data

The Output:

```
data/processed/video_features.csv
```

Act 2: Understanding the Audience

From Comments to Community:

Comments are individual voices, but together they form a chorus. We need to understand both the individual and the collective.

What We Build:

13. `comment_length` - The depth of expression

  - Word count per comment

  - Long comments = deep engagement; Short comments = quick reactions

14. **Top Commenters Analysis** - The community architects

  - Who comments most frequently?

  - Are there power users driving the conversation?

  - This reveals community structure

Feature	Description
`comment_length`	Number of words in comment
`top_commenters`	Count of comments per user
`like_per_comment`	Like count on comment

Graph 3.2: Top Commenters Bar Chart

- Location: After top commenters calculation
- Shows: Top 20 commenters by comment count
- Purpose: Visualize community concentration

The Outputs:

`data/processed/comment_features.csv`

`data/processed/top_commenters.csv`

Act 3: The Words Themselves

From Transcripts to Linguistic Patterns:

Note: This section requires manual transcript data at `data/raw/transcripts/only_1_minute_manual.csv`*

Transcripts are the blueprint of content. Every word was chosen, every sentence crafted. We analyze this craftsmanship.

The Linguistic Metrics:

15. `word_count` - Content volume

  - Total words in transcript

  - More words = more information density

16. `sentence_count` - Narrative structure

  - Sentences separated by `.`, `?`, `!`

  - Shows pacing and structure

17. `words_per_second` - Speaking rate

  - Formula: `word_count / duration_sec`

  - Measures information density over time

  - Too fast = overwhelming; Too slow = boring

Feature	Description
`word_count`	Total words in transcript
`sentence_count`	Total sentences (split by `.`, `?`, `!`)
`words_per_second`	`word_count/duration_sec`

Graph 3.3: Transcript Features Summary

- Location: After transcript feature engineering
- Shows: Word count, sentence count, words_per_second distributions
- Purpose: Understand content linguistic characteristics

The Output:

```
data/processed/transcript_features.csv
```

What We've Learned: Raw data was just the beginning. We've now created a language of metrics, a vocabulary that lets us ask deeper questions. Every metric is a lens through which we can view the data differently.

Chapter 4: The First Explorations

Notebook: `04_exploratory_analysis.ipynb`

The Story: With our refined metrics in hand, we begin our first real explorations. Every graph we create answers a question. Every question matters.

The Philosophy

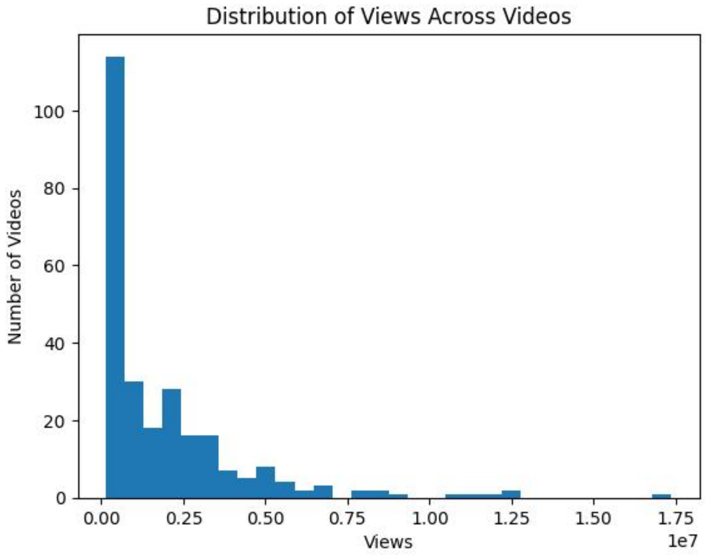
> *"Every plot must answer a business question. If a plot does not inform a decision, it does not belong in a professional analysis."*

We're not just making graphs for the sake of graphs. Each visualization tells part of the story. Each one reveals something that changes how we think about content strategy.

Question 1: "How uneven is attention across videos?"

The Investigation:

We create a histogram of views distribution. This simple visualization reveals a fundamental truth about digital content: the distribution is rarely fair.



Graph 4.1: Distribution of Views Across Videos

- Type: Histogram
- X-axis: Views
- Y-axis: Number of Videos
- Title: "Distribution of Views Across Videos"

The Story It Tells:

- Right-skewed distribution → hit-driven strategy (few blockbusters, many underperformers)

- Narrow distribution → predictable, consistent performance

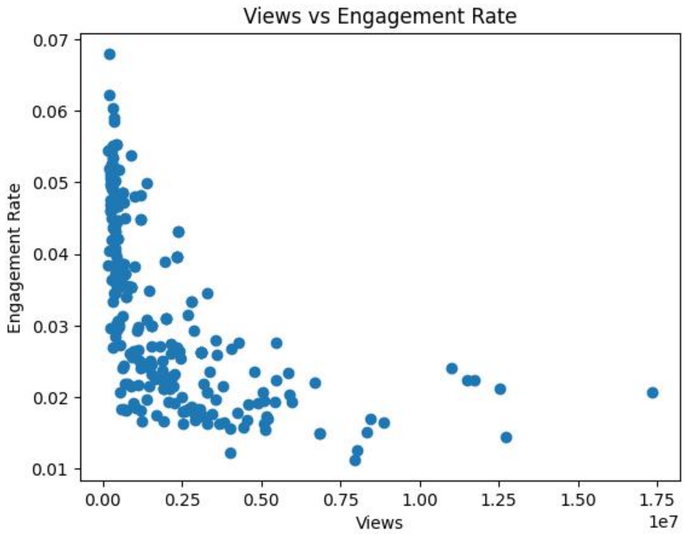
- This answers: "Should we rely on viral hits or build consistency?"

The Business Insight:

If the curve is steep and right-skewed, you're in a hit-driven world. A few videos carry the weight. If it's flatter, you have consistency. Both are strategies, but they require different approaches.

Question 2: "Are views translating into interaction?"
The Investigation:

Views tell us reach, but engagement tells us connection. We scatter views against engagement rate to see if bigger audiences mean better engagement.



Graph 4.2: Views vs Engagement Rate

- Type: Scatter Plot
- X-axis: Views
- Y-axis: Engagement Rate
- Title: "Views vs Engagement Rate"

The Story It Tells:

- High views + low engagement → passive consumption (wide reach, shallow connection)

- Moderate views + high engagement → community value (targeted reach, deep connection)

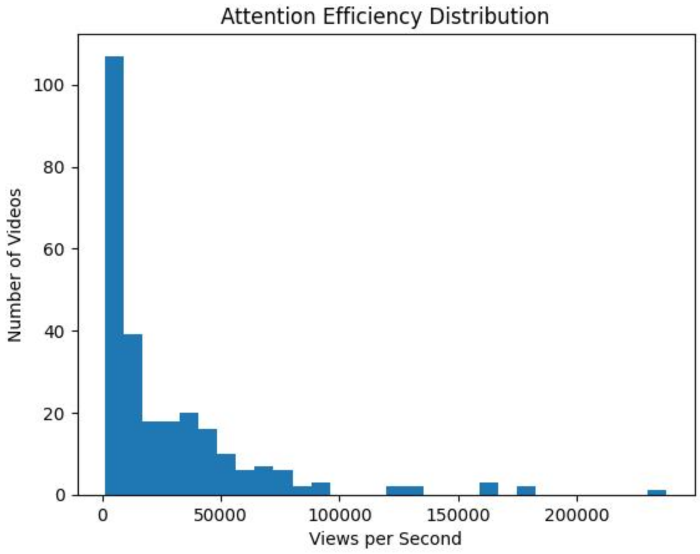
- This reveals: "Which videos create true fans vs. casual viewers?"

The Business Insight:

Not all views are created equal. A video with 1M views and 0.5% engagement rate is different from 100K views with 5% engagement rate. The second has a stronger community connection.

Question 3: "Is the 1-minute format actually efficient?"
The Investigation:

The core value proposition: maximum value in minimum time. We test this by looking at views per second—does the format deliver disproportionate attention?



Graph 4.3: Attention Efficiency Distribution

- Type: Histogram
- X-axis: Views per Second
- Y-axis: Number of Videos
- Title: "Attention Efficiency Distribution"

The Story It Tells:

- High views/sec → strong format-product fit (the format works!)

- Wide variance → topic matters more than format (format alone isn't enough)

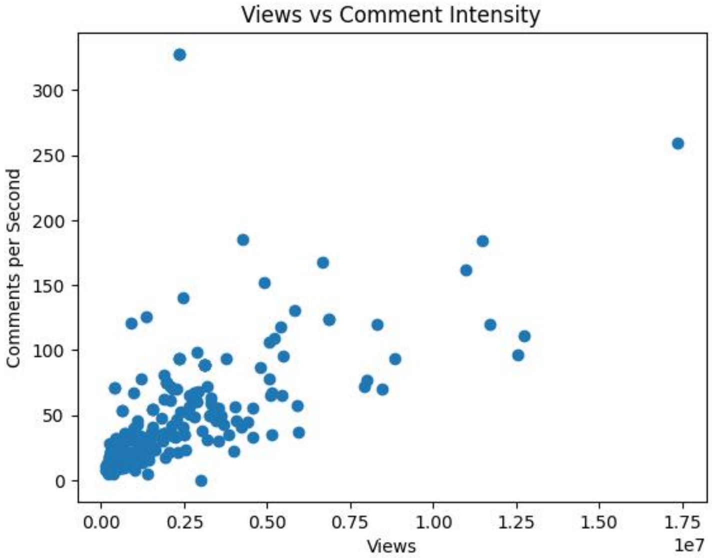
- This validates: "Is the 1-minute constraint an advantage or limitation?"

The Business Insight:

If most videos have high views-per-second, the format itself is valuable. If there's huge variance, then content quality matters more than format efficiency.

Question 4: "Which videos generate conversation, not just clicks?"
The Investigation:

Comments are the gold standard of engagement—they represent active, not passive, viewers. We plot views against comments per second.



Graph 4.4: Views vs Comment Intensity

- Type: Scatter Plot
- X-axis: Views
- Y-axis: Comments per Second
- Title: "Views vs Comment Intensity"

The Story It Tells:

- Some videos invite discussion independent of views

- These are high brand affinity assets—they create community

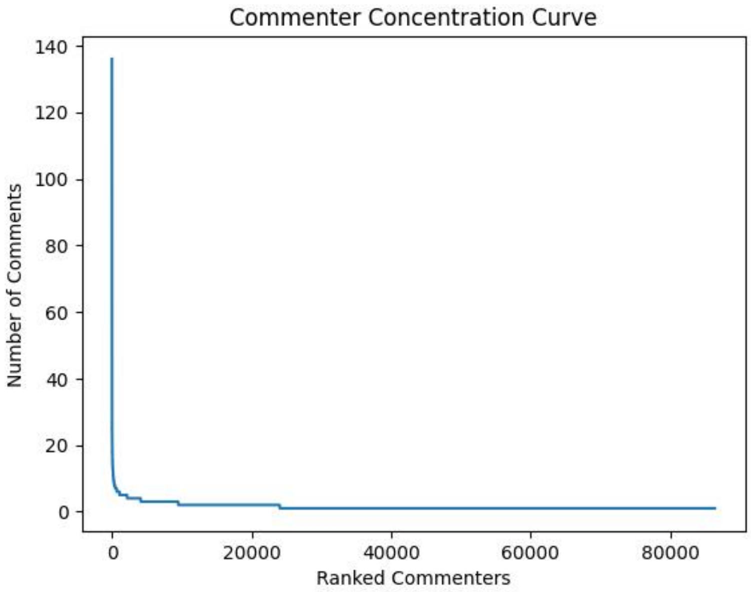
- This identifies: "Which videos turn viewers into participants?"

The Business Insight:

Videos that generate comments disproportionate to views are special. They create community. They build brand. They're worth more than their view count suggests.

Question 5: "Is engagement broad or driven by a few users?"
The Investigation:

Healthy communities have many voices. Unhealthy ones have few dominant voices. We plot the commenter concentration curve.



Graph 4.5: Commenter Concentration Curve

- Type: Line Plot
- X-axis: Ranked Commenters (by comment count)
- Y-axis: Number of Comments
- Title: "Commenter Concentration Curve"

The Story It Tells:

- Steep curve → small group dominates (unhealthy concentration)

- Flat curve → diverse audience participation (healthy community)

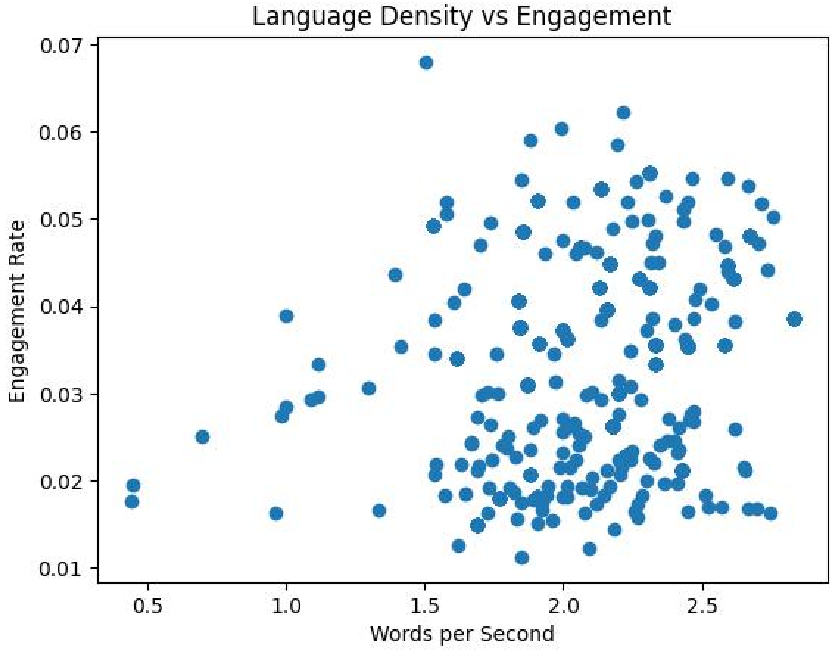
- This measures: "Is our community democratic or oligarchic?"

The Business Insight:

A steep drop-off means you're relying on superfans. That's fine, but fragile. A flatter curve means broader engagement—more resilient, more sustainable.

Question 6: "Does faster speech reduce or increase engagement?"
The Investigation:

Information density is a double-edged sword. Too much = cognitive overload. Too little = boredom. We plot words per second against engagement rate.



Graph 4.6: Language Density vs Engagement

- Type: Scatter Plot
- X-axis: Words per Second
- Y-axis: Engagement Rate
- Title: "Language Density vs Engagement"

The Story It Tells:

- Bell-shaped curve → optimal pacing exists (there's a sweet spot!)

- Negative trend → overloading viewers (more isn't better)

- Positive trend → viewers want more information

- This reveals: "What's the optimal speaking pace?"

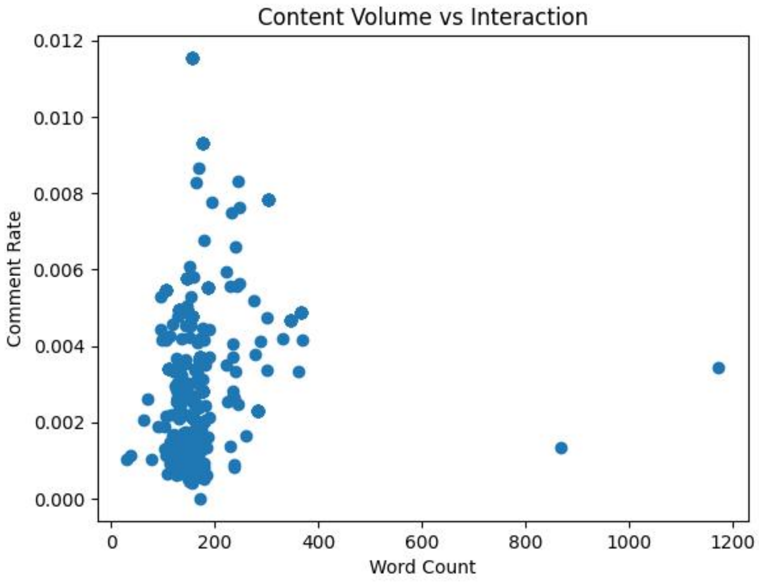
The Business Insight:

There's likely an optimal range. Too fast = viewers can't process. Too slow = viewers lose interest. The data will show us where that sweet spot is.

Question 7: "Does more content lead to more interaction?"

The Investigation:

Does word count correlate with engagement? Is more always better?



Graph 4.7: Content Volume vs Interaction

- Type: Scatter Plot
- X-axis: Word Count
- Y-axis: Comment Rate
- Title: "Content Volume vs Interaction"

The Story It Tells:

- Higher word count with lower interaction → diluted messaging (more words, less impact)

- Optimal range → efficient storytelling (quality over quantity)

- This answers: "Should we pack more words or focus on clarity?"

The Business Insight:

More words don't guarantee more engagement. In fact, they might dilute your message. There's likely an optimal word count range that maximizes interaction without overwhelming.

What We've Learned: Every visualization answered a question. Every question matters. We're not just observing patterns—we're building a strategic understanding of what works, what doesn't, and why.

Chapter 5: The Strategic Synthesis

Notebook: `05_strategic_analysis.ipynb`

The Story: Now we move from observation to action. Every analysis here answers: "If I were Nas Daily's content or growth team, what would I do differently?"

The Transformation from Analysis to Strategy
Exploratory analysis showed us what exists. Strategic analysis shows us what to do about it. This is where insights become actions.

Strategy 1: Content Archetype Identification
The Big Question: "Are all 1-minute videos the same from a business perspective?"

The answer is no. Even within one playlist, videos fall into distinct performance archetypes. Identifying these is like creating a taxonomy of success.

The Method:

We use K-Means clustering to group videos by behavior patterns. Three key metrics define our clusters:

- `views_per_second` - Attention efficiency
- `engagement_rate` - Connection strength
- `comment_rate` - Conversation generation

The Process:

18. We standardize the features (because views and rates are on different scales)
19. We run K-Means with 4 clusters (the number reveals itself through analysis)
20. Each video gets assigned to a cluster
21. We interpret what each cluster means

Graph 5.1: Content Cluster Visualization

- Type: 3D Scatter Plot (or 2D projections)
- Axes: views_per_second, engagement_rate, comment_rate
- Colors: Cluster assignments
- Title: "Content Performance Archetypes"

The Story It Tells:

- Cluster 0: Viral reach drivers (high views, moderate engagement)

- Cluster 1: Community discussion starters (high comments, high engagement)

- Cluster 2: Consistent performers (balanced across metrics)

- Cluster 3: Underperformers (low across all metrics)

Graph 5.2: Cluster Comparison Bar Chart

- Type: Grouped Bar Chart
- X-axis: Cluster ID

- Y-axis: Average metric values (normalized)
- Series: views_per_second, engagement_rate, comment_rate
- Title: "Average Metrics by Content Cluster"

The Story It Tells: Clear visual comparison of cluster characteristics

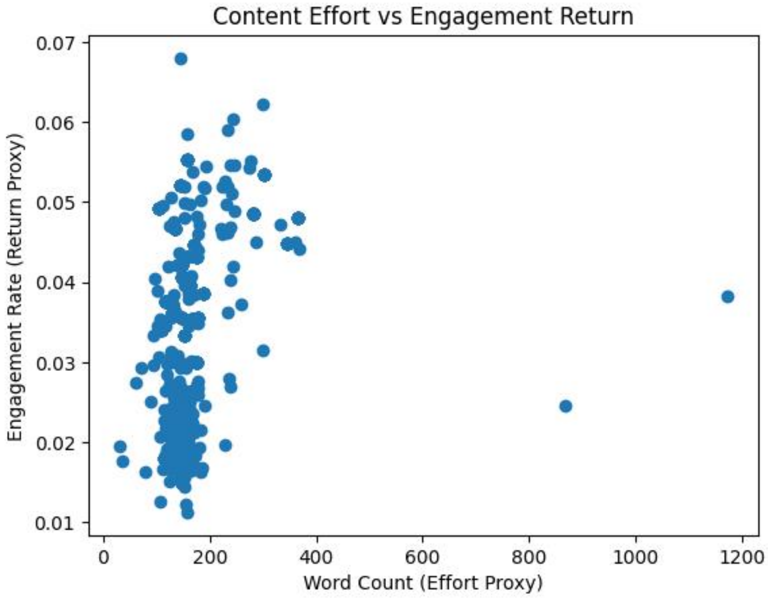
The Strategic Output:

- Each video labeled with its archetype
- Templates for content creation based on successful archetypes
- Promotion strategies tailored to each type

Strategy 2: Effort vs Return Analysis

The Big Question: "Where should we invest our content creation effort?"

Content creation costs time and energy. We need to know where that investment pays off.



Graph 5.3: Content Effort vs Engagement Return

- Type: Scatter Plot
- X-axis: Word Count (Effort Proxy)
- Y-axis: Engagement Rate (Return Proxy)
- Title: "Content Effort vs Engagement Return"

The Story It Tells:

- Upper right quadrant → high effort, high return (worth the investment)

- Upper left quadrant → low effort, high return (efficiency winners!)

- Lower right quadrant → high effort, low return (rethink these)

- Lower left quadrant → low effort, low return (filler content)

- Trend line shows if more words = more engagement (or not)

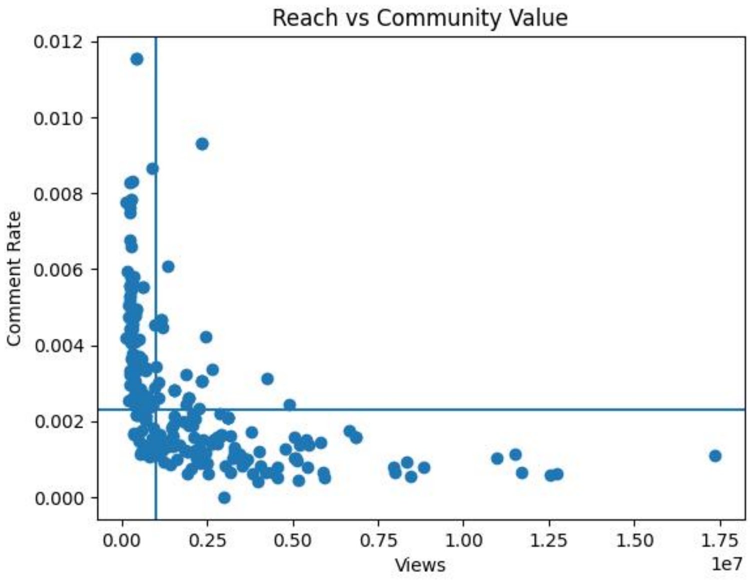
The Business Insight:

If there's a positive correlation, invest more words in high-engagement topics. If there's a negative correlation or no correlation, focus on clarity and efficiency over volume.

Strategy 3: Reach vs Community Value

The Big Question: "How do we balance wide reach with deep community building?"

Not every video needs to go viral. Some videos should build community, even if they don't reach millions.



Graph 5.4: Reach vs Community Value

- Type: Scatter Plot with Reference Lines
- X-axis: Views
- Y-axis: Comment Rate
- Features: Median lines (horizontal and vertical)
- Title: "Reach vs Community Value"

The Story It Tells:

- Quadrant I (high views, high comments) → Best of both worlds

- Quadrant II (low views, high comments) → Community builders (valuable for engagement)

- Quadrant III (low views, low comments) → Need improvement

- Quadrant IV (high views, low comments) → Passive consumption (good for reach, weak on community)

The Business Insight:

Different videos serve different purposes. Some should maximize reach. Others should maximize community value. Both are valid strategies, but they require different approaches.

Strategy 4: Speaking Pace Optimization
The Big Question: "What speaking rate maximizes engagement?"

We analyze engagement by words-per-second bins to find the optimal pacing.

Graph 5.5: Engagement Rate by Speaking Pace

- Type: Bar Chart or Line Plot
- X-axis: Words per Second (binned into ranges)
- Y-axis: Average Engagement Rate
- Title: "Optimal Speaking Pace Analysis"
- **The Story It Tells:**

- Peak engagement at specific pace range → optimal pacing identified

- Steep drop-offs → pace matters significantly

- Flat curve → pace doesn't matter much

The Business Insight:

There's likely a sweet spot. Too slow = boring. Too fast = overwhelming. The data reveals where that sweet spot is.

Strategy 5: Audience Concentration Metrics
The Big Question: "How concentrated is our community engagement?"

We calculate what percentage of comments come from the top 10 commenters.

Graph 5.6: Top Commenters Share Analysis

- Type: Pie Chart or Stacked Bar
- Shows: Share of comments from top 10 vs. rest of community
- Title: "Comment Distribution: Top 10 vs. Community"

The Story It Tells:

- High concentration (e.g., >50% from top 10) → fragile community

- Low concentration (<20% from top 10) → healthy, diverse community

The Business Insight:

High concentration means you're dependent on superfans. That's valuable, but risky. Broad distribution means a more resilient community.

What We've Learned: We're no longer just observing—we're strategizing. Every analysis points to a decision. Every decision improves content strategy.

Chapter 6: The Language of Content

Notebook: `06_contextual_transcript_analysis.ipynb`

The Story: Beyond metrics lie the words themselves. What linguistic patterns create success? How does the actual language used affect engagement?

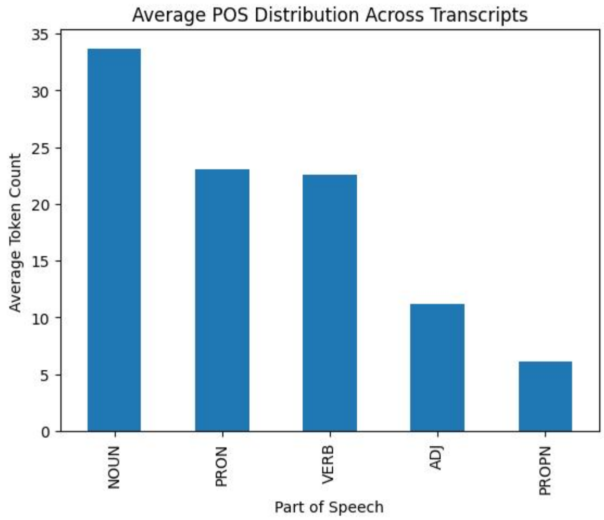
The Deep Dive into Language

While metrics tell us what happened, linguistics tells us how it happened. This chapter explores the actual words, structures, and patterns that make content work.

Linguistic Discovery 1: Narrative Style Analysis

The Investigation:

Every speaker has a style. Some use more nouns (descriptive), some use more verbs (action-oriented). We analyze Part-of-Speech distributions.



Graph 6.1: Average POS Distribution Across Transcripts

- Type: Bar Chart
- X-axis: Part of Speech Tags (NOUN, VERB, ADJ, ADV, etc.)
- Y-axis: Average Token Count

- Title: "Average POS Distribution Across Transcripts"
- **The Story It Tells:**

- Noun-heavy → descriptive, informative style

- Verb-heavy → action-oriented, dynamic style

- Adjective-heavy → emotional, evocative style

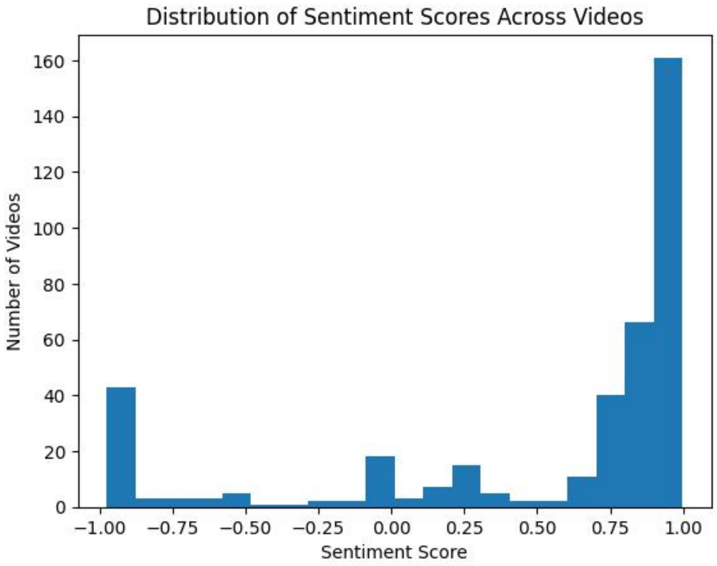
- This reveals: "What linguistic style defines the content?"

The Business Insight:

Different styles resonate differently. Action-oriented might work for tutorials. Descriptive might work for storytelling. The data shows which style correlates with success.

Linguistic Discovery 2: Sentiment Patterns
The Investigation:

Using VADER (optimized for social media), we analyze the sentiment of each transcript. Is positive sentiment correlated with performance?



Graph 6.2: Distribution of Sentiment Scores Across Videos

- Type: Histogram
- X-axis: Sentiment Score (-1 to 1)
- Y-axis: Number of Videos
- Title: "Distribution of Sentiment Scores Across Videos"

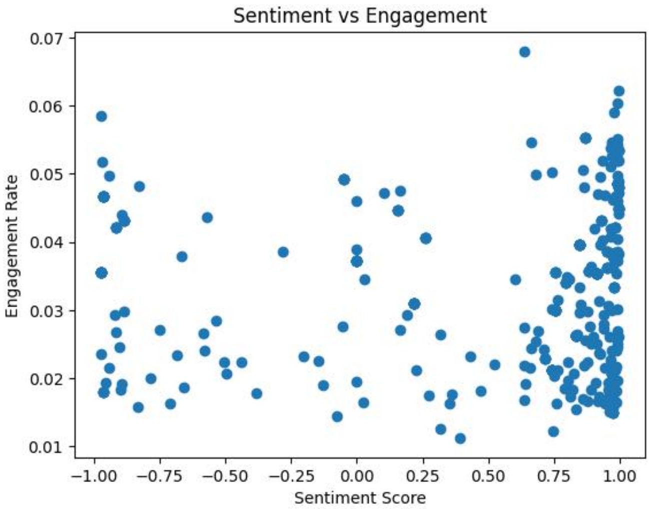
- **The Story It Tells:**

- Positive skew → optimistic content overall

- Negative skew → serious, critical content

- Bimodal → mix of tones

- This shows: "What's the emotional tone of the content?"



Graph 6.3: Sentiment vs Engagement

- Type: Scatter Plot
- X-axis: Sentiment Score
- Y-axis: Engagement Rate
- Title: "Sentiment vs Engagement"
- **The Story It Tells:**

- Positive correlation → positive content performs better

- Negative correlation → serious/critical content resonates

- No correlation → sentiment doesn't affect engagement

- This answers: "Does being positive help or hurt?"

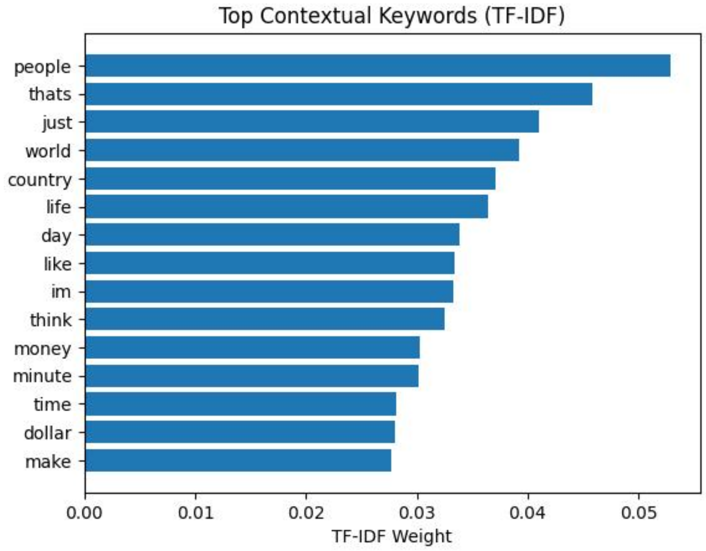
The Business Insight:

If positive sentiment correlates with engagement, lean into optimism. If not, authenticity matters more than forced positivity.

Linguistic Discovery 3: Keyword Extraction

The Investigation:

What words define successful content? TF-IDF analysis reveals the keywords that make content stand out.



Graph 6.4: Top Contextual Keywords (TF-IDF)

- Type: Horizontal Bar Chart
- X-axis: TF-IDF Weight
- Y-axis: Keywords (top 20-30)
- Title: "Top Contextual Keywords (TF-IDF)"

The Story It Tells:

- High TF-IDF words = distinctive, important terms

- These are the words that make content unique

- This reveals: "What language makes content distinctive?"

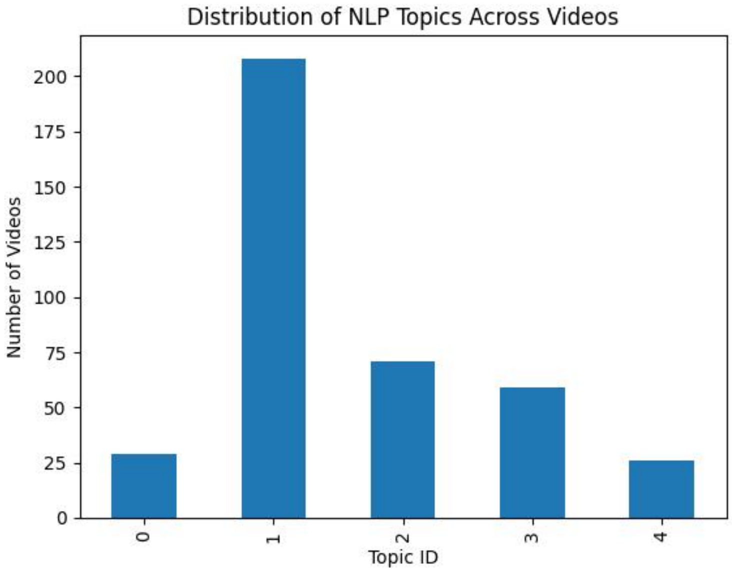
The Business Insight:

Keywords that appear in high-performing videos but not in others are your secret sauce. These are the terms that resonate with your audience.

Linguistic Discovery 4: Topic Modeling

The Investigation:

What are the underlying topics in the content? LDA or BERTopic reveals thematic clusters.



Graph 6.5: Distribution of NLP Topics Across Videos

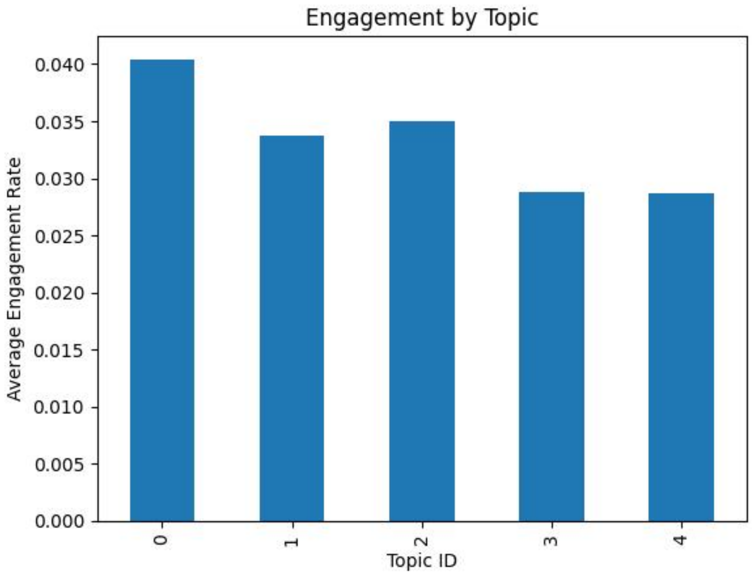
- Type: Bar Chart or Histogram
- X-axis: Topic ID
- Y-axis: Number of Videos
- Title: "Distribution of NLP Topics Across Videos"

The Story It Tells:

- Even distribution → diverse content

- Concentrated → focused content strategy

- This shows: "What topics dominate the playlist?"



Graph 6.6: Engagement by Topic

- Type: Bar Chart
- X-axis: Topic ID
- Y-axis: Average Engagement Rate
- Title: "Engagement by Topic"
- **The Story It Tells:**

- High-engagement topics → focus more on these

- Low-engagement topics → reconsider or improve

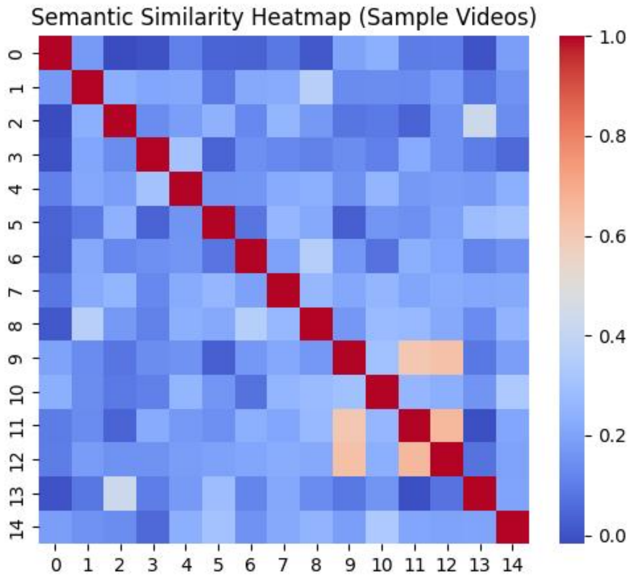
- This answers: "Which topics resonate most?"

The Business Insight:

Some topics are winners. Others need work. Topic modeling shows you where to focus your content creation efforts.

Linguistic Discovery 5: Semantic Similarity
The Investigation:

Which videos are semantically similar? This reveals content redundancy and opportunities.



Graph 6.7: Semantic Similarity Heatmap

- Type: Heatmap
- Axes: Video IDs (sample of 20-30 videos)
- Color intensity: Similarity score
- Title: "Semantic Similarity Heatmap (Sample Videos)"
- **The Story It Tells:**

- Dense clusters → similar content (maybe too similar?)

- Isolated videos → unique content

- This reveals: "Are we repeating ourselves or staying diverse?"

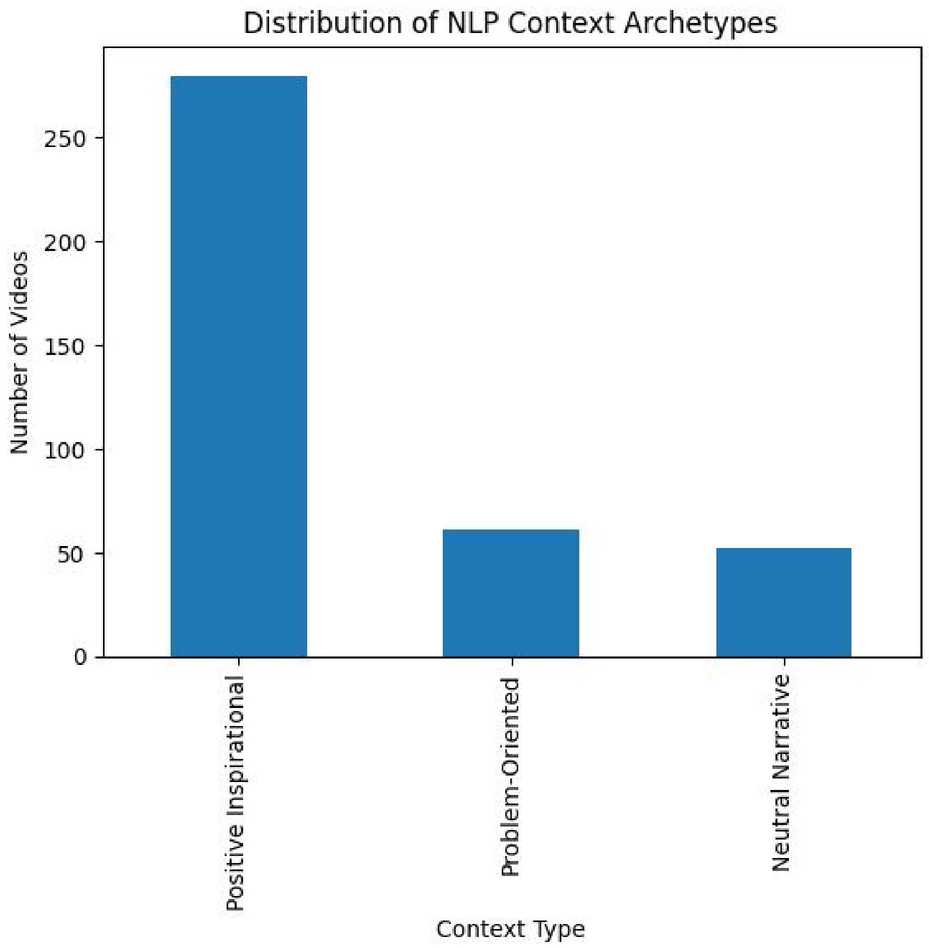
The Business Insight:

High similarity might mean you're finding your voice (good) or repeating yourself (bad). Low similarity means diversity (good) or lack of focus (could be bad).

Linguistic Discovery 6: Context Archetypes

The Investigation:

Similar to content archetypes, but based on linguistic features. What linguistic patterns define successful content?

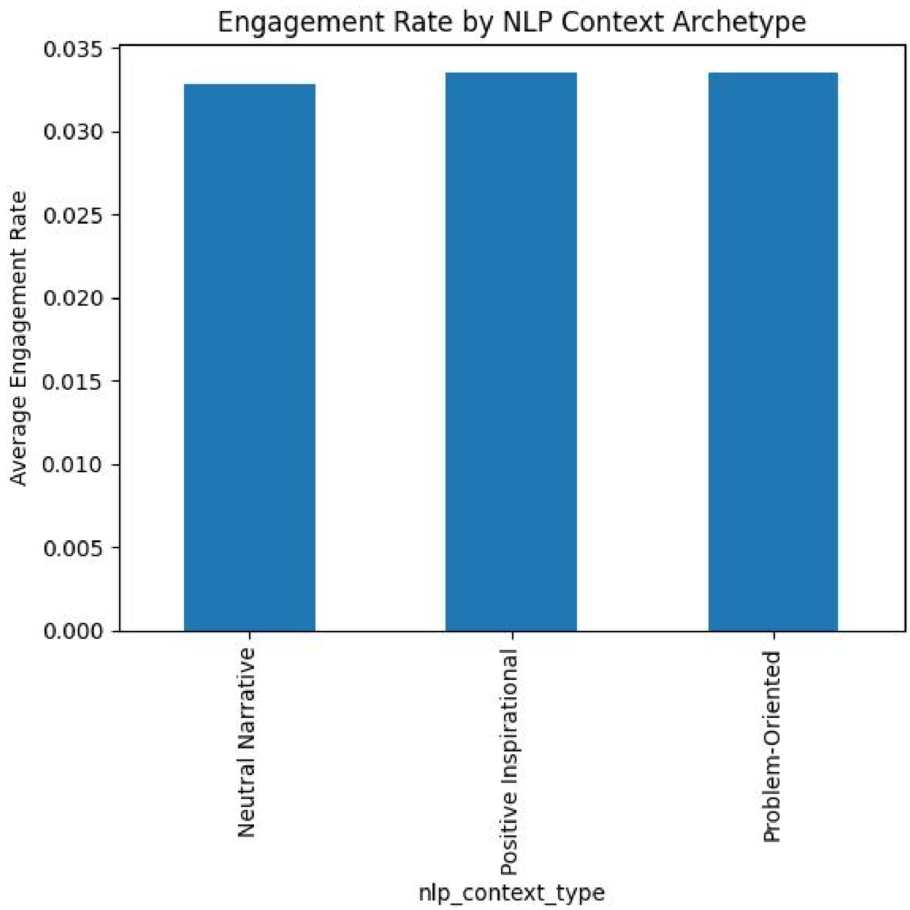


Graph 6.8: Distribution of NLP Context Archetypes

- Type: Bar Chart
- X-axis: Context Type (archetype labels)
- Y-axis: Number of Videos
- Title: "Distribution of NLP Context Archetypes"
- **The Story It Tells:**

- Shows which linguistic styles are most common

- Reveals if there's a dominant style or healthy diversity



Graph 6.9: Engagement Rate by NLP Context Archetype

- Type: Bar Chart
- X-axis: Context Type
- Y-axis: Average Engagement Rate
- Title: "Engagement Rate by NLP Context Archetype"
- **The Story It Tells:**

- Which linguistic archetypes perform best

- Clear guidance on which style to emulate

What We've Learned: Language matters. Not just what you say, but how you say it. The linguistic patterns that correlate with success become our content creation template.

Chapter 7: The AI Revolution

Notebook: `07_ai_based_sentiment_emotion_topic_intelligence_using_nlp_model.ipynb`

The Story: We've used statistics and linguistics. Now we unleash AI—pretrained models that understand sentiment and emotion at a depth beyond traditional analysis.

The Power of Transformers

This is where we go beyond what humans can easily observe. AI models trained on millions of examples see patterns we might miss. They understand nuance, context, and emotional subtext.

AI Analysis 1: Video-Level Sentiment

The Tools:

- **Sentiment Model:** `distilbert-base-uncased-finetuned-sst-2-english`

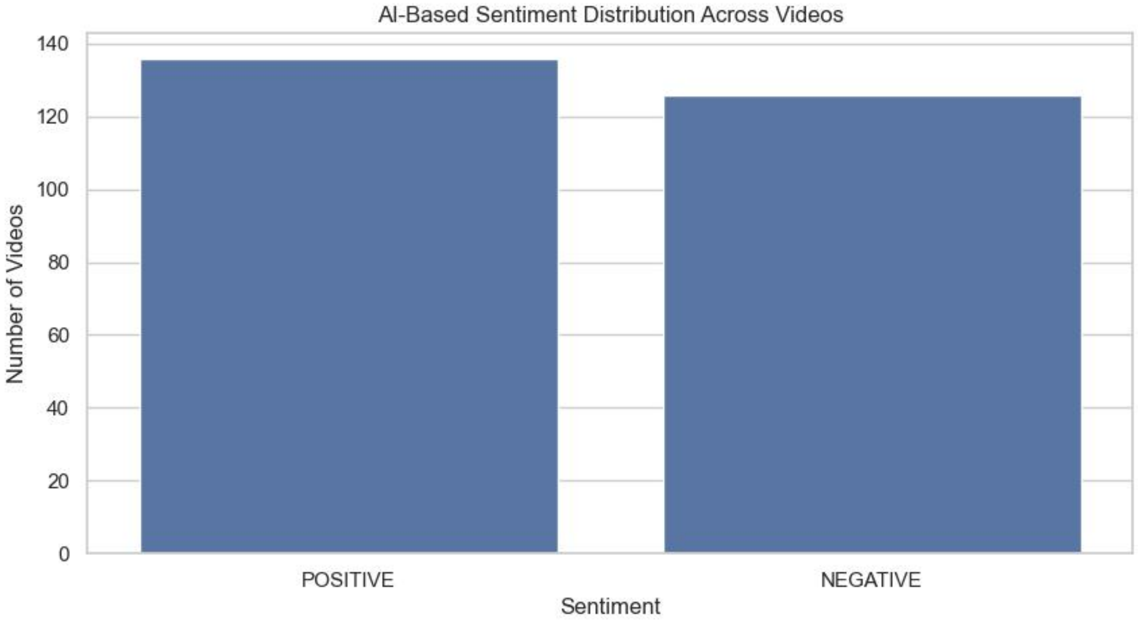
- Binary classification: POSITIVE or NEGATIVE

- Confidence scores (0 to 1)

- Optimized for social media text

The Process:

Each transcript is fed through the model. The AI doesn't just count positive words—it understands context, nuance, and implied sentiment.



Graph 7.1: AI-Based Sentiment Distribution Across Videos

- Type: Count Plot / Bar Chart
- X-axis: Sentiment Label (POSITIVE / NEGATIVE)
- Y-axis: Number of Videos
- Title: "AI-Based Sentiment Distribution Across Videos"
- **The Story It Tells:**

- Proportion of positive vs. negative videos

- Overall emotional tone of the playlist

- This reveals: "Is the content generally positive or does it tackle serious topics?"

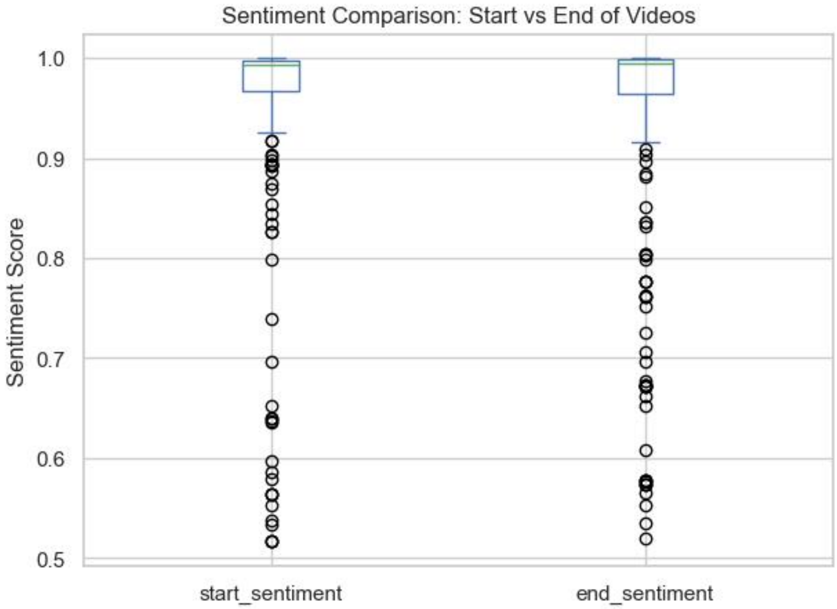
Technical Note: The model handles long transcripts intelligently, using proper token truncation (400 words ≈ 520 tokens) to stay within model limits while preserving meaning.

AI Analysis 2: Sentiment Evolution
The Investigation:

How does sentiment change throughout a video? Does it start positive and end positive? Or is there a narrative arc?

The Method:

- Analyze first 20% of transcript (the hook)
- Analyze last 20% of transcript (the conclusion)
- Compare start vs. end sentiment



Graph 7.2: Sentiment Comparison: Start vs End of Videos

- Type: Box Plot (side by side)
- Series: Start Sentiment Scores, End Sentiment Scores
- Title: "Sentiment Comparison: Start vs End of Videos"
- **The Story It Tells:**

- If start > end → videos get more serious/conclusive

- If end > start → videos build to positive conclusion

- If similar → consistent tone throughout

- This reveals: "What's the narrative arc pattern?"

The Business Insight:

Videos that start positive and end positive might be feel-good content. Videos that start neutral and end positive might be problem-solution narratives. The pattern matters.

AI Analysis 3: Audience Emotion Intelligence
The Tools:

- **Emotion Model:** `j-hartmann/emotion-english-distilroberta-base`

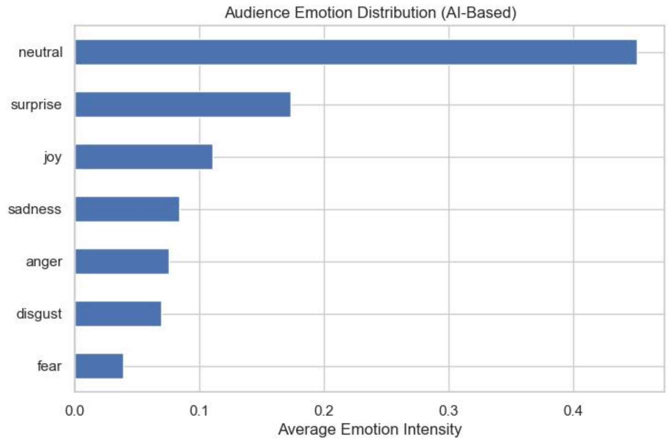
- 7 emotions: anger, disgust, fear, joy, neutral, sadness, surprise

- Probability scores for each emotion

- Understands emotional nuance beyond simple positive/negative

The Process:

Every comment is analyzed. Not just "is this positive?" but "what emotion is this person feeling?"



Graph 7.3: Audience Emotion Distribution (AI-Based)

- Type: Horizontal Bar Chart
- X-axis: Average Emotion Intensity
- Y-axis: Emotions (anger, disgust, fear, joy, neutral, sadness, surprise)
- Title: "Audience Emotion Distribution (AI-Based)"
- **The Story It Tells:**

- Which emotions dominate audience reactions

- Joy high + fear low → positive engagement

- Surprise high → content is unexpected/engaging

- Anger high → controversial content (might be intentional)

- This reveals: "What emotions does our content evoke?"

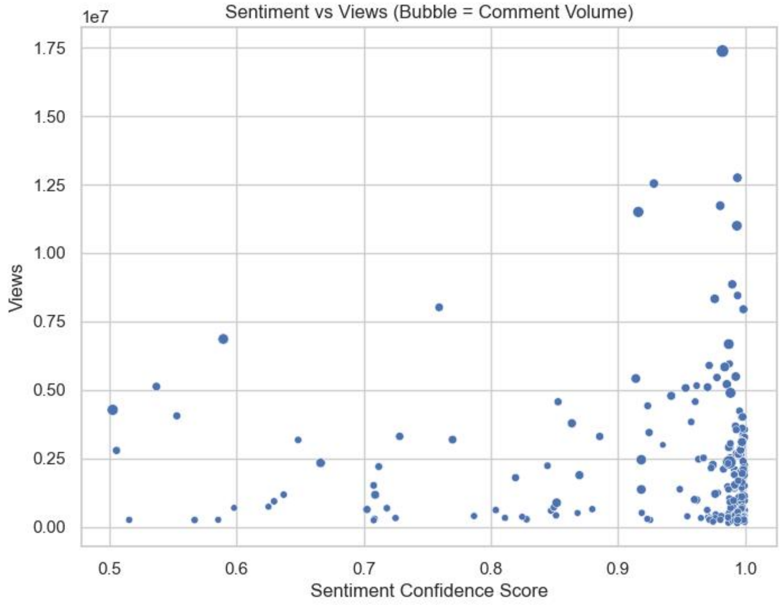
The Business Insight:

Different emotions drive different behaviors. Joy might drive shares. Surprise might drive comments. Understanding emotional impact helps optimize content for desired outcomes.

AI Analysis 4: Sentiment-Performance Correlation

The Investigation:

Now we connect AI insights to business metrics. Does AI-detected sentiment correlate with views and engagement?



Graph 7.4: Sentiment vs Views (Bubble = Comment Volume)

- Type: Scatter Plot with Size Encoding
- X-axis: Sentiment Confidence Score
- Y-axis: Views
- Size: Comment Count (bubble size)
- Title: "Sentiment vs Views (Bubble = Comment Volume)"
- **The Story It Tells:**

- Positive correlation → positive sentiment = more views

- Negative correlation → serious/negative content performs better

- No correlation → sentiment doesn't drive views

- Bubble size shows which videos generate discussion

- This answers: "Does AI sentiment predict performance?"

The Business Insight:

If positive sentiment correlates with views, lean into positivity. If not, authenticity and topic selection matter more than forced positivity.

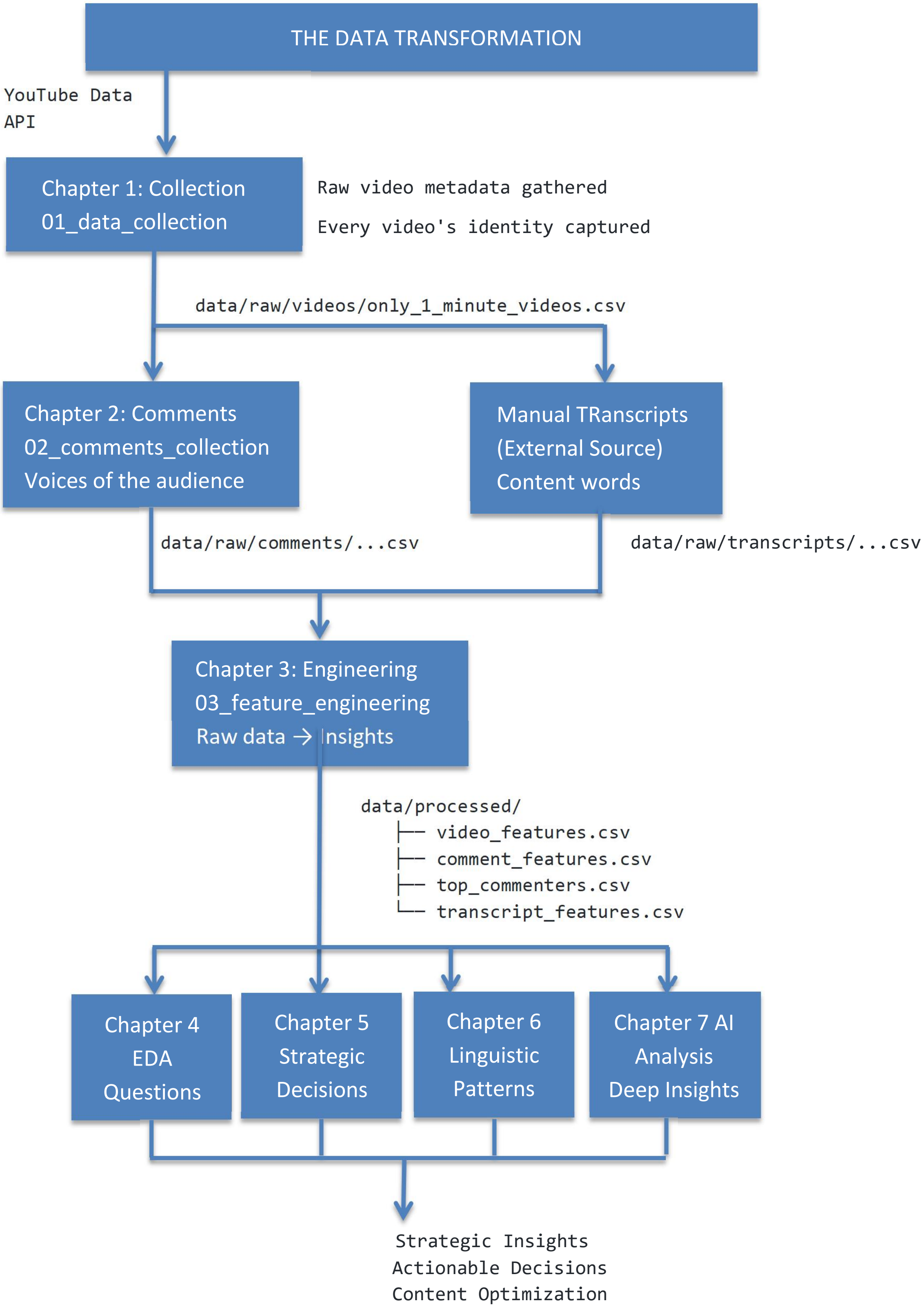
The Outputs:

`data/processed/ai_transcript_analysis.csv` - Transcripts enriched with AI sentiment

`data/processed/ai_comment_emotions.csv` - Comments with 7-emotion scores

What We've Learned: AI doesn't replace human insight—it amplifies it. These models see patterns in sentiment and emotion that traditional analysis might miss. They give us a new lens through which to understand content performance.

The Complete Journey: Data Flow Visualization



The Execution Roadmap (Each chapter builds on the previous ones.)

The Sequential Path:

22. **Chapter 1: Data Collection**

- Prerequisites: YouTube API key, internet connection
- Creates: Raw video metadata
- Time: ~20-40 minutes (depending on playlist size)

23. **Chapter 2: Comments Collection**

- Prerequisites: Chapter 1 complete
- Creates: Raw comment data
- Time: ~30-60 minutes (API rate limits apply)

24. **Chapter 3: Feature Engineering**

- Prerequisites: Chapters 1, 2, and manual transcripts file
- Creates: Processed feature datasets
- Time: ~5 minutes

25. **Chapter 4: Exploratory Analysis**

- Prerequisites: Chapter 3 complete
- Creates: Visualizations and insights
- Time: ~10 minutes

26. **Chapter 5: Strategic Analysis**

- Prerequisites: Chapter 3 complete
- Creates: Clusters, strategic insights
- Time: ~10 minutes

27. **Chapter 6: Contextual Transcript Analysis**

- Prerequisites: Raw transcripts file
- Can run: Independent of other chapters (after Chapter 1)
- Creates: Linguistic insights
- Time: ~15-20 minutes

28. **Chapter 7: AI-Based Analysis**

- Prerequisites: Raw transcripts and comments
- Can run: Independent of other chapters (after Chapters 1, 2)
- Creates: AI-powered sentiment and emotion analysis
- Time: ~60-120 minutes (model inference time)

The Questions We Answer

Throughout this journey, every analysis answers a business question:

- **Performance Distribution:** Is attention evenly distributed or hit-driven?
  - Chapter 4, Graph 4.1

- **Engagement Efficiency:** Do views translate to meaningful interaction?
  - Chapter 4, Graph 4.2

- **Format Efficiency:** Does the 1-minute format deliver value?
  - Chapter 4, Graph 4.3

- **Community Building:** Which videos generate conversation, not just clicks?
  - Chapter 4, Graph 4.4

- **Community Health:** Is engagement broad or concentrated?
  - Chapter 4, Graph 4.5

- **Optimal Pacing:** What speaking rate maximizes engagement?
  - Chapter 4, Graph 4.6; Chapter 5, Graph 5.5

- **Content Investment:** What content effort level yields best returns?
  - Chapter 4, Graph 4.7; Chapter 5, Graph 5.3

- **Content Archetypes:** What types of videos perform best?
  - Chapter 5, Graphs 5.1, 5.2

- **Strategic Balance:** How do we balance reach vs. community value?
  - Chapter 5, Graph 5.4

- **Sentiment Patterns:** How does content sentiment affect performance?
  - Chapter 6, Graph 6.3; Chapter 7, Graph 7.4

- **Audience Emotions:** What emotions do videos evoke in audiences?
  - Chapter 7, Graph 7.3

- **Narrative Structure:** How does sentiment flow through videos?
  - Chapter 7, Graph 7.2

- **Linguistic Style:** What language patterns create success?
  - Chapter 6, Graphs 6.1, 6.4, 6.5, 6.6

Troubleshooting Guide

Every journey has obstacles. Here's how to overcome them:

Issue 1: "ModuleNotFoundError: No module named 'src'"
The Problem: Python can't find your source modules.

The Solution: The first cell in each notebook adds the project root to `sys.path`. Make sure you run it first!

Issue 2: "FileNotFoundError: only_1_minute_manual.csv"
The Problem: Transcript file is missing.

The Solution: Create `data/raw/transcripts/only_1_minute_manual.csv` with columns: `video_id`, `transcript`. This file requires manual transcript entry or extraction.

Issue 3: "HttpError 403: API quota exceeded"
The Problem: YouTube API daily quota is exhausted.

The Solution: YouTube API has daily limits. Wait 24 hours, or use multiple API keys, or request quota increase from Google Cloud Console.

Issue 4: "RuntimeError: Token length exceeds 512"
The Problem: Text too long for AI models.

The Solution: Already fixed! Models use proper truncation (400 words max). If you still see errors, check that you're using the updated code.

Issue 5: "Missing dependencies"
The Problem: Required packages not installed.

The Solution: Install all packages in your virtual environment:

```
pip install pandas google-api-python-client youtube-transcript-api isodate tqdm scikit-learn matplotlib seaborn transformers spacy nltk
```

The Output Files Archive

Raw Data (The Foundation)

- `data/raw/videos/only_1_minute_videos.csv` - The starting point
- `data/raw/comments/only_1_minute_comments.csv` - Audience voices
- `data/raw/transcripts/only_1_minute_manual.csv` - The actual words (manual input)

Processed Features (The Refinement)

- `data/processed/video_features.csv` - Enhanced video metrics
- `data/processed/comment_features.csv` - Comment-level features
- `data/processed/top_commenters.csv` - Community leaders
- `data/processed/transcript_features.csv` - Linguistic metrics

AI Analysis (The Deep Insights)

- `data/processed/ai_transcript_analysis.csv` - AI-powered sentiment
- `data/processed/ai_comment_emotions.csv` - Emotion intelligence

Beyond the Analysis: Next Steps

The journey doesn't end with the last notebook. This is where insights become actions:

Immediate Actions:

29. **Review Visualizations**

 - Go through each graph (referenced above)

 - Identify patterns that surprise you

 - Note correlations that matter

30. **Apply Cluster Insights (Chapter 5)**

 - Use content archetypes to inform production

 - Create templates based on successful clusters

 - Tailor promotion to archetype

31. **Optimize Speaking Pace (Chapter 4, 5)**

 - Identify optimal words-per-second range

 - Train content creators to match successful pace

 - Monitor new content against this benchmark

32. **Balance Content Strategy (Chapter 5)**

- Allocate effort: some videos for reach, others for community

- Use reach vs. community value matrix to guide decisions

- Don't try to make every video do everything

33. **Leverage Linguistic Patterns (Chapter 6)**

- Identify high-performing keywords

- Emulate successful linguistic styles

- Use topic modeling to focus content themes

34. **Harness Emotional Intelligence (Chapter 7)**

- Understand what emotions drive engagement

- Optimize content for desired emotional outcomes

- Monitor sentiment evolution patterns

Long-Term Strategy:

- **Build Dashboards:** Create monitoring dashboards using these metrics
- **A/B Testing:** Use insights to design content experiments
- **Content Templates:** Develop templates based on successful archetypes
- **Community Engagement:** Focus on top commenters for community building
- **Continuous Learning:** Re-run analysis periodically to track changes

The Complete Graph Reference

For easy navigation, here's where to find each visualization:

Chapter 4: Exploratory Analysis

Chapter 5: Strategic Analysis

Chapter 6: Contextual Transcript Analysis

Chapter 7: AI-Based Analysis

Chapter 3: Feature Engineering (Data Validation)

The Final Word

This journey from raw data to strategic insight is more than analysis, it's a transformation. Each notebook is a chapter in a story of discovery. Each graph answers a question. Each insight drives a decision.

The data tells a story. Your job is to listen, understand, and act.

Welcome to data-driven content strategy. Welcome to the NAS Daily Analysis.

Document Version: 2.0 (Storytelling Edition)

Last Updated: 2025

Journey Author: NAS Daily Analysis Project

Graph Count: 23 visualizations across 7 chapters

"In data we trust. In insights we act. In strategy we succeed."

Technical Appendix: GitHub Repository

Click here: [Github_Repository_DBA_Project_by_Team_Decision_Dynamics **].**