Maximizing Success on YouTube's Trending Page:

Content Optimization Strategies Based on Analysis of the Top 200 Videos Over the Last Four Years

•••

LIS 4761 | Group 1

December 8, 2024

ROADMAP

- Project Proposal
- The Dataset
- Data Cleaning
- Descriptive Statistics
- Sentiment Analysis
- Association Mining
- Recommendations for Creators
- Limitations

Project Proposal

Research Question: What trends are common among top-trending YouTube videos in the United States, and how do various factors impact their success?

Research Goal: To analyze the daily top 200 trending YouTube videos in the U.S. from August 2020 to April 2024 and identify patterns contributing to their success.
Success will be evaluated based on the following:

Engagement Ratio	Trending Speed	Trending Retention
(Likes + Comments) / Views	How Quickly did the video trend?	How long did the video remain on the trending page?

Intended Audience: This project is aimed at YouTube creators looking to improve their success on the trending page.

The Dataset

This dataset consists of data from each day’s top 200 trending YouTube videos from August 2020 to April 2024. This dataset is sourced from YouTube’s API and was updated daily until 7 months ago.

Data Cleaning

Cleaning:

- Normalized column names.
  - Reorganized columns.
- Fixed data types.
- Cleaned up NA and incorrect values.
- Added category names with a JSON file.

New Data Frame with Success Metrics:

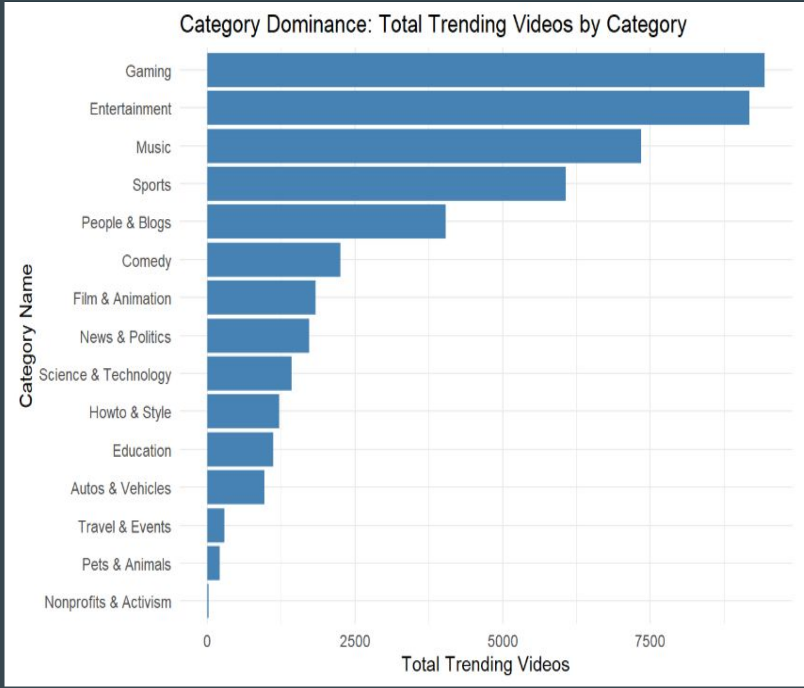
- Added:
  - Max views
  - Max likes
  - Max dislikes
  - Max comments
  - Publication date
  - First date on trending
  - Last date on trending
  - Days until trending
  - Trending retention
  - Engagement ratio

Methods and Techniques

- Descriptive Statistics
- Sentiment Analysis
- Association Mining

Descriptive Statistics
Key Insights for Success Metrics

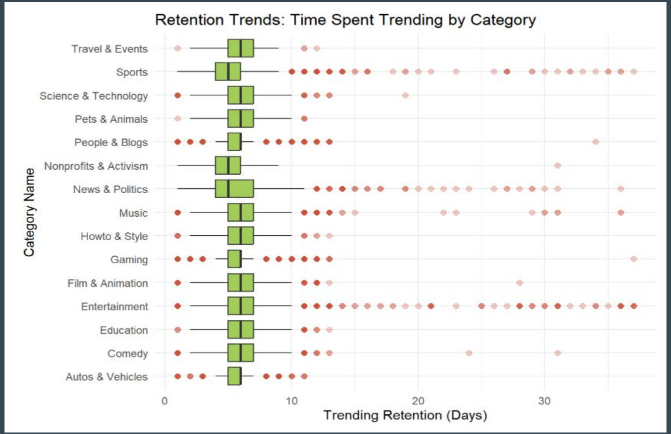
- The dataset analyzed 268,704 videos, totaling over 731 billion views
- The average view count was 2.73 million views, with a median of 937,255.
- We Found no correlation between any success metrics.
- The most successful creators, such as the NFL and NBA channels, dominate with consistent trending content, averaging significant views and high retention
- Top Categories = Gaming, Entertainment, Music
- Companies tend to outperform independent creators in terms of trending speed and visibility.

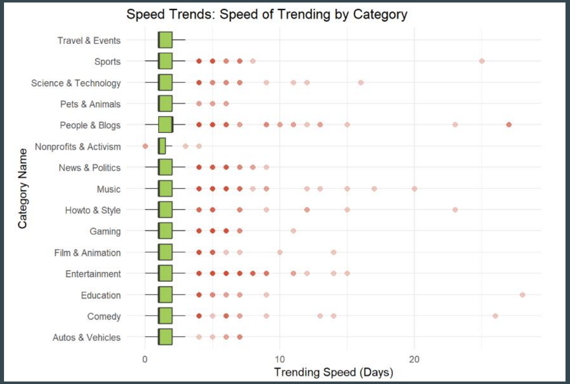


Descriptive Statistics

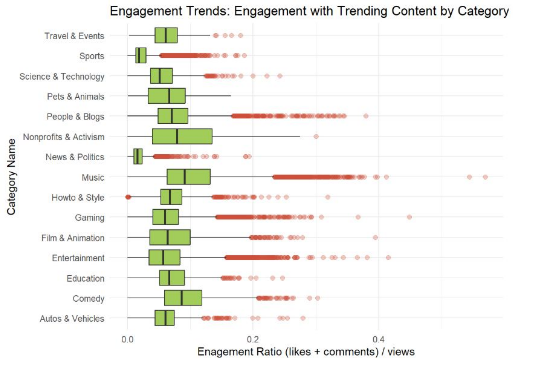
Category-Specific Success Metrics

The retention trends below show where the averages of each categories lie, so creators can gauge where they are in comparison to other creators in the same category as themselves





Speed trends show that some categories achieve visibility faster, while others rely on more consistency, so it shows different ways to be successful based on your content type



Categories with strong engagement ratios indicate strong audience participation which can also drive success

Conclusion

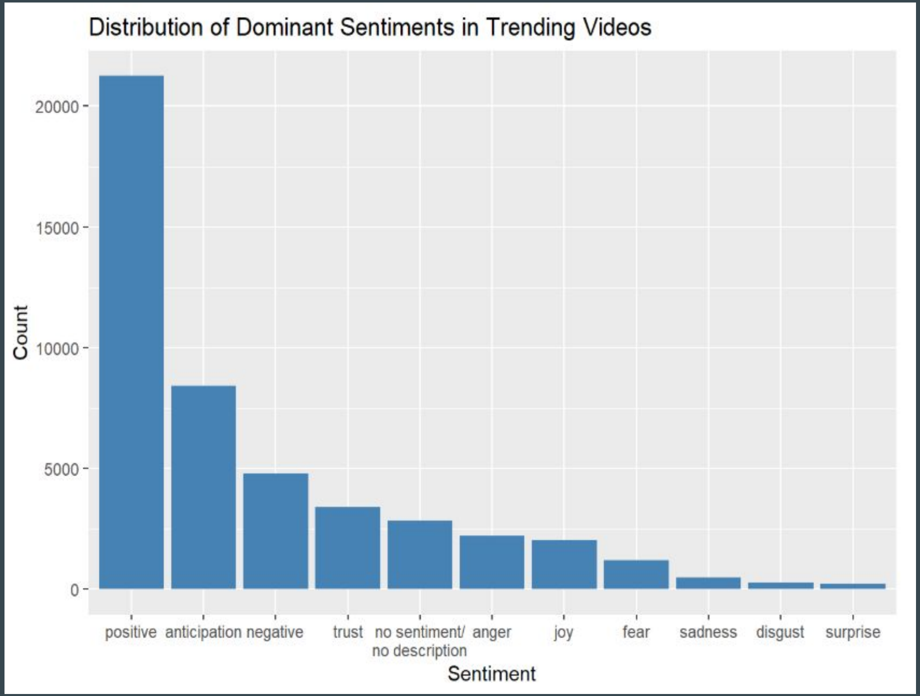
All of this suggest that creators should tailor their strategies to the categories their videos are apart of, and their categories strengths

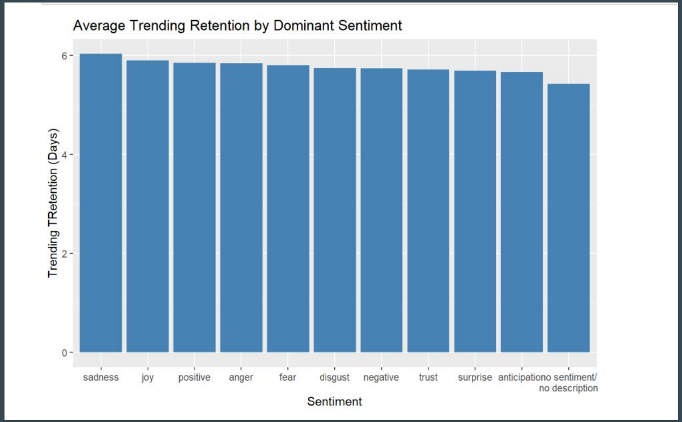
Sentiment Analysis

- Used NRC database to select key emotion sentiments
- Summed all the elements of the description then applied dominant sentiment of the then found the Dominant sentiment
- Trending success was very consistent regardless of emotion
- Faced severe computational limitations for our study

Dominant Sentiment Average Success Metrics

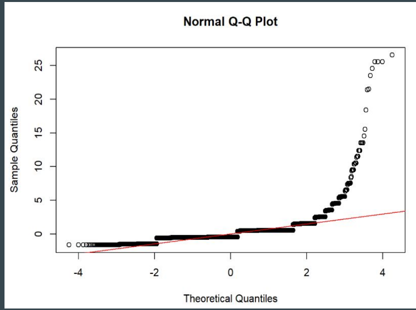
dominant_sentiment	total_videos	avg_trending_speed	avg_trending_retention	avg_engagement_ratio
positive	21246	1.46	5.85	0.064
anticipation	8430	1.48	5.66	0.062
negative	4803	1.51	5.74	0.069
trust	3401	1.46	5.71	0.069
no sentiment/ no description	284			

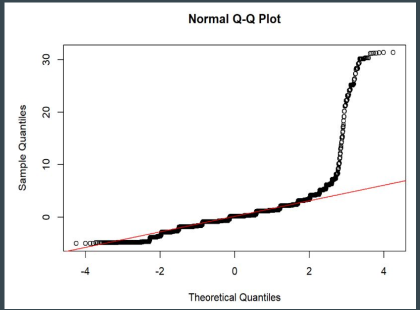


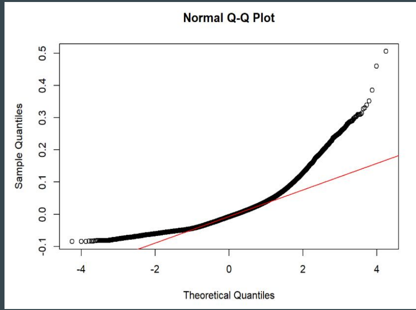


Sentiment Analysis - ANOVA

- Conducted an ANOVA analysis
- All Three ANOVA tests returned P-values below 0.05
- However, the Q-Q Plots showed violations of Normality, which calls into question the reliability of ANOVA
- ANOVA result was most likely a Type I error (False Positive)
- Therefore, we cannot confidently reject our null hypothesis at this time, and must conclude that Dominant Sentiment does not have a significant impact on trending success







Association Mining

- We ran 3 batches of association mining aggregated by Channel Category
  - Descriptive Variable Lengths (Tag Usage, Description Length, Title length)
  - Time of Posting (Time of Posting, Day of Week, Time of Month, Season)
  - Dominant Sentiment

- Association Mining works best with categorical data
- Used Tertile-Based Classification due to outliers identified in the Descriptive Statistics Section
- Creation of an “Overall Success” Categorization

Engagement Ratio

Poor = Bottom ⅓ Tertile: 0.0000 - 0.0413 Engagement Ratio
Good = Top ⅓ Tertile: 0.0765 - 0.5694 Engagement Ratio

Trending Speed

Poor = Bottom ⅓ Tertile: 1 - 5 Days Trending
Good = Top ⅓ Tertile: 6 - 37 Days Trending

Trending Retention

Poor= Top ⅓ Tertile: 2 - 28 Days

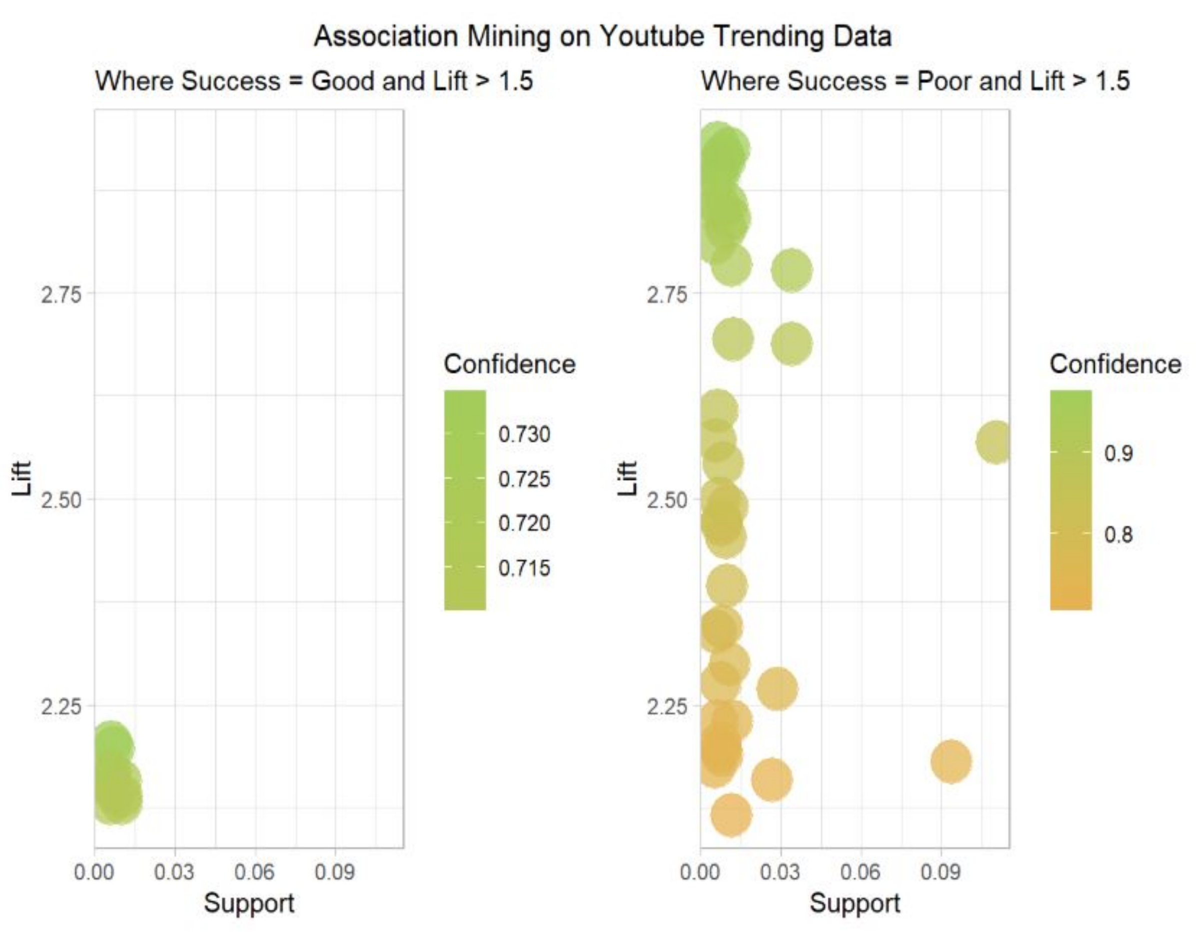
Good = Bottom ⅓ Tertile: 0 - 1 Days

Association Mining

Total of 51 rules identified where Confidence > 70% and Lift > 1.5

- 10 Rules related to "Good" Success
- 41 Rules related to "Poor" Success

Average Confidence for "Good" Success was 72.0%
Average Confidence for "Poor" Success was 85.1%



Recommendations to Clients

- Keep the majority of videos focused on positivity and then try to add in aspects of anticipation and trust unless stated otherwise.
- There aren't really any factors that directly increase the success of a video, but there are some shown to inhibit the success of a video:
  - Don't post during the evenings in Summer.
  - Comedy channels should keep the number of tags low and titles short.
  - Music channels should keep their titles short and have a detailed description, with some sentiments of trust.
  - News and politics channels should use lots of tags and have a detailed description, but avoid anticipation sentiments.
  - Sports channels should use lots of tags, have short tiles, and a detailed description, but avoid trust and anticipation sentiments.

Study Limitations

Unfortunately, we encountered numerous limitations in this study, primarily related to computing power and data availability. Data processing took ~32 minutes per run due to the dataset size and technique intensity.

Data Gaps:

- Limited data to distinguish Companies vs. Independent Creators (manual classification revealed 43 companies vs. 7 independent creators in the top 50 channels).
- No comparative data for videos that didn't trend, restricting insights into how to trend, focusing instead on maximizing success once trending.

Future Directions:

- Comparative analysis would require significantly more computing power.
- Sophisticated models (e.g., LLMs and NLPs) could provide deeper insights into sentiment and success relationships.

Thank You!

Alina Hagen

Team Leader
Association Mining

Joshua Motte

Data Munging
Conclusions
Recommendations

Brodie Vancil

Descriptive Statistics

Sherman LaCost

Sentiment Analysis