YouTube Trending Video Analytics with Sentiment Analysis

- General analysis of factors affecting the popularity of the videos
- Sentiment Analysis to check if there is any correlation between the number of views, likes, dislikes, and comments
- Sentiment Analysis to check which type of video instigates more discussion - whether videos with positive sentiment instigate more discussion or videos with negative sentiment instigate more discussion
- Visualization of trends over time - is violence or negative sentiment increasing over time? How are the music preferences changing over time - planning to integrate Spotify dataset for this task

Problem Statement

How does the popularity of videos affect the sentiments of the general public? What affects the genre preferences of music over time?

Users

Individual content creators, media companies and tabloids, artists, YouTube (age restrictions based on sentiments and type of content), Government

Existing solutions and shortcomings

Before the advent of machine learning, individual content producers, YouTube, or media companies would have combed through the most watched or most liked videos and manually or automatically used scripts that find a specific word and associate it with a sentiment to analyze the general mood of the viewers. Manually sorting through a large amount of data is challenging and time-consuming. In order to tackle complicated problems and identify trends that are difficult to spot while manually skimming through the data, data visualization and the application of machine learning, where needed, are helpful. With data visualizations, we can spot new trends, correlations, and patterns.

Data Visualizations aid in improving our understanding of the data, and if application of machine learning is needed they may also be used to aid in decision-making over the features to be used. If we have specific questions in mind, they might also assist us in determining the answer to a complex question.

After using machine learning models, visualizations can also be produced. Following the use of machine learning, visualizations may aid in the interpretation of the findings as well as how the model got at them and which features were crucial in doing so. They aid in our comprehension of how the machine learning models operate.

Collection Method: Scrape trending video data from the official APIs

- Youtube Data API (https://developers.google.com/youtube/v3)

Data attributes and Types for YouTube dataset

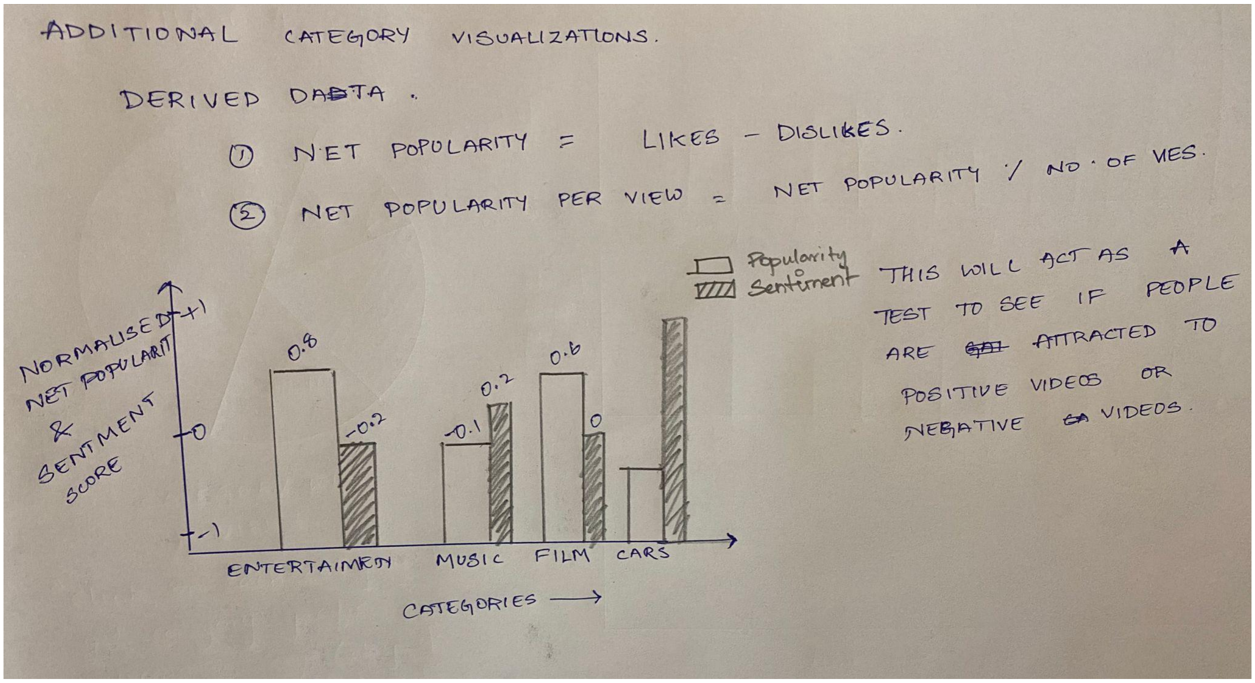
Data Attribute	**Type of the Attribute**
Video ID	Categorical
Title	Categorical
Trending Date	Ordinal/Categorical
Published Date	Ordinal/Categorical
Channel Title	Categorical
Category ID	Categorical
Tags	Categorical
Description	Categorical
Views	Quantitative
Likes	Quantitative
Dislikes	Quantitative
Comments	Categorical
Comment Count	Quantitative
Comments Disabled	Categorical

Design

Youtube Data Analysis Visualizations

Design 1: Net Popularity versus Sentiment Score

Visualization - Bar Chart



ADDITIONAL CATEGORY VISUALIZATIONS.

DERIVED DA~~B~~TA.

① NET POPULARITY = LIKES - DISLIKES.

② NET POPULARITY PER VIEW = NET POPULARITY / NO. OF VIES.

THIS WILL ACT AS A TEST TO SEE IF PEOPLE ARE ~~SA~~ ATTRACTED TO POSITIVE VIDEOS OR NEGATIVE ~~SA~~ VIDEOS.

In this approach, we are initially calculating a derived data from the given data using the following formula:

Net Popularity of a video = No. of Likes - No. of Dislikes

Normalized Net Popularity of a video = Net Popularity / No of Views

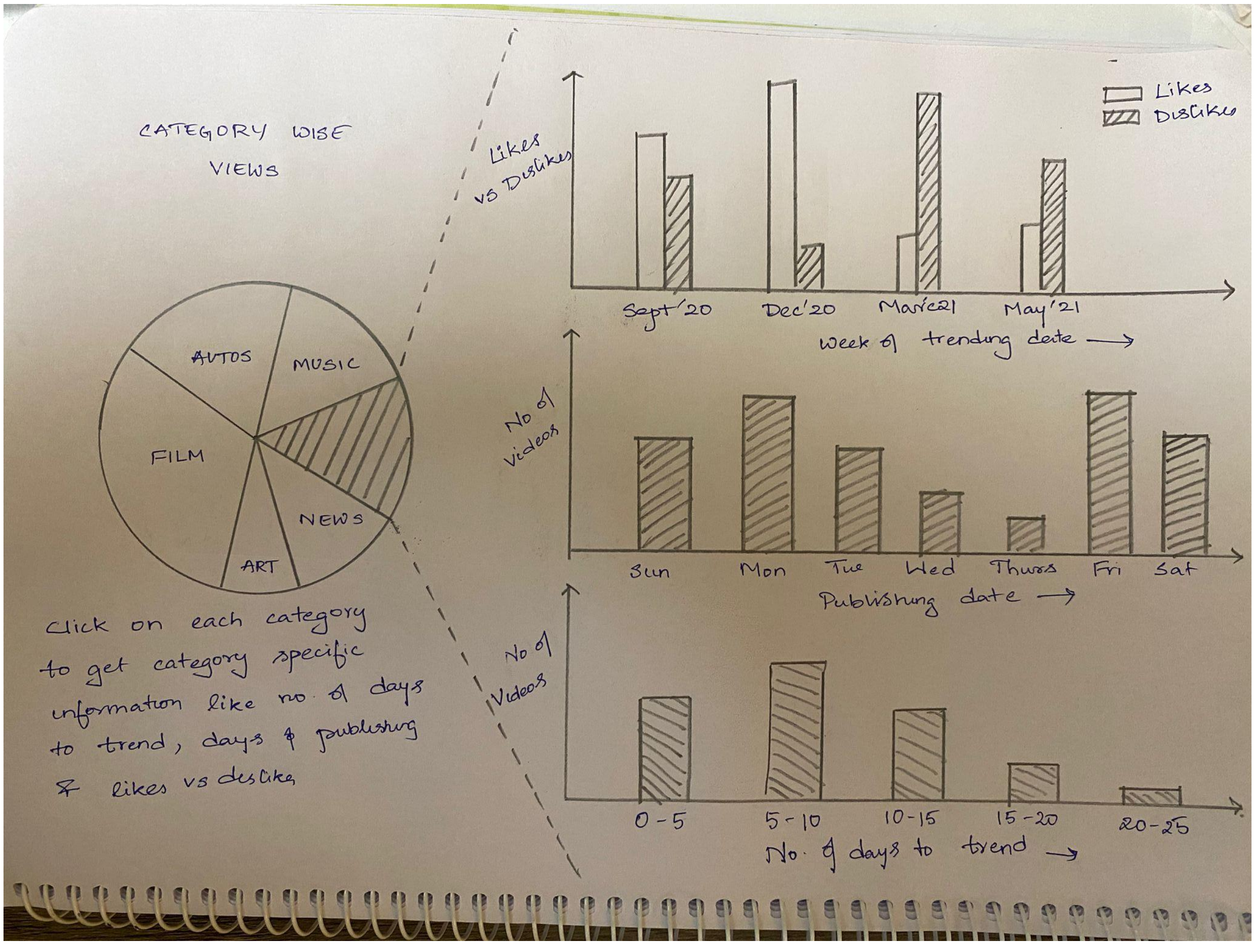
In the chart shown above, unshaded segments represent the actual sentiment score from our dataset whereas the shaded segments represent the derived data (Normalized Net Popularity)

The idea is to normalize the value of calculated popularity in the range of -1 to +1 and map it against the average sentiment score of each category. This visualization will answer the crucial question of whether the general crowd is more inclined towards positive videos or the negative videos.

Interactive Analytics Visualization

Design 2: Category based Interactive Visualization

Visualization - An Interactive series of bar charts

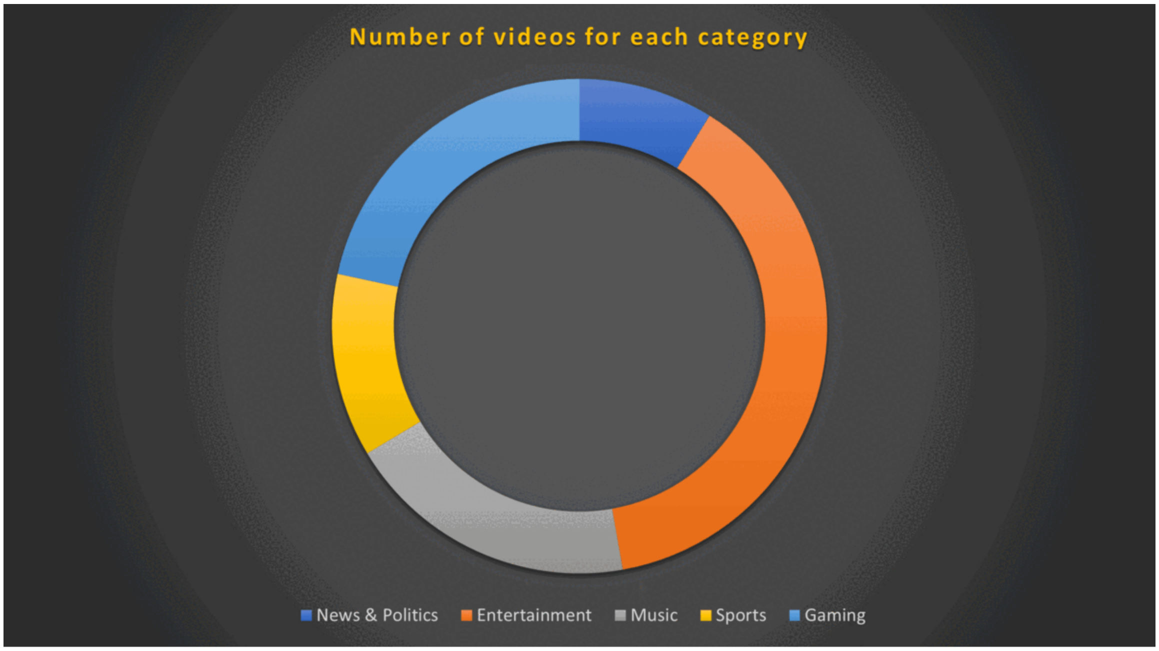


The idea is to make the pie chart of the category-wise split interactive. i.e., if the user clicks on one of the sections, our visualization model prompts them to three graphs specific to that category. The user can have a holistic view of that particular category by viewing the Likes vs Dislikes, Day of Publishing & the number of days to trend. As an extension, clicking on individual components of the bar graph can give a much more detail oriented view of that particular case. For example, if the user clicks on the bar along Sunday, the remaining two graphs could change to show the division of the videos that were released on Sunday in that category.

Sentiment Analysis Visualizations

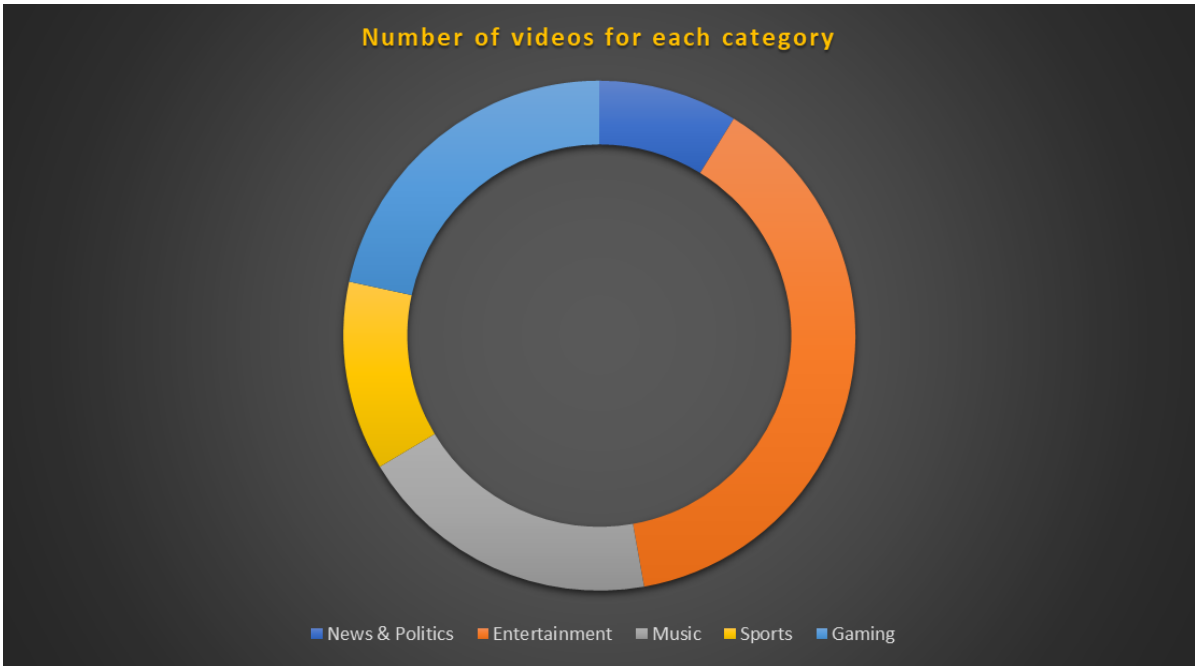
Design 3: Category-wise Analysis of Sentiment Scores

Visualization - Nested Donut / Multi-Level Donut chart

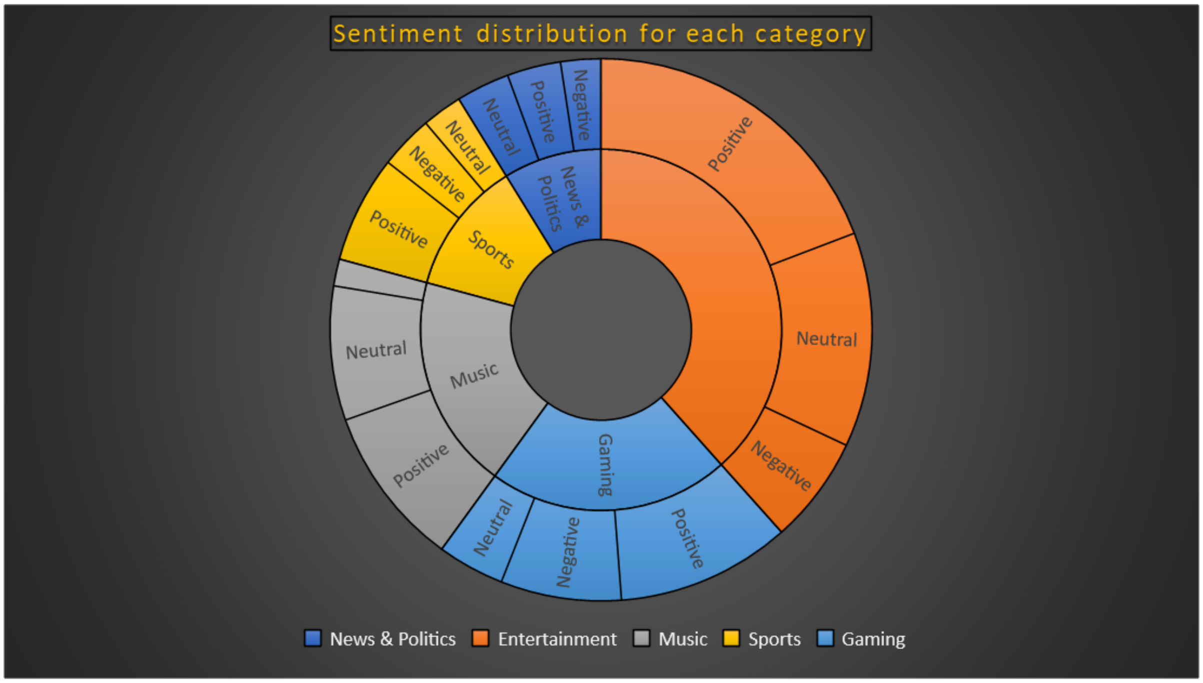


There are 3 layers in the nested donut chart to showcase the distribution of sentiment scores under each category. The size of each sector in the donut chart represents the total number of videos under each category.

- **Categories:** The first level in our chart shows the video categories. For the sake of illustration, we are just showing 5 categories to give an idea.

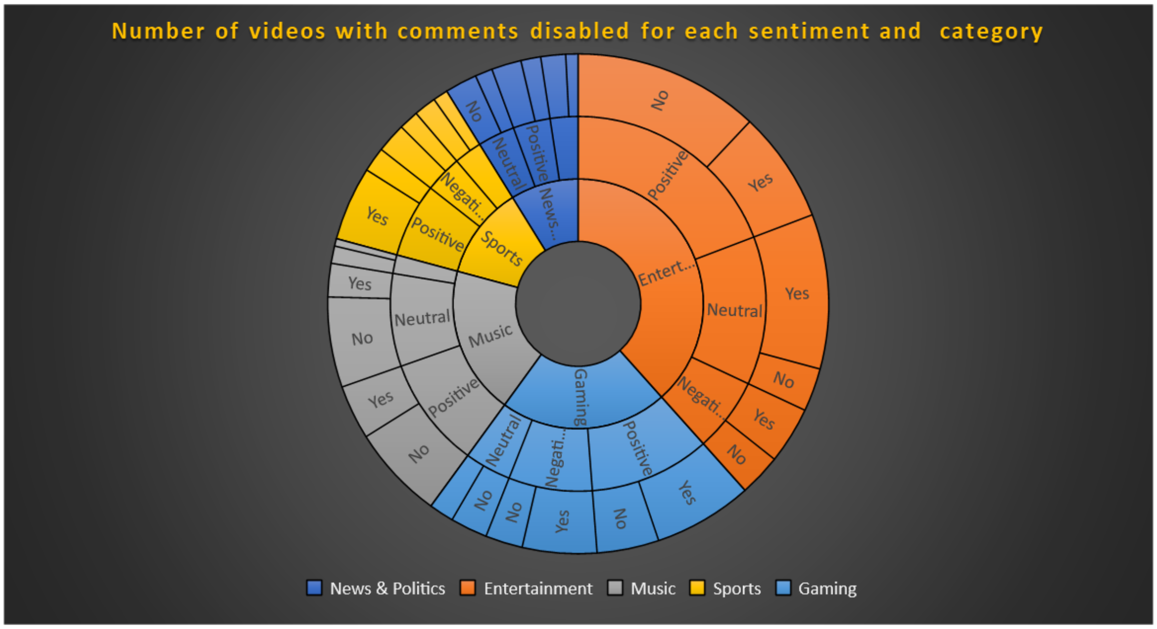


- **Sentiments:** On selecting a category, the second layer of the donut chart shows us the size of the sentiments (positive, negative, and neutral) based on the number of videos in that category



- **Interaction:** On selecting a particular sentiment type, the final layer of the chart shows us the distribution of a derived variable - which functions as a Boolean indicator based on the user interaction for a given video

- If a video has more views and comments than the average number of views and comments (in the dataset), then we assume that the video has a higher user interaction
- The number of views and dislikes can be aggregated and compared with the average in the same way - this can be a third condition to check for, to determine if a video has more user interaction or not
- This gives the user an idea of whether the sentiment scores correlate with the raw statistics of videos that might indicate their popularity



This visualization attempts to answer the following domain question -

“Sentiment Analysis to check which type of video instigates more discussion - whether videos with positive sentiment instigate more discussion or videos with negative sentiment instigate more discussion”

Design Flow

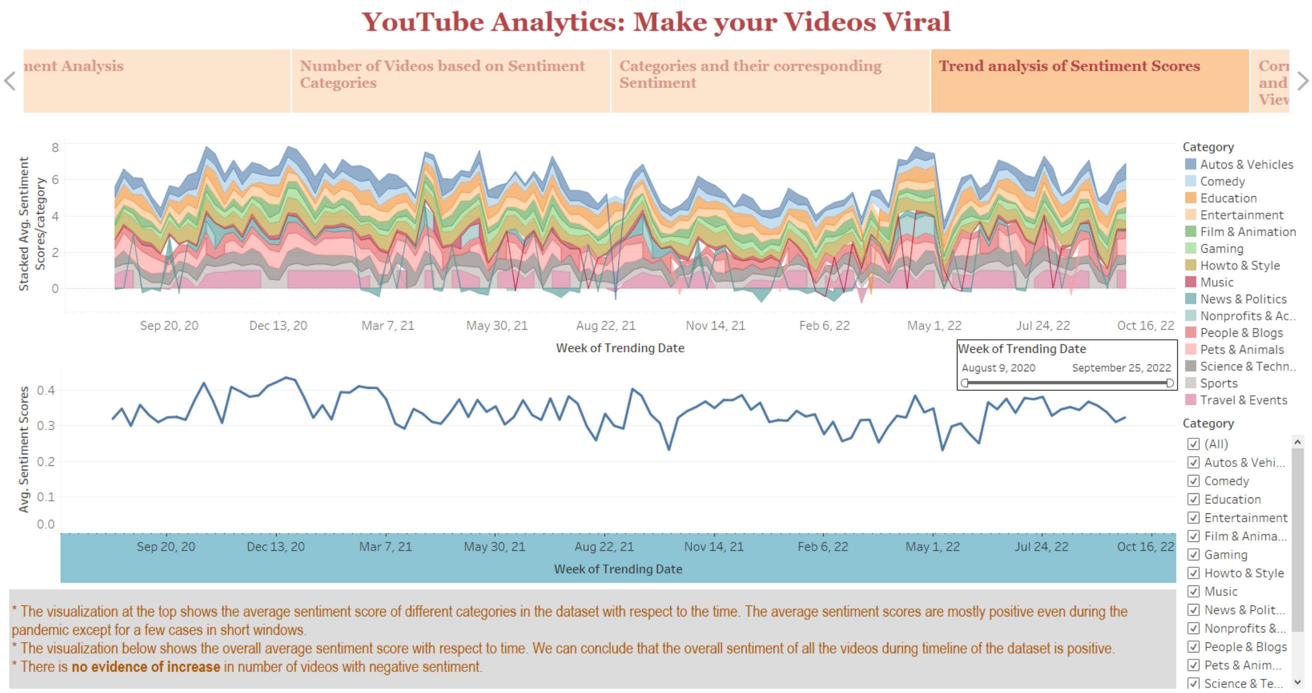
- Selecting a category in the first stage of the donut chart shows the second stage
- If the user selects a sector in the second stage, the third stage is shown

Design 4: Trend Analysis of Sentiment Scores

In this design there are 2 components - an area chart and a calendar chart.

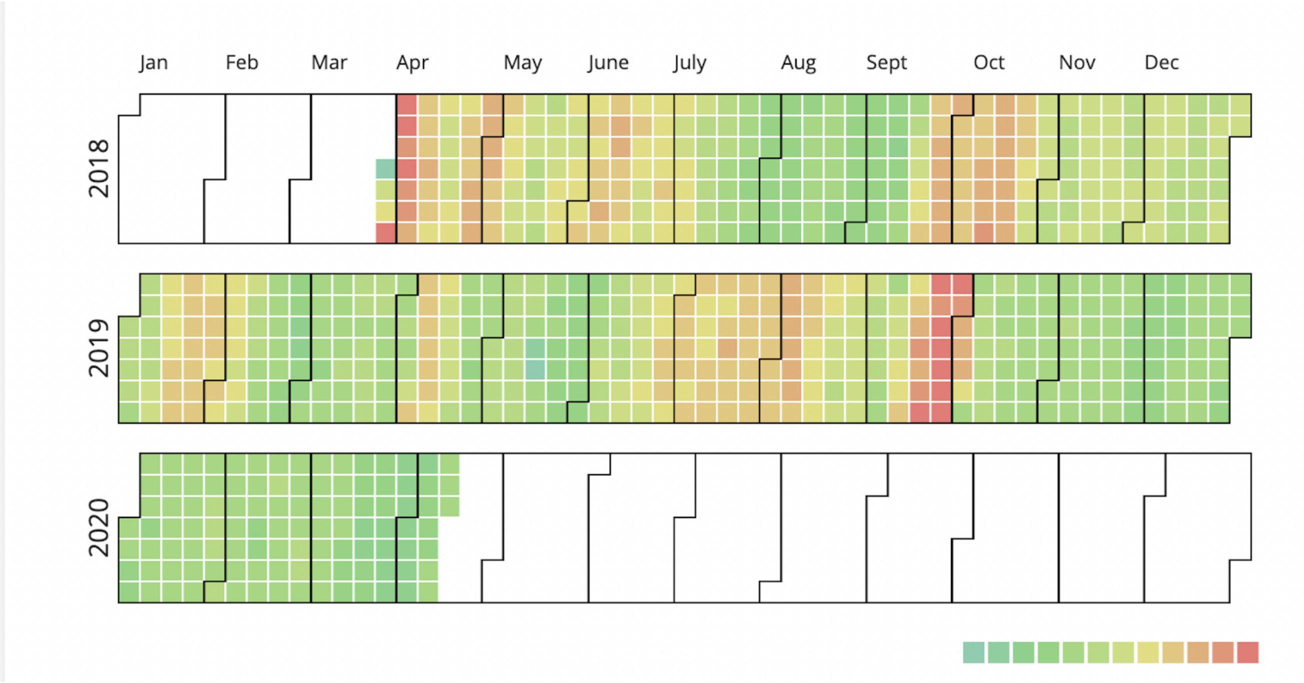
- Area chart
  - The area chart shows the average sentiment scores over time.
  - Brushable interaction is used where the user can select a range of data.
- Calendar chart
  - The calendar chart shows the sentiment scores mapped to a scale (heatmap) for dates that are selected in the area chart
  - The magnitude of heatmap is based on the average sentiment score for the selected dates
  - Details on demand are displayed for the selected data.

The individual components and the final design that integrates both these visualizations are discussed below.



* The visualization at the top shows the average sentiment score of different categories in the dataset with respect to the time. The average sentiment scores are mostly positive even during the pandemic except for a few cases in short windows.
* The visualization below shows the overall average sentiment score with respect to time. We can conclude that the overall sentiment of all the videos during timeline of the dataset is positive.
* There is **no evidence of increase** in number of videos with negative sentiment.

This is a screenshot of the Tableau design (area chart shown for the duration of a year).

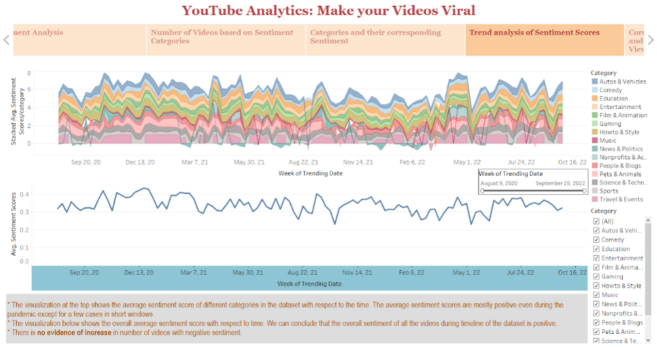


The idea for the calendar chart is similar to the design shown above. For the period of time that has been selected in the area chart (in this case April 2018 - April 2020 has been selected), the average scores would be displayed in the format as seen here. We are also planning to add a fade-in effect to the selected dates, and fade-out the remaining dates.

This gives the user an idea of how videos with positive or negative sentiments correlate to the months in a year or even the days - for example infer the sentiments of videos on weekdays or during holidays.

Design Flow

Area chart - sentiment scores vs. time



YouTube Analytics: Make your Videos Viral

nent Analysis | Number of Videos based on Sentiment Categories | Categories and their corresponding Sentiment | Trend analysis of Sentiment Scores | Cor and Viev

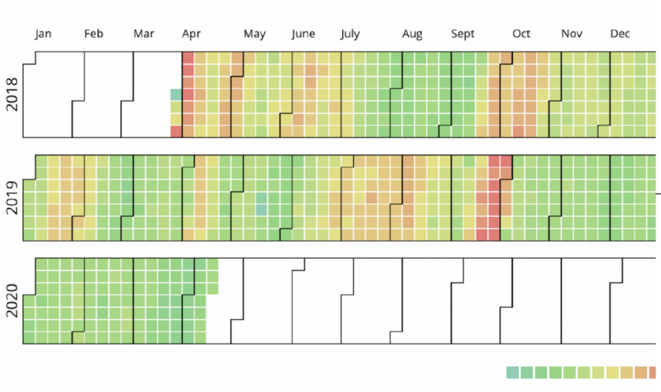
Category

- ☑ (All)
- ☑ Autos & Vehi...
- ☑ Comedy
- ☑ Education
- ☑ Entertainment
- ☑ Film & Anima...
- ☑ Gaming
- ☑ Howto & Style
- ☑ Music
- ☑ News & Polit...
- ☑ Nonprofits &...
- ☑ People & Blogs
- ☑ Pets & Anim...
- ☑ Science & Te...

* The visualization at the top shows the average sentiment score of different categories in the dataset with respect to the time. The average sentiment scores are mostly positive even during the pandemic except for a few cases in short windows.
* The visualization below shows the overall average sentiment score with respect to time. We can conclude that the overall sentiment of all the videos during timeline of the dataset is positive.
* There is **no evidence of increase** in number of videos with negative sentiment.

Interaction
Brushable

User selects a range



Calendar chart

Interaction
Details on Demand

User selects a date

Overall statistics for the selected date