Analyzing YouTube Virality with Content Categorization

Anuj Sinha, Mahir Jain, Megan Ly, Sungha Kang

The Goal of the Research

What are we trying to do? What is the problem? Why is it hard?

"Are specific content categories more likely to go viral than others?"

What is the Problem?

Currently, the YouTube algorithm arranges users' feeds across channels, however, there is a lack of algorithm rearranging the user feed **within a channel**.

Why is it hard?

- Video transcripts can be **long**
- Video transcripts can be **multi-categorical**
- **Virality Score Formula**

Relevant Scholarship

How is it done today, and what are the limits of current practice?

- Previous research typically considered **view counts, engagement metrics** (likes, comments, shares, and subscriber growth), **watch time, etc.** as primary indicators of content virality.

- However, there are very few **video-based** research analyses that consider **transcripts** in calculating the virality of the content.

The Innovation in Methodology

What's new in our approach?

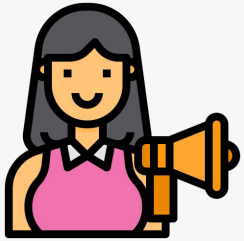
- Using **transcripts**, we will be able to conduct **in-depth content analysis** of YouTube videos, allowing us to analyze the actual words spoken or displayed.
- It will provide valuable **insights into the topics and keywords** that are associated with viral videos.
- To normalize the parameter across videos, we will use the number of subscribers and length of the video.

Stakeholder Interest & Significance

Who cares?

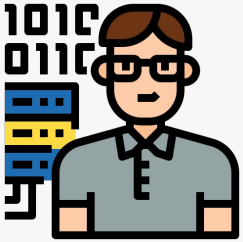
Content Creators

The study can help them on how to structure their videos within their channel.



Marketers & Advertisers

The study can help them utilize the YouTube platform for promoting products and services.



Research Community

The study using transcripts can contribute valuable insights to academic literature and suggest future studies in this area.

YouTube

The platform can devise algorithms to better design within the channel feed.

Anticipated Success & Impact Assessment

If the research is successful, what difference will it make? What impact will success have? How will it be measured?

Virality Scores

Across categories will help us understand what kind of content has the popularity

Demography of Content Consumers

Promoted Content within Each Category

Observation of **statistically significant difference** between virality scores across categories

Curating Within -Channel Content

for a particular cross-cutting high-content-providing channel.

Data Collection

01

Google Search

Select multi-categorical YouTube channels.

02

Transcription API

Extract transcript for videos across X channels.

03

YouTube API

Extract metadata including ‘Likes,’ ‘Comments,’ ‘Shares,’ and ‘Subscriptions’ for each video.

04

API Summarization

Use a summarization tool on the transcript.

05

Hugging Face LLM Model

Use LLM or categorization API on the summarized transcript.

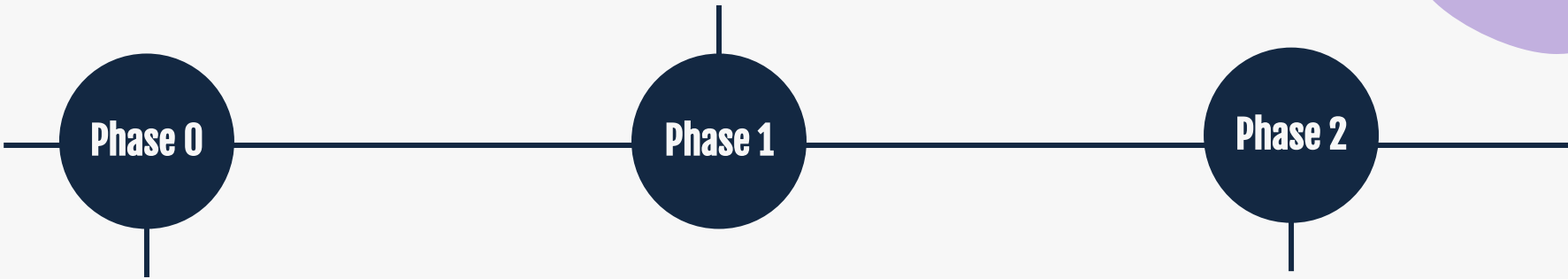
06

Jupyter Script

Store all the above-extracted information into CSV.

Project Phases

Led by Mahir
- **Extraction of Metadata**
- **Extraction of Transcripts**



- **Selection of YouTube Channels for Video Extraction**
- **Document Drafting & Slide Design**
  : Sungha & Anuj
- **Presentation Delivery**
  : Mahir & Megan

Led by Anuj
- **Summarization of Transcripts**
- **Categorization of Transcripts**
- **Data Normalization and Finalization into CSV**

Project Phases

Led by Megan & Sungha

- **Devising of Virality Calculation Formula**
- **Exploratory Data Analysis on the Data**
- **Analysis of the Distribution of User Engagement**

Phase 3

Phase 3*

Led by Sungha & Megan

- **Data Visualization of Analysis**
- **Calculation of Virality Score across Categories**
- **Composition of Research Report**