# Analyzing YouTube Trending Videos: Exploring Factors that Contribute to Video Popularity and Cross-Cultural Understanding

**Team 102: Ahmed Ali, Anthony Del Rosso, Garrison Winter, Jong Pil Park, Min Hsuan Lin, Richard Su**

**Abstract/Introduction:**
YouTube is one of the world's largest social media and video sharing sites, with over 2.6 billion monthly active users worldwide. Understanding what makes a video trend on YouTube is crucial for content creators, marketers, and business owners who wish to increase their channel awareness and monetization (Q4). However, the current practice of analyzing trending videos has limitations in accurately predicting video relevance beyond a couple of days or to a single country (Q2). In this project, we propose a new approach to analyze YouTube trending videos across major English speaking countries and explore the relationships between video content, tags, and topics that drive video trends (Q1). Our approach leverages machine learning such as NLP, regression modeling and network recommendation, allowing us to analyze large volumes of data faster and more accurately.

**Project Description:**
This project aims to analyze YouTube trending videos across English speaking countries, specifically the US, Great Britain, and Canada. The primary objective is to determine what factors lead some videos to trend and to evaluate the relationship between the trend level, which we are determining based on number of views, and different factors such as likes, dislikes, category, tags, and title. The project will also analyze how cultural and geopolitical differences may affect trend factors in different countries.

To analyze YouTube trending videos, we will be using an unique approach to categorize videos by genre, look at views and engagement, and analyze keywords. We will also perform statistical analysis, NLP, and network theory to examine data distribution, correlations, and patterns, to predict video trends and popularity. The success of the project will be measured by the quality and relevance of the data collected and the extent to which the findings can be adopted by relevant stakeholders (content creators, marketers, and business owners). The project's impact will be significant, improving global understanding of YouTube trends across countries and benefiting content creators and marketers through better targeting of audiences (Q5).

**Project Innovation:**
Our project incorporates innovative methods to analyze YouTube trending videos. First, we will categorize trending videos by genre and analyze the popularity of specific content types to identify gaps in the market. Second, we will evaluate the number of views, likes, comments, and engagement of each trending video to identify videos that are popular among the audience. Third, we will analyze the keywords in the title, description, and tags of the trending videos to identify the topics that are currently popular on YouTube. Finally, we will perform statistical analysis to examine data distribution, correlations, and patterns, and use machine learning models to predict video trends and popularity (Q3).

Our innovation seeks to understand the factors that make a video trend on YouTube, and how this varies across different countries. By analyzing YouTube trending videos across English speaking countries, specifically the US, Great Britain, and Canada; we hope to promote cross-cultural understanding and

global awareness of diverse perspectives and cultures. **In addition, we aim to benefit content creators and marketers (Q3) by providing insights into how to effectively create YouTube content that will resonate with cultural mosaic countries.**

However, utilizing this approach carries potential risks, as the data and models we employ may result in erroneous findings due to unconscious bias. Such outcomes could be harmful to content creators, marketers, and other stakeholders who rely on our research. Nevertheless, our project has the potential to offer significant benefits to these stakeholders by providing insights into what drives YouTube video trends and how these trends vary across countries. Such insights can lead to increased engagement and monetization opportunities for businesses and individuals, while also contributing to cross-cultural awareness and broader social and cultural benefits (Q6).

**Project Costs:**
The project's cost will be minimal as we plan to leverage open-source tools and resources to analyze YouTube trending videos. The only costs involved will be for human resources and data processing tools (Q7).

**Literature Survey:**
In summary, the references cover various topics related to analyzing and predicting user behavior on YouTube. The topics include sentiment analysis [5][16][17], video ranking algorithms [1][11][14][18], filtering spam and cyber-bullying comments [8], predicting the popularity of videos using regression analysis and machine learning[6][7][13][15], analyzing the impact of social information on video enjoyment [2][3][4][12], and creating kid-friendly viewing models [9][10]. Each reference highlights the potential usefulness of the methods and models discussed for creators, marketers, and researchers seeking to optimize video performance, increase visibility and engagement, and understand user behavior. However, the limitations and shortcomings of each approach are also discussed, such as limited data quality, siloed approach, and the inability to account for external factors.

**Data Sourcing:**
In addition to the data we used from Kaggle, we also developed and evaluated other data sourcing methods to enrich our data sets. We developed web scraping code to extract other pertinent information from youtube videos such as comments, video duration, channel count of subscribers, and transcript. We also successfully utilized Google's API to extract additional metadata not available through web scraping. We ultimately selected the data we needed based on our models explained below as well as considerations for performance and latency of our data sourcing needs.

1. **Regression Model**
    a. **Proposed Method:**

Our method involved creating and assessing various Regression models to identify the factors that contribute to a video's trendiness on YouTube. We experimented with models such as linear regression, generalized additive models, polynomial regression, k-nearest neighbors, and support vector machines. Unfortunately, generalized additive models showed poor accuracy, and polynomial regression was not a viable option due to the distribution not following any particular Polynomial function. Additionally, the support vector machine took too much time to run (up to 5 hours) due to the size of the data. As a result,

we focused on the models that provided the most promising results, which were the linear regression model and k-nearest neighbor model. However, even after using the elbow method to determine the optimal value of K, we still encountered a high mean absolute error, making the model unusable. Ultimately, we selected the linear model as it showed the highest R-squared values and accuracy with a relatively fast run time.

**b. Experiments/Evaluation:**

Initially, our linear regression model had an R-squared value as low as 0.68. In order to improve the model's performance, we made several data manipulations, including:

- Changing the comment_disabled and rating_disabled variables to binary variables.
- Expanding the title into both Number of characters and Number of words, and expanding the description into both Number of characters and Number of words.
- Changing category_Id into a categorical variable.
- Splitting the publishedAt date into year, month, week, and day.

These changes led to a significant improvement in the model's accuracy, achieving an impressive R-squared value of 0.74 on the testing set and between 0.72 and 0.75 R-squared on datasets from the countries we tested. The efficiency and effectiveness of the linear model make it a practical solution for our purposes.

To further evaluate the model's performance, we tested it on other English and non-English speaking countries, such as Great Britain, Canada, France, Japan, Korea, Mexico, and India. The linear model showed significant impact on 6 out of the 7 countries tested, with India being the outlier. We conducted T-tests and further analysis to determine whether this was due to cultural differences or data issues. The results revealed that the outlier was more related to data and language processing issues rather than cultural interests. As a result, we decided to exclude India from the linear regression analysis.

**c. Conclusion/Discussion:**

To conclude, through our model we found that a linear regression with the factors such as likes, dislikes, comment count, category ID among other factors was the most effective in predicting video view counts across different countries and cultures. Through visualization using Tableau, we gained additional insights into the data and discovered that different countries have varying tastes and interests when it comes to certain categories such as Entertainment, News & Politics, and Non-profit & Activism with France showing the highest deviation. Dashboard Link

However, we also acknowledge the importance of conducting further investigation to explore the potential impact of cultural differences on YouTube video view counts. Thus, moving forward, to better understand the potential impact of cultural differences, it is advised to conduct further research and surveys on the ground. This will help us gain a more nuanced understanding of the cultural differences and preferences of viewers in different countries, and ultimately improve the accuracy of our models in predicting view counts. Our findings have significant implications for content creators, marketers, and advertisers who rely on YouTube as a platform to reach their target audience, and we hope to continue our research to help them achieve their goals.

**2. Suggestive Title Maker**

**a. Proposed Method**

One of the key decisions facing content creators is the phrasing of a title that grabs attention of viewers. The right title not only gives context to the video, but can also be used to elicit reactions and comments which is one of the variables that determines if a video will trend. Our other ambitious goal of this project, after determining factors to predict views, is using NLP to create a random text generator of the best related words that can be used in titles, based on keywords inputted by the content creator. We decided on the use of Long Short-Term Memory (LSTM) as LSTMs are predominantly used to learn, process, and classify sequential data because these networks can learn long-term dependencies between time steps of data.

**b. Experiment/Evaluation Process**

To begin, the hardest part was to pre-process the data. We created functions to remove redundant spaces, clean up numbers, correct misspelled words, clean repeat words, clean emojis and punctuations, and clean up stop words. After the data cleanup was finished, we tokenized the titles. The tokenizer model created a list of 37,471 unique words and the model was saved as a pickle file for faster access in subsequent code. For every title, a sequence was created which consisted of a list of pairs of words, where the first word in the pair was the start and the second word the next word. For example, the title "Dog Flipping Over Ball" would consist of [Dog, Flipping], [Flipping, Over], [Over Ball].

The next stage of the model was to prepare the features and labels, where the first word would be the feature and the last word the label. Both features and labels were converted to binary representation through the use of one-hot encoding. We utilized the Sequential Model from tensorflow.keras.models and trained the model using the following parameters: epochs=50, batch_size=4, validation_split= 0.3, shuffle=True. The epochs to use typically hovered around 50 and not as close to 100-150 since the accuracy would actually spout in the wrong direction for some category models. We utilized Callbacks in order to control the training phase and noticed with a lower batch size we could still include a decent amount of randomness within the model. With higher batch sizes the model tended to overfit which showed with an accuracy of 100%. We utilized typical parameters such as monitor = 'loss' & 'accuracy', factor=0.2, patience=3, min_lr=0.1 to not stray far from the default parameters. We started with an overfitted average score of ~95% and through multiple iterations of model fitting, achieved an average accuracy score of ~65% and average loss of ~1.5 using 50 as our epoch value.

We used the LSTM model we created and applied a category filter, and then ran the category model selected on the data for each video category which would surface a list of the top 5 suggestions for the next possible sequence of words based on the current word. The final result was a user generated prompt that a user can input a word and view the suggestions immediately demo.

**c. Conclusion/Discussion**

Based on our LSTM results, the overall accuracies would vary based on category due to specificity found within each category. An example we found was within the Sports category the difference between NBA and FIFA was typically getting mixed together. This we believe was also due to sheer frequency differences within our dataset for soccer and basketball in comparison to football. With this being said a category such as Games, related words were distinct and resided within their own probabilities based on the most recent word. Implications for content creators, is to decide from the onset if their content is intended for general audiences or a very specific segment of viewers such as sports fans. The usage of domain specific keywords in titles would be important in order to capture the attention of target viewers.

**3. Tagging Network Recommendations**

### a. Proposed Method

Creators and brands alike want to know how their videos fit into existing communities and video genres and how those communities/genres vary from region to region. Such knowledge can inform how brands and creators market their videos, how to vary their marketing from region to region, and how to create content in the future. We've created a tool that allows a user to input a series of keywords and returns a list of similar videos; additionally, the user can take that video and use it to generate recommendations; this process can be repeated iteratively to specifically tailor recommendations to a user's taste. We also give the user the option to select eight different regions, facilitating comparisons between recommendations in different regions.

### b. Experimentation/Evaluation

We used tags on a video as a summarization of a video's content and as the building blocks of our recommendation model. After extracting the video id, country, and tags for each video, we built two undirected graphs: the first had vertices of tags with edges between tags attached to the same video and the second had vertices of video ids with edges between video ids with common tags (we removed duplicate edges if two videos had multiple tags in common). Using these graphs, we created two functions: **like_videos** and **recommendations**. Like_videos takes a set of keywords, a user's location, and a number of recommended videos as arguments. The function works in two steps: first we execute a PageRank algorithm using the tags vertices graph for that user's location with the keywords as a personalization vector to generate a page rank of each tag. Then, for each video in that user's location with at least one tag that matches a supplied keyword, we take the sum of the PageRanks of the tags which match keywords. These sums form a personalization vector for a PageRank algorithm executed with the video vertices graph. We order the videos in descending order by PageRank and keep only the top n as determined by the user supplied number. Recommendations works similarly to like_videos, but takes a video_id as an argument instead of a set of keywords and also takes a user_id argument. The user_id argument allows us to store a user's preferences. A user can repeatedly use the recommendations function and build up a set of preferences, leading to more finely tailored video recommendations. Users interact with these functions through an R Shiny application and have their results returned in the form of a table and an interactive graph network so that they can visualize the connections between their recommendations.

To improve the performance of the functions, we tested two additions: a penalization algorithm that modified the personalization weights of videos to lower the PageRank of videos to similar to a user's taste and an exponential smoothing algorithm to calculate personalization weights overtime. We examined four versions of a penalization algorithm

- (N-p)/ N * weight
- (N- (1- (p/C))*p)/N * weight
- ((N-p)/N + 1-(p/C)) * weight
- 4*(N-p)/N*(1-((N-p)/N))*(1-(p/C))*weight

where weight is the sum of the tags on a given video in common with the watched video or keywords N is the number of tags on the watched video or number of keywords, p is the number of tags shared between the given video and the watched video or list of keywords and C is the number of tags on the given video or list of keywords. We selected a set of ten videos and ran each of them through the recommendation like_videos function with different versions of the penalization algorithm and compared them to the base version. After examining the results, we determined that each version of the penalization algorithm

produced worse recommendation than the base version. We repeated this test with different exponential smoothing weights for the recommendations algorithm and found a 80/20 split between the existing weight and new weight produced the best results.

**c. Conclusion**

The advantage of this model lies in its simplicity–using basic graph analysis and publicly available data, we created a powerful recommendation system that facilitates robust inter regional analysis and allows users to discover content they would not have otherwise. This system does not rely on proprietary or creator specific data (such as watch time) nor does it employ computationally costly algorithms. It's a trivial matter to collect more recent data from Kaggle or through the YouTube API, build a larger database, and scale up the functions to create a better performing system. However, the model does struggle when a video has very few tags which are commonly used, defaulting to broad recommendations. To improve the quality of the model, we would look beyond tags. Users could refine recommendations using non-tag information like video genre or like-dislike ratio as a measure of video quality. demo

**Final Conclusions:**

Our linear regression model showed that factors such as likes, dislikes, and the video category played the most important role in predicting the number of views of a video. In addition, the distribution of videos that trend are consistent in some countries, while differing in others, which speaks of the cultural similarities/differences that drive engagement. For example, Music dominated the categories in all countries, while France had a smaller percentage of videos that trended in the Entertainment and Gaming categories compared with the United States. In our suggestive title maker, we were able to associate the most likely to follow words associated with the prior word and found that in certain video categories, the diversity of the words created more challenges in crafting a title, especially in Sports where there can be words that are only related to that sport such as team names. As such, in creating a title of a video, content creators should take it into consideration in order to reach their target audience. Lastly, in our Tagging Network, we were able to implement a recommendation engine with penalization/decay to give the users ability to discover content outside their original categories and also across borders and cultures. Our suite of 3 tools together enables content creators on YouTube to analyze trending videos by categories and engagement to select the best video category; create titles based on trained datasets for optimal reach; and find other tags that can be used in order to allow more discoverability.

**All team members have contributed fairly to the project.**

**References:**


[1] Krafft, P. M., Chen, Y., Gopal, R. D., & Kannan, P. K. (2019). Video ranking in a user-generated content platform: Evidence from YouTube. Proceedings of the National Academy of Sciences, 116(41), 20378-20383.

[2] Vilares, I., & Banchs, R. E. (2021). The YouTube Corpus of User-Generated Short Videos for Research in Multimodal Understanding and Commonsense Reasoning. arXiv preprint arXiv:2102.07484.

[3] Nanath, Krishnadas, Kaitheri, Supriya, Malik, Sonia, Mustafa, Shahid. Examination of fake news from a viral perspective: an interplay of emotions, resonance, and sentiments.

[4] Amar Krishna, Joseph Zambreno, Sandeep Krishnan. (2014). Polarity trend analysis of public sentiment on YouTube. Journal of Information Science.

[5] C. Weismayer. (2021). Investigating the affective part of subjective well-being (SWB) by means of sentiment analysis. International Journal of Social Research Methodology, 24(6), 697-712.

[6] Gill, K., Arlitt, M., Li, Z., & Mahanti, A. (2007). Describing and forecasting video access patterns. IEEE/ACM Transactions on Networking.

[7] Peter Braun, Alfredo Cuzzocrea, Lam M.V. Doan, Suyoung Kim, Carson K. Leung, Jose Francisco A. Matundan, Rashpal Robby Singh. (2017). Enhanced prediction of user-preferred YouTube videos based on cleaned viewing pattern history. IEEE Transactions on Multimedia, 19(10), 2230-2239.

[8] Shreyas Aiyar, Nisha P Shetty. (2017). N-Gram Assisted Youtube Spam Comment Detection. IEEE Transactions on Multimedia, 19(8), 174-182.

[9] Reddy, S., Srikanth, N., & Sharvani, G. S. (2020). Development of kid-friendly YouTube [3] access model using deep learning. Journal of Intelligent & Fuzzy Systems, 39(2), 2871-2878.

[10] Balakrishnan, J., & Griffiths, M. D. (2017). Social media addiction: What is the role of content in YouTube? Journal of Behavioral Addictions, 6(3), 364-377.

[11] Foster, D. (2018). Factors influencing the popularity of YouTube videos and users' decisions to watch them.

[12] Anne Marthe Möller, Mark Boukes. (2021) Online social environments and their impact on video viewers: The effects of user comments on entertainment experiences and knowledge gained during political satire consumption. New Media & Society, pages 146144482110159.

[13] Rui, L. T., Afif, Z. A., Saedudin, R. R., Mustapha, A., & Razali, N. (2019). A regression approach for prediction of Youtube views. Bulletin of Electrical Engineering and Informatics, 8(4), 1502-1506.

[14] Barjasteh, I., Liu, Y., & Radha, H. (2014). Trending videos: Measurement and analysis. arXiv preprint arXiv:1409.7733.

[15] Li, Y., Eng, K., & Zhang, L. (2019). YouTube Videos Prediction: Will this video be popular?

[16] Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 1-18.

[17] Singh, R., & Tiwari, A. (2021). YouTube comments sentiment analysis. International Journal of Scientific Research in Engineering and Management, 5(5), 517-521.

[18] Justin, J. (2018, November 16). Reverse engineering YouTube Search. Briggsby.