Trendsetters Across Borders: Decoding YouTube's Viral Landscape 2024

Arkadii Bessonov
Matrikelnummer 7038413
arkadii.bessonov@student.uni-tuebingen.de

Kirian Fink
Matrikelnummer 5698172
kirian.fink@student.uni-tuebingen.de

Neagoe Mihai Alexandru
Matrikelnummer 7023684
mihai-alexandru.neagoe@student.uni-tuebingen.de

Abstract

In this study, we aimed to identify country- and region-specific characteristics reflected in trending YouTube videos. The motivation behind our research is to analyze cultural and social indicators manifested in popular video content, which can contribute to a better understanding of audience preferences alongside the real-world attributes of different countries. Specifically, we investigated the popularity of emoji usage in video titles as well as the relationship between title length and the Human Development Index (HDI). We also analyzed the popularity of various video categories across countries. Thus, our work tests several hypotheses based on real data, with the aim of uncovering patterns in trending video content and their association with the social characteristics of countries.

1 Introduction and Dataset description

We selected a dataset of YouTube trending videos from 113 countries, collected via the YouTube API and openly available through [4]. This dataset is highly suitable given its appropriate scale and credibility. The dataset comprises daily snapshots of trending videos for the entire year of 2024, providing a reasonable timeframe for comparative analysis across different countries.

The dataset includes key YouTube API features: data collection date, country, video title, view count, and publication date. We enriched it with video categories from the YouTube API and 2022 Human Development Index (HDI) data from Wikipedia, chosen for its stability over two years, aligning with our study's needs.

To ensure dataset quality and relevance, we conducted thorough preprocessing. We assessed accuracy by analyzing video counts, distribution patterns, missing values, and outliers. We also filtered out live streams and excluded "shorts" to minimize API anomalies and maintain consistency with our analysis goals.

2 Methods

We carefully approached the theoretical justification of statistical tests. We used efficient and well-known statistical tools implemented in python libraries such as "scipy", "statsmodels". To ensure strong theoretical guarantees, we made data post-processing, to enhance independence of observations as discussed in section 4. Specifically, we used following methods in our main experiments:

Project Report for *Data Literacy* 2024/25

1. For emoji usage experiment 3.1 to confirm patterns, we used Welch's t-tests [7] (suitable for unequal variances) with one-sided alternative hypothesis $\mu_{\text{Middle East country}} \geq \mu_{\text{EA country}}$, applying Benjamini-Yekutieli [1] correction to all country pairs (correction is suitable for case of dependent pairs).
2. For exploring HDI/title dependency (3.2) we used Shapiro-Wilk [5] ($p = 0.00$) test for normality of residuals to check that we can't rely on linear regression. Taking into account this fact, we didn't use Pearson correlation and regression analysis. We showed that there is possibly monotonic dependency. We therefore employed dual non-parametric measures – Kendall's Tau [3] (τ-b) and Spearman's Rho [6] (ρ) – to assess ordinal concordance and monotonic trend. This dual approach balances robustness (via Kendall) with traditional interpretability (via Spearman).
3. In subsection 3.3 formal hypothesis testing employed proportion z-tests with Bonferroni correction [2] for multiple comparisons. The experimental framework evaluated:

$$H_0 : p_{C_1,cat_1} = p_{C_2,cat_2} \quad \text{(Equal category prevalence)}$$

$$H_1 : p_{C_1,cat_1} < p_{C_2,cat_2} \quad \text{(Reduced prevalence in test country)}$$

Where p values below the $\alpha_{bonferroni} = \alpha/n$ threshold indicated statistically significant differences.

For geographic identification, we employed ISO-2 country codes as this standardized coding system. Throughout all experiments, we maintained a fixed significance level of $\alpha = 0.05$ to ensure statistical rigor in our analyses.

3 Results

3.1 Cultural Patterns in Title Emoji Usage

We tested whether Middle Eastern countries systematically use more emojis in trending video titles than Euro-Atlantic counterparts —a feasible, culturally relevant comparison using available data. We are comparing 15 countries (8 Middle Eastern, 10 North-Euro-Atlantic, see 1b for specific countries) with videos filtered to specific region (details: Section 4). Indeed, Figure 1a shows Middle Eastern titles contain 2-3$\times$ more emojis.

Results in Figure 1b confirm significant differences ($p < 0.05$) for about 80% of Middle East vs. West comparisons. Important exceptions emerged: Czech Republic showed no significant differences with Middle East along with Qatar and Bahrein in comparisson with Euro-Atlantic countries, suggesting unique regional trends[1].

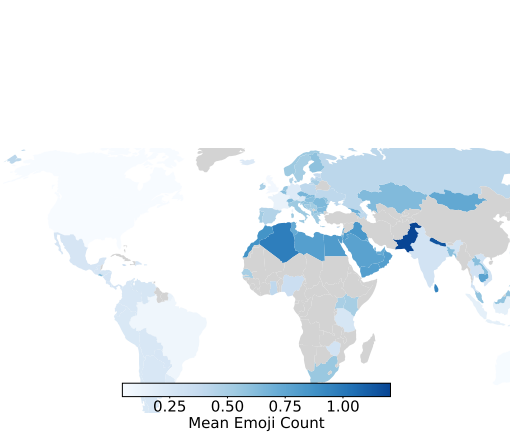
Overall, these results provide insight into culturally specific trends in emoji usage in trending video titles across regions. The countries mentioned above, however, highlight the need for granular regional analysis. Overall, our findings indicate that many large Middle Eastern countries tend to use more emojis in their trending titles, potentially reflecting culturally driven patterns in video naming practices.

3.2 The Human Development Index (HDI) and title length

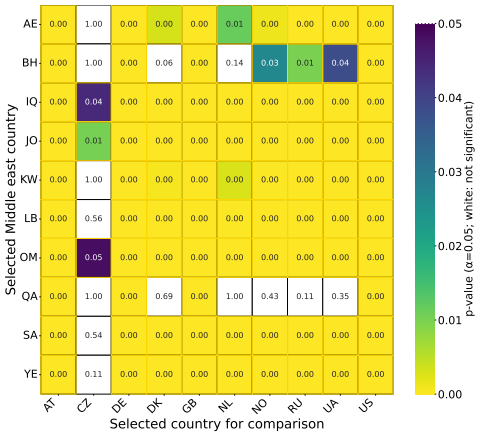
The HDI provides a multidimensional assessment of national development, education, income, and life expectancy. Our study investigates potential monotonic associations between HDI levels and digital content patterns, specifically examining whether socioeconomic development correlates with linguistic concision in video titles. Title length might reflect either communicative efficiency in high-literacy contexts or lexical compensation strategies in resource-constrained environments and is simple indicator.

Following the preprocessing steps in section 4 to ensure sample independence, mean title lengths were computed per country. The Kendall Tau implementation function in python revealed a statistically significant negative association ($\tau = -0.24$, $p = 2 \cdot 10^{-4}$) as well as Spearman R ($\rho = -0.37$,

[1]We also explored that East European countries have more emojis than Western and Bahrein's p-values decreased after correction.



(a) Mean emojis per title. Darker shades indicate higher usage.



(b) Statistical significance (BY-corrected) between country pairs. Colored: Middle East > West ($p < \alpha$).

Figure 1: Emoji usage analysis shows consistent cultural differences, with Middle Eastern titles containing significantly more emojis.

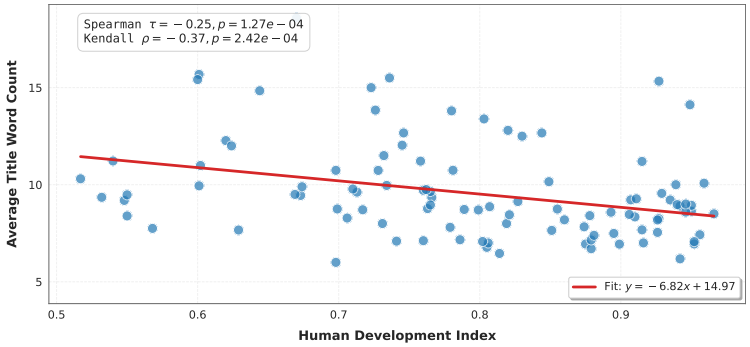


Figure 2: Relationship Between National HDI and Mean Video Title Length. Linear approximation presented here only to show visual trend direction.

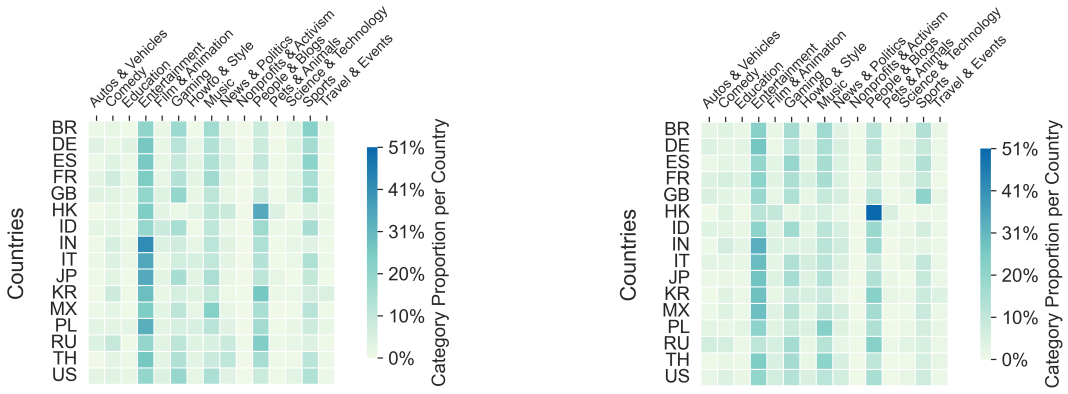
$p = 1 \cdot 10^{-4}$), visualized in Figure 2. Small negative values of coefficients demonstrate decreasing dependency with moderate effect.

This inverse monotonic relationship supports two plausible interpretations: Higher HDI nations may adopt brevity conventions to enhance information density in competitive attention economies, or countries with lower literacy rates might employ circumlocution strategies where limited vocabulary necessitates multi-word expressions. However, we can't argue about causality and can only state statistical significance in monotonic dependency.

3.3 Cross-Country Analysis of YouTube Content Categories

We used the same preprocessing as described in section 4. To better understand the differences in the trending video categories between different countries we plotted the category distributions per country on a normalized heatmap (Figure 3). This is using a blue-gradient color scheme where intensity corresponds to category frequency inside the corresponding country. The visualization highlights distinct regional preferences - notably in Hong Kong (HK) the category "Poeple & Blogs" seems to appear especially often in the trends.

With the above specified method we compared the category People & Blogs in Hong Kong with Germany resulting in a $p = 7.22 \times 10^{-76}$ indicating a reduced prevalence in Germany. Another interesting comparison contrasted German and Russian comedy content. The statistical test yielded $p = 1.82 \times 10^{-6}$. This indicates Russian trending lists contain significantly more unique comedy content, potentially reflecting cultural preferences for humor styles or algorithmic promotion patterns.



(a) Duplicates not filtered out.

(b) Duplicates across time and country filtered out.

Figure 3: Heatmap of normalized category distributions. Colors represent proportional prevalence, with darker blues indicating higher relative frequencies.

The analysis revealed interesting insight when comparing automotive content in the Germany versus sports in the United States videos. With a $p = 5.82 \times 10^{-5}$ value we reject the null hypothesis of equal popularity to the alternative that the popularity of Sports in the US is higher than the popularity of Autos & Vehicles in DE. Comparing the unfiltered and filtered data in (Figure 3) shows that the overall trend is not influenced to much by the filtering.

4 Limitations and Discussion

For statistical tools, independence is crucial. We employed post-processing to address this issue. In our analysis, we primarily focus on unique videos for each country by excluding those that appear in multiple countries. This approach allows us to concentrate on local content and treat videos within each country as independent. In pairwise comparisons, this filtering step results in the loss of less than 15% of the dataset, so its impact is minimal. By removing global videos, we focus on country-specific content, ensuring independence and capturing regional behavior. Additionally, we selected one random video per channel if more than one video from the same channel was trending, further ensuring data independence.

Our study is based solely on 2024 data, so the findings are specific to this year and should not be generalized to other years without additional research. However, it is likely that these trends reflect enduring cultural aspects rather than temporary changes.

All conclusions in this work are based on trending videos. While these likely reflect the main interests and preferences of viewers in each country, they may not fully represent overall viewing habits or the full range of available content.

The dataset consists of daily snapshots, which are suitable for most analyses. However, some videos may appear or disappear within the same day, making our dataset a subsample rather than a complete capture of all trending videos.

Given more resources, these limitations could be addressed by using more granular data or expanding the analysis over multiple years to capture temporal dynamics and assess the stability of trends.

5 Statement of Contributions

Arkadii Bessonov conducted the experiments in Sections 3.1 and 3.2, helped with Section 3.3, and contributed to EDA of languages. He also the primary author of the report. Kirian Fink utilized the YouTube API to add categories to dataset, authored Section 3.3, and performed EDA on categories. Mihai Neagoe created the GitHub repository, refactored all the code, and organized the project into structured files. Additionally, he conducted EDA on language usage. For grammar correction, DeepSeek-V3 was used on a limited basis.

References

[1] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[2] C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, 8:3–62, 1936.

[3] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[4] A. Rathnayake. Trending youtube video statistics (113 countries). Accessed: 2025-01-01.

[5] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[6] C. Spearman. The proof and measurement of association between two things. 1961.

[7] B. L. Welch. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.