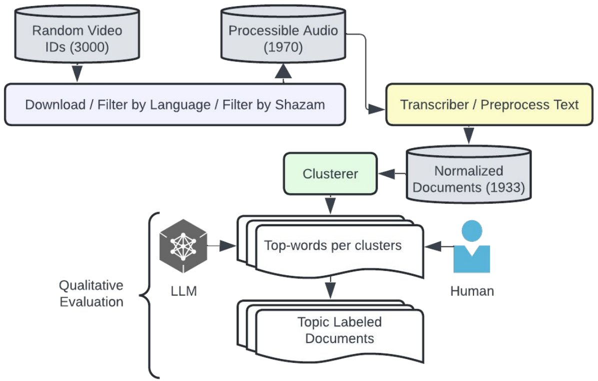
University of Massachusetts Amherst

YouTube Topic Modeling

Ozel Yilmazel, Ryan McGrady, Virginia Partridge, Ethan Zuckerman

Abstract

This study explores **large-scale, unsupervised topic modeling on YouTube** beyond pre-labeled categories. Using a random URL discovery technique, we curated a dataset of 3,000 high-confidence English videos. We applied and evaluated four clustering strategies, selecting the most effective for qualitative analysis. We used human annotators and emerging LLM labeling techniques. Our findings show that topic modeling on randomly sampled YouTube content is feasible and give a sense of content distribution. We also highlight challenges and suggest future work, including expanding this method to other languages for broader, cross-linguistic analysis of YouTube's content landscape.



Methods

Preprocessing: We began by filtering out music-only videos using Shazam, then transcribed the audio with Whisper. Videos with 37 words or fewer were excluded to ensure sufficient textual content for analysis.

Clustering: We experimented with four clustering approaches:

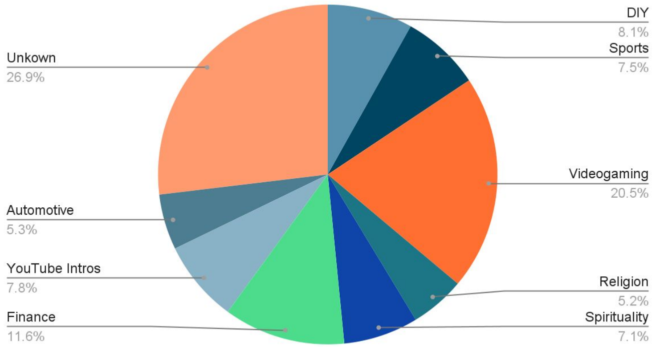
- Approach 1: TF-IDF with KMeans++
- Approach 2: Sentence Transformer embeddings with KMeans++
- Approach 3: TF-IDF-weighted word embeddings with KMeans++
- Approach 4: Latent Dirichlet Allocation (LDA).

For each method, **we explored between 10 and 30 clusters**, using KMeans++ initialization to improve stability.

Topic Labeling: Key terms were extracted using a PMI × TF weighting. Cluster topics were then labeled with input from iDPI Lab staff and several large language models, including Grok, GPT-4o-mini, and Claude.

Our machine learning and evaluation pipeline is presented in the workflow diagram.

Video distribution by topic



Results

Evaluation: Statistical metrics showed moderate success, with Approaches 2 and 3 yielding the strongest quantitative performance. This shows that learning correct representations for videos is crucial for good clustering. However, unsupervised evaluation requires both statistical and qualitative assessment. We selected Approach 2 for in-depth qualitative analysis.

Findings: Approximately 80% of clusters were deemed coherent by human annotators. There was substantial overlap between human and LLM-generated labels. The pie chart illustrates the predicted topic distribution from both annotation sources, with agreement on all labels except "Unknown." The pie chart shows our predicted distribution based on our sample.

Next Steps

- Expand our sample size to up to 40,000 videos.
- Broaden our analysis by including other widely spoken languages, such as Spanish and Turkish.
- Test out more human annotation tasks, such as intruder detection.
- Fine-tune sentence transformer to create clusters that improve statistical measures.

References: Kozlowski, D., Pradier, C., & Benz, P. (2024, August 13). *Generative AI for automatic topic labelling*. arXiv.org. https://arxiv.org/abs/2408.07003, Von Luxburg, U., Williamson, R. C., Williamson, R. C., & Guyon, I. (2012). Clustering: science or art? In I. Guyon, G. Dror, V. Lemaire, G. Taylor, & D. Silver (Eds.), *JMLR: Workshop and Conference Proceedings* (Vol. 27, pp. 65–79). https://proceedings.mlr.press/v27/luxburg12a/luxburg12a.pdf, McGrady, R., Zheng, K., Curran, R., Baumgartner, J., & Zuckerman, E. (2023). Dialing for Videos: a random sample of YouTube. *Journal of Quantitative Description Digital Media*, *3*. https://doi.org/10.51685/jqd.2023.022

University of Massachusetts Amherst

YouTube Topic Modeling

Ozel Yilmazel, Ryan McGrady, Virginia Partridge, Ethan Zuckerman

We present top-10 words for 5 example clusters. Numbers next to words represent how many times it has appeared in the cluster. Before you look at the labels, we suggest trying to come up with a topic on your own.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
federal: 18 economic: 15 government: 35 tax: 24 market: 34 agency: 16 state: 59 property: 23 economy: 15 individual: 35	resonates: 13 wand: 10 pentacle: 8 emotional: 16 emotion: 17 forefront: 7 resonate: 7 reality: 22 nurturing: 6 passion: 17	amen: 47 bible: 47 scripture: 38 verse: 47 worship: 35 pray: 53 church: 52 praise: 39 christ: 52 sin: 41	championship: 22 coach: 31 defender: 18 matchup: 14 defensive: 22 league: 28 opponent: 20 referee: 11 playoff: 11 offense: 21	transmission: 16 rear: 18 engine: 22 automatic: 16 steering: 13 passenger: 14 wheel: 23 liter: 10 vehicle: 21 cylinder: 12
Human: Finance LLM: Economics	Human: Spirituality LLM: Spirituality	Human: Christianity, Religion LLM: Christianity	Human: Sports LLM: Sports, Football	Human: Car Reviews, Cars LLM: Automotive

Clusters are color-coded to pie-chart in previous slide.