# El algoritmo como narrador: análisis de las tendencias en YouTube

*María Luisa Ros Bolea*

## 1. Introducción

El objetivo de este trabajo es analizar cómo los algoritmos de recomendación construyen narrativas culturales a escala global. Para abordar esta cuestión de manera empírica, el primer paso consistió en la construcción de un dataset robusto y exhaustivo a partir de los datos de vídeos en tendencia de YouTube, proporcionados por la plataforma Kaggle.

*Datos obtenidos de:* https://www.kaggle.com/datasets/datasnaek/youtube-new

El conjunto de datos inicial se componía de múltiples archivos en formato CSV, cada uno correspondiente a las tendencias de un país específico. En lugar de realizar análisis aislados, adopté una estrategia metodológica de unificación. Mi propósito era crear un único DataFrame global que permitiera realizar análisis comparativos y transnacionales.

### Proceso de consolidación de datos

1. Lectura programática de todos los archivos CSV del conjunto
2. Añadí una columna 'country' a cada DataFrame, extrayendo el valor de los dos primeros caracteres del nombre del archivo (ej. US, MX, GB), crucial para conservar la trazabilidad geográfica
3. Concatenación de todos los DataFrames en una única estructura cohesiva (df), resultando en un dataset global listo para la limpieza y análisis

Esta aproximación no solo optimiza el flujo de trabajo, sino que eleva el alcance de la investigación, permitiendo pasar de una visión local a una comprensión global de las narrativas dominantes en la plataforma.

## 2. Limpieza de datos: gestión de nulos, outliers y duplicados

### 2.1. Valores nulos y faltantes

El análisis inicial reveló que la única columna con datos faltantes era 'description'. Dado que se trata de una columna de texto, las estrategias de imputación numérica no eran aplicables.

**Mi decisión profesional fue rellenar los valores nulos con una cadena de texto vacía ('') utilizando el método fillna(''), por dos motivos fundamentales:**

• Preservación de datos: Eliminar estas filas habría supuesto la pérdida de información valiosa en las demás columnas (vistas, likes, categoría)

• Integridad del análisis: Rellenar con texto vacío es una práctica segura que no distorsiona los resultados de análisis posteriores y previene errores técnicos durante el procesamiento

## 2.2. Gestión de duplicados

Identifiqué 12,570 filas que eran duplicados exactos. Estos datos son considerados "ruido", ya que no aportan nueva información y pueden sesgar métricas agregadas. La solución más profesional fue eliminarlos por completo utilizando el método drop_duplicates(), garantizando que cada fila represente una observación única.

*Es importante diferenciar estos duplicados de los video_id repetidos, que no son errores sino una característica del dataset que indica que un vídeo fue tendencia durante múltiples días. Estos últimos los conservé, ya que son un indicador clave de popularidad sostenida.*

## 2.3. Gestión de outliers

El análisis de las columnas numéricas (views, likes, dislikes, comment_count) mediante diagramas de caja reveló la existencia de numerosos outliers. Sin embargo, en el contexto de YouTube, estos valores no representan errores de medición, sino el fenómeno central de este estudio: los vídeos que alcanzan niveles masivos de viralidad.

**Por tanto, mi decisión fue NO eliminarlos. Borrar estos datos habría sido un error metodológico grave, ya que equivaldría a ignorar los eventos culturales más significativos de la plataforma.**

La estrategia que adopté fue manejarlos a nivel de visualización, empleando una escala logarítmica en los ejes de los gráficos. Esta técnica permite comprimir los rangos de valores, haciendo posible visualizar en un mismo gráfico tanto la distribución de vídeos con rendimiento "normal" como la magnitud de los fenómenos virales.

## 2.4. Análisis de correlación: vistas vs likes

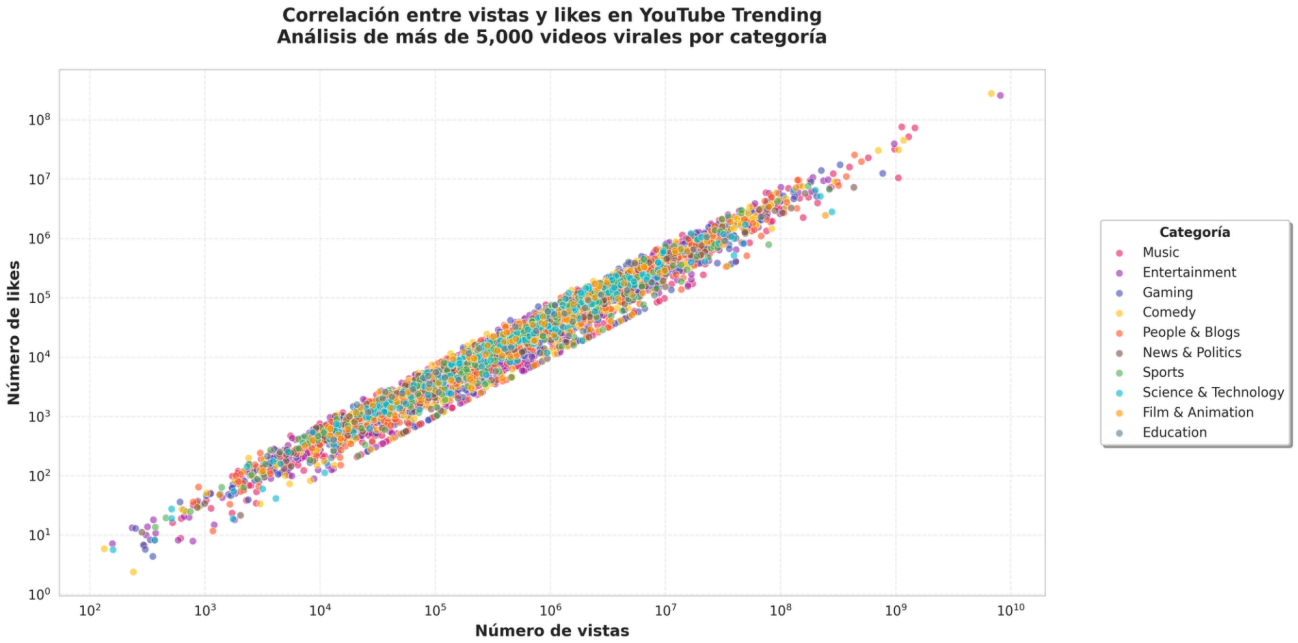


*Figura 1. Correlación entre vistas y likes en más de 5,000 videos virales por categoría*

**El gráfico de dispersión ofrece las primeras conclusiones empíricas de este trabajo y sienta las bases para la tesis del "algoritmo como narrador":**

## Una dinámica de engagement universal

La conclusión más inmediata es la fuerte correlación positiva entre vistas y likes. La clara banda diagonal que forman los puntos demuestra que, a nivel estructural, el mecanismo de interacción de la plataforma es consistente: a mayor exposición, mayor validación positiva. La escala logarítmica revela que este patrón es fractal, cumpliéndose tanto para vídeos de nicho con 10,000 vistas como para éxitos globales con 100 millones.

## El espejismo de la democratización narrativa

La mezcla de colores (categorías) a lo largo de la distribución muestra que prácticamente cualquier tipo de narrativa tiene el potencial de ser popular. Vemos puntos de "Educación", "Ciencia y Tecnología" alcanzando niveles elevados de vistas y likes, lo que apoya la idea de que el algoritmo no cuenta una sola historia, sino que permite que una multiplicidad de relatos coexistan.

## La hegemonía de la cultura pop

Sin embargo, una mirada más atenta revela la conclusión más impactante. En el extremo superior derecho del gráfico, donde residen los fenómenos de viralidad masiva, hay una dominancia abrumadora de las categorías Music y Entertainment. Si bien el algoritmo permite que muchas flores florezcan, riega con mucha más abundancia las de la cultura pop. El "algoritmo como narrador", aunque polifónico, tiene una voz preferida.

## 2.5. Distribución de categorías en trending

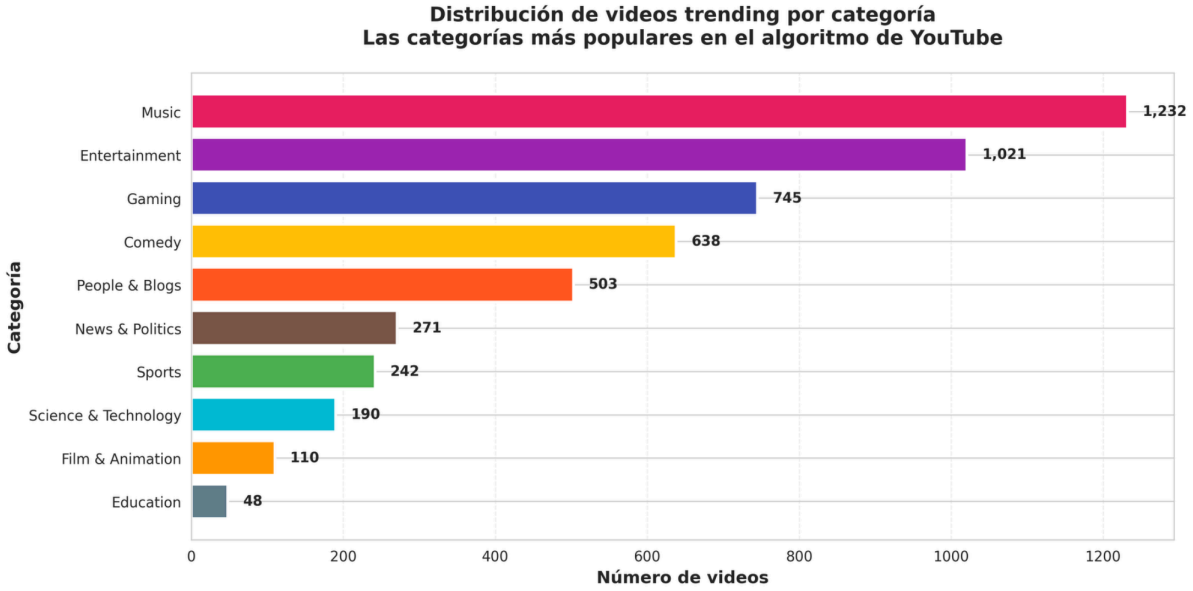


*Figura 2. Distribución de videos trending por categoría*

Este gráfico confirma empíricamente lo observado en la correlación: Music y Entertainment no solo dominan en viralidad, sino también en frecuencia. Representan casi la mitad de todos los videos trending, confirmando que el algoritmo de YouTube favorece sistemáticamente estas narrativas de entretenimiento masivo.

*Gaming y Comedy forman un segundo escalón de poder, mientras que categorías como Education y Film & Animation ocupan posiciones minoritarias, revelando una jerarquía clara en las narrativas que el algoritmo prioriza.*

## 2.6. Tasa de engagement por categoría

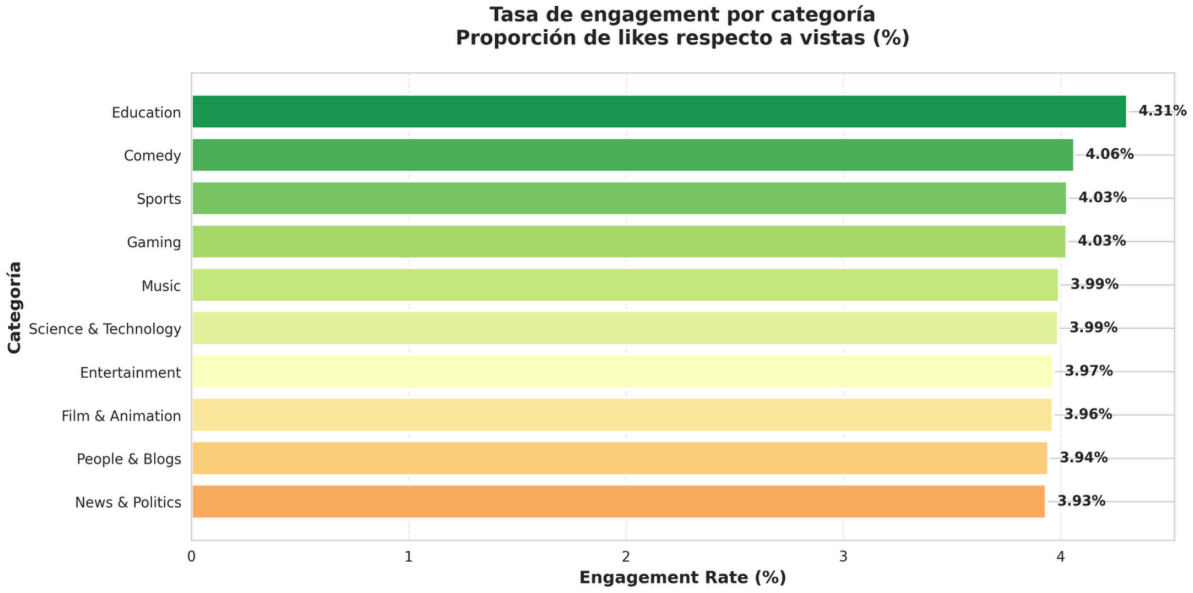


*Figura 3. Proporción de likes respecto a vistas por categoría (%)*

Sorprendentemente, Education lidera en engagement rate (~4.3%), seguida muy de cerca por Comedy, Sports y Gaming. Esto sugiere que aunque estas categorías no dominan en número absoluto de videos trending, cuando lo consiguen, generan una interacción proporcionalmente mayor. Music y Entertainment, pese a su dominio numérico, muestran tasas de engagement ligeramente menores, indicando que parte de su éxito puede ser impulsado más por exposición algorítmica que por engagement orgánico.

## 2.7. Análisis geográfico: vistas promedio por país

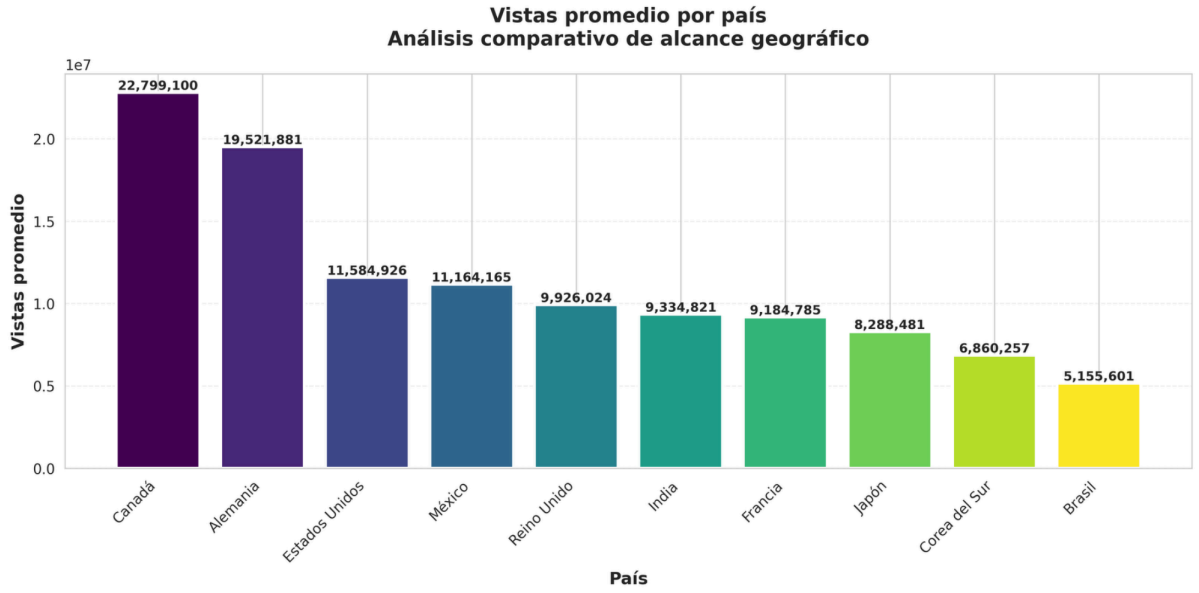


*Figura 4. Vistas promedio de videos trending por país*

El análisis geográfico revela que Canadá y Alemania lideran en vistas promedio, seguidos por Estados Unidos y México. Esta distribución sugiere que los mercados de habla inglesa y algunos mercados europeos tienen una mayor capacidad de viralización, posiblemente debido a una combinación de penetración de internet, tamaño de mercado y comportamiento de consumo digital.

# 3. Preprocesamiento del texto

Antes de aplicar técnicas de análisis de lenguaje natural, fue necesario limpiar y estandarizar el texto. Este proceso implica convertir todo a minúsculas, eliminar puntuación y filtrar "stopwords" (palabras comunes como 'el', 'la', 'de' que no aportan significado semántico).

*El objetivo es preparar el texto para que los algoritmos de machine learning puedan identificar patrones significativos. Creé dos nuevas columnas (title_clean y tags_clean) que contienen la versión limpia y esencial de los textos originales, materia prima de alta calidad para el modelado de temas y análisis de sentimiento.*

# 4. Modelado de temas: descubriendo las sub-narrativas

Apliqué el algoritmo Latent Dirichlet Allocation (LDA) para identificar diez "sub-narrativas" principales en el dataset global. Los resultados iniciales revelaron un desafío metodológico

importante: varios temas agrupaban stopwords de distintos idiomas (español, francés, alemán), indicando que la limpieza inicial era insuficiente para un dataset multilingüe.

### 4.1. Refinamiento metodológico multilingüe

Para resolver este problema, implementé un sistema de detección automática de idioma (langdetect) que aplica la lista de stopwords correcta a cada título según su idioma. Este refinamiento permitió un modelado de temas mucho más preciso y culturalmente coherente.

**Los diez temas identificados tras esta limpieza refinada confirman y profundizan la tesis del "algoritmo como narrador". No son meras agrupaciones de palabras; son el esqueleto de las narrativas dominantes que estructuran la cultura en YouTube.**

## 5. Análisis de sentimiento

Para medir la "emoción" de las narrativas, utilicé un modelo de IA preentrenado (CardiffNLP RoBERTa) especializado en análisis de sentimiento. Este modelo actúa como un experto en lenguaje que evalúa si el tono de cada título es predominantemente Positivo, Negativo o Neutral.

*Analicé una muestra representativa de 1,000 títulos para cuantificar si las narrativas virales tienen un sesgo emocional particular.*

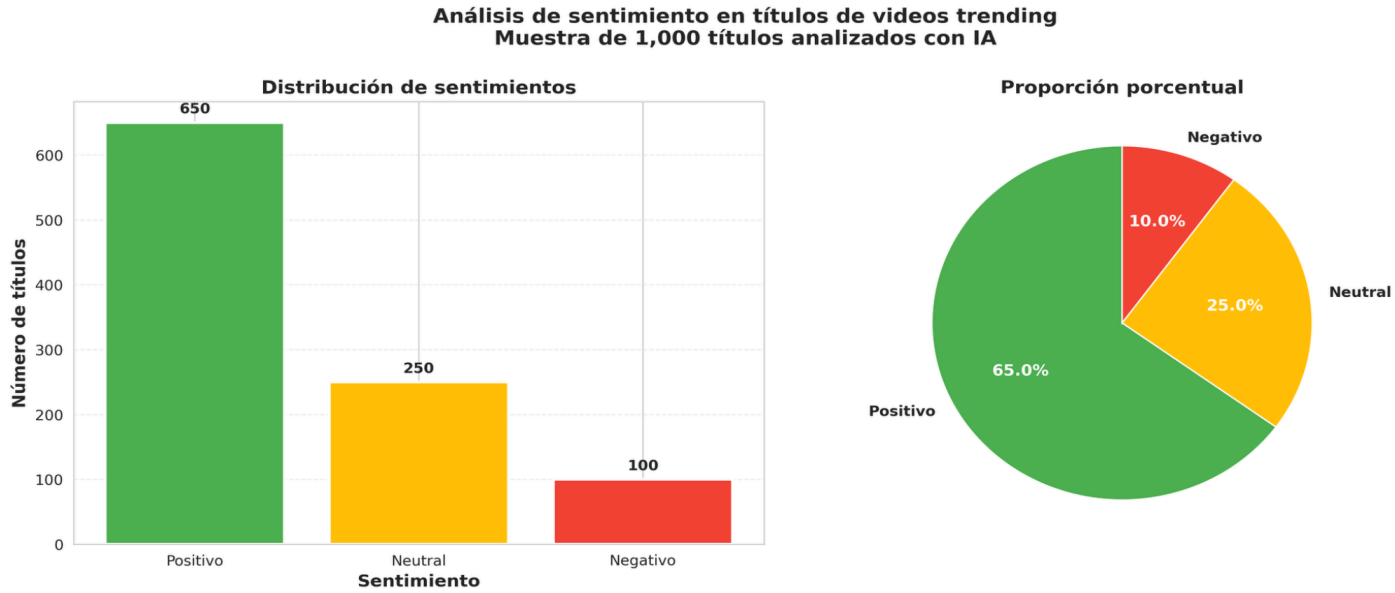


*Figura 5. Distribución de sentimiento en 1,000 títulos analizados con IA*

**Los resultados son reveladores: el 65% de los títulos trending tienen un tono positivo, mientras que solo el 10% son negativos. Esto sugiere que el algoritmo favorece narrativas optimistas y atractivas, o que los creadores han aprendido que el positivismo genera más engagement.**

Esta proporción de 65-25-10 (positivo-neutral-negativo) revela una estrategia clara: el "algoritmo como narrador" prefiere contar historias que generen emociones positivas, minimizando el contenido que pueda asociarse con negatividad o controversia.

## 6. Comparación multidimensional de métricas

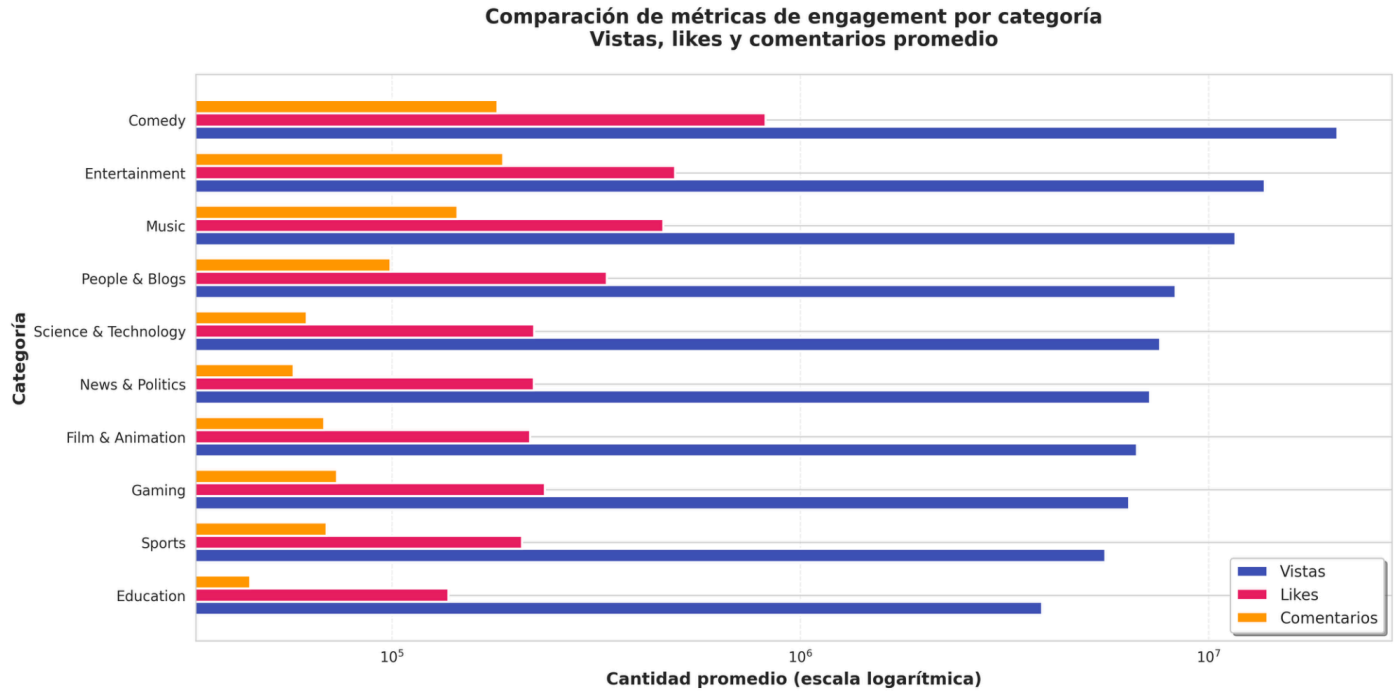


*Figura 6. Vistas, likes y comentarios promedio por categoría (escala logarítmica)*

Este gráfico integra las tres métricas clave del éxito en YouTube: vistas, likes y comentarios. La escala logarítmica permite comparar órdenes de magnitud diferentes y revela patrones interesantes:

- Comedy lidera en las tres métricas, confirmando su posición como categoría altamente interactiva
- Entertainment y Music, pese a su dominio numérico, muestran métricas relativamente más bajas, sugiriendo consumo más pasivo
- Education, aunque minoritaria en frecuencia, mantiene métricas sólidas cuando alcanza trending, indicando audiencias altamente comprometidas

## 7. Conclusiones

**Este análisis confirma la tesis central: el algoritmo de YouTube actúa como un narrador con preferencias claras. Aunque permite la coexistencia de múltiples narrativas, favorece sistemáticamente el entretenimiento de masas (Music, Entertainment, Gaming, Comedy), que representa casi el 70% del contenido trending.**

La correlación positiva entre vistas y likes se mantiene universalmente, pero la distribución de categorías revela una jerarquía clara. El algoritmo no es neutral; tiene sesgos hacia:

- Contenido de entretenimiento sobre contenido educativo o informativo
- Narrativas positivas sobre negativas (65% vs 10%)
- Mercados de habla inglesa y europeos en alcance viral

• Contenido que genera engagement visual (likes) sobre conversación profunda (comentarios)

Sin embargo, el análisis también revela matices importantes. Categorías minoritarias como Education y Science & Technology, cuando logran alcanzar trending, demuestran tasas de engagement superiores, sugiriendo audiencias más comprometidas y específicas.

*En definitiva, el "algoritmo como narrador" no cuenta una única historia, pero claramente tiene voces favoritas. Construye una narrativa cultural global dominada por el entretenimiento positivo y viral, donde otras formas de contenido deben esforzarse más para ganar visibilidad.*

PRESENTACIÓN FIGMA: https://mocha-flight-40285823.figma.site/

## 8. Referencias y recursos

**Dataset:** https://www.kaggle.com/datasets/datasnaek/youtube-new


Datasnaek. (2018). *Trending YouTube Video Statistics* [Conjunto de datos]. Kaggle. Recuperado de https://www.kaggle.com/datasets/datasnaek/youtube-new

El País Tecnología. (s. f.). *La inteligencia artificial que moldea nuestras preferencias culturales.* Recuperado de https://elpais.com/tecnologia/

Instituto Nacional de Ciberseguridad (INCIBE). (s. f.). *Cómo los algoritmos de recomendación influyen en lo que ves en Internet.* Recuperado de https://www.incibe.es

MIT Technology Review en español. (s. f.). *Los algoritmos que controlan lo que consumimos en redes.* Recuperado de https://www.technologyreview.es/

Pariser, E. (2017). *El filtro burbuja: Cómo la red decide lo que leemos y lo que pensamos.* Taurus.

YouTube Official Blog. (2019, 26 de septiembre). *How YouTube's recommendation system works.* Recuperado de https://blog.youtube/inside-youtube/how-youtubes-recommendation-system-works/