Propuesta de investigación: Modelo predictivo de viralidad de un video de youtube

Alumnos: Strejilevich Francisco, Giacobelli Francisco

Profesores: María Noelia Castillo, Ignacio Spiousas

Materia: Ciencia de Datos

7/12/2024

Introducción

Se entiende por videos virales a videos que se vuelven populares por las personas que lo comparten entre sus amigos y parientes (Broxon, et al, 2011)
En los últimos años Youtube se ha vuelto una de las redes sociales más importantes del mundo, cualquiera puede pasar a ser una celebridad simplemente por el hecho de que tuvieron un vídeo viral, lo mismo puede ocurrir con compañías startup. Promocionar su producto mediante un video que termina siendo viral le permite al mundo conocer la compañía y sus productos, lo que puede posteriormente incentivar a posibles inversionistas a apoyar a la compañía. En el caso de compañías más grandes, cuando se está promocionando un producto nuevo o un cambio dentro de la compañía un video viral puede resultar en el producto presentándose a una población superior de la que se esperaba inicialmente, con el video alcanzando personas que no conocen/consumen el producto.

Es por esto que consideramos de suma importancia a la hora de realizar un video como un método promocional entender cuales son las características de un video que promueven la viralización, debido a que un video puede resultar costoso de hacer y si no rinde adecuadamente su implementación puede resultar irrelevante para la empresa/persona o incluso puede tener consecuencias negativas para el producto o persona, por ejemplo resultando en que la población objetivo termine ignorando el producto una vez sale a mercado.

En este trabajo nos proponemos generar un modelo predictivo de viralización de un video de youtube utilizando las características “demográficas” de los videos,

entendiendo las características demográficas de los videos como las etiquetas (tags) que definen de qué se trata el video.

Literatura previa

El que hace que un contenido del internet se haga viral y cómo predecirlo ha sido un objeto de estudio importante en las últimas décadas.

Estudios previos han abordado este problema al centrarse en las emociones que genera el contenido y como esto afecta la viralidad del contenido. Berger y Milkman (2012) Hacen un análisis de qué emociones afectan la posibilidad de que un artículo se empiece a compartir en masa y por ende se vuelva viral. En su estudio utilizan artículos del New York Times a los que les asigna una valencia de emocionalidad dictada por un panel de sujetos de prueba y observando posteriormente cuáles artículos eran los que más se veían, utilizando la lista de “Most Emailed List”. Se encontró que los artículos que tenían una valencia emocional más positiva tenían una mayor probabilidad de encontrarse en la lista de más buscados.

Por otro lado, un estudio de Pinto et al. (2013) hace un análisis de predicción popularidad de un video utilizando cantidad de visitas en dos modelos diferentes, un modelo de regresión lineal multivariada, en donde se usa como variables predictoras cantidades de visitas en intervalos específicos de tiempo, y un modelo al que llaman MRBF en el que se analiza probabilidad de popularidad mediante comparaciones entre el video a predecir y videos que están en un training set. Encontrando resultados favorables para ambos modelos a la hora de predecir qué tan popular va a ser un video.

A comparación de estos dos estudios y para agregar un factor más, nuestro estudio se centra en las características del video en sí, centrándose en cual es el contenido que está presentando, basándonos en las etiquetas (tags) que cada video tiene, con el objetivo de observar si ellas son capaces de medir la popularidad de un video. Consideramos que analizar la cantidad de visitas resulta irrelevante ya que se trata de una variable que se explica post viralización antes que pre (el video se compartió mucho y por ende tiene muchas visitas)

Base de datos

La base de datos que se utilizará se dividirá en dos partes, en primer lugar, una base de datos de Kaggle (Sharma, 2023) que contiene todos los datos de los videos virales de youtube en diferentes años y regiones, nosotros nos centraremos en la base de Estados Unidos. El método de obtención de estos datos es utilizando la Youtube API que permite obtener la información de los videos más trending de youtube en el momento. La base de datos contiene información como el nombre del video, el canal que lo publicó, el año de publicación, las tags utilizadas en el video, que funcionan principalmente como una especie de filtro (definiendo de qué se trata el video y hacia qué público apunta), la cantidad de visitas que tiene el video, la cantidad de likes, dislikes y comentarios que tiene el video y variables categóricas de si se permite visualizar los likes y los dislikes y si se puede dejar comentarios en el video.

Si bien es verdad que estos datos ya se pueden utilizar para un análisis de frecuencia y categorización, si se desea hacer un modelo predictivo de probabilidad de que un video sea trending es necesario obtener datos de videos que no son

trending, para ello mediante la misma API de youtube se realizará otra base de datos con las mismas variables utilizando los datos de videos que no son trending. Para ello, se obtendrá mediante web scraping una lista de links a videos que no se encuentran actualmente en la lista de videos trending en youtube y posteriormente se los compilará en una lista igual a la de la otra base de datos utilizando la función list de la API de youtube (Google, 2024). Además, para evitar confusiones se creará una variable “es_viral” que mediante el promedio de la base inicial se obtendrá un valor que si se supera o es similar se considerará a un video como viral, de esa manera se podrá identificar correctamente qué cantidad de visitas definen a un video como viral.

Por último antes de cualquier análisis se separará la variable tag en múltiples variables categóricas, eliminando aquellas que se repiten pocas veces (algunas son los nombres de los artistas o de los canales, por lo que se pueden eliminar si se tratan de casos muy específicos), para luego utilizarlas el en análisis como nuestras variables predictoras.

Metodología

Para el análisis exploratorio inicial se utilizarán gráficos e histogramas para analizar la frecuencia de los datos. Por ejemplo se puede observar las frecuencias de los likes, los dislikes y la cantidad de visualizaciones para tener una idea de cual es la proporción de visitas dentro de la muestra, para poder observar la distribución de vídeos virales y no virales. Para continuar con ese análisis se realizará un gráfico de

barras con la variable creada “es_viral” para observar exactamente cuántos videos son virales y cuantos no dentro de la muestra.

Para complementar el análisis, se realizará también un proceso de clustering, que permitirá identificar grupos naturales dentro de los datos y explorar patrones emergentes entre los videos virales y no virales. Específicamente, se utilizará el algoritmo de K-means o K-medias, conocido por su eficiencia en grandes volúmenes de datos y su capacidad para dividir los videos en grupos con características similares.

El número de clusters podremos pre-definirlos de forma arbitraria o utilizar la técnica denominada “método del codo", que analiza la disminución de la inercia intra-cluster a medida que aumenta el número de clusters, identificando el punto óptimo donde dicha disminución se estabiliza. Los resultados del clustering nos permitirán tener una segmentación clara de los datos, permitiendo explorar diferencias entre videos basados en sus características intrínsecas y categorizarlos en grupos representativos de patrones de viralización.

Para el análisis estadístico y predictivo se realizará dos modelos logísticos que serán penalizados de maneras diferentes uno mediante Lasso y el otro mediante Ridge, con el objetivo de encontrar el mejor modelo posible sin tener un overfit de la muestra de entrenamiento. El lambda se obtendrá mediante un cross validation con K=10 y posteriormente se comparará los dos modelos mediante curvas ROC, AUC, accuracy y error cuadrático medio (MSE), quedándonos con el mejor modelo. El modelo logístico se trata de un modelo que permite, mediante odds, predecir cuál es la probabilidad de que algo ocurra o no, en nuestro caso utilizando la variable “es_viral” como nuestra variable dependiente y las demás variables como variables

independientes, salvo cantidad de likes/dislikes y visualizaciones debido a que las consideramos variables que representan si un video es viral o no en vez de si puede llegar a ser viral (predecir que un video con un millón de visitas o cien mil likes es viral es algo asegurado). Un video trending por definición va a tener mucha interacción por lo que si analizamos mediante estas variables, que definen interacción en el video, se espera que el modelo siempre predice que la probabilidad de que un video sea viral va a aumentar conforme aumenta esta variable y siempre va a predecir que un video con muchas visitas va a ser viral.

Conclusiones y limitaciones

Con la implementación de los análisis presentados en este escrito esperamos poder generar un mejor entendimiento de que es lo que se presenta en un video que lo puede llegar a hacer viral, el análisis de clusters permite diferenciar dentro de la base de datos los diferentes tipos de videos mientras que el modelo logístico permite generar un modelo de evaluación y predictivo para todos los videos.

Se espera encontrar varios resultados clave. En primer lugar, se espera encontrar diferentes agrupaciones de videos virales, insinuando que hay más de una manera por la que un video se puede llegar a hacer viral, ya sea por pertenecer a diferentes categorías de videos (videos musicales, algún videojuego que está de moda, algún evento importante). En segundo lugar, se espera encontrar un patrón de tags que permitan predecir una mayor probabilidad de que un video sea viral, mediante el modelo predictivo que revelen cuáles son las tags y por ende las categorías de videos que mayor probabilidad de viralización tienen, ya sea una agrupación de tags

o incluso una sola. Por último, se espera que este modelo se pueda utilizar para análisis de mercado posteriores, observando diferencias y similitudes entre los videos que son actualmente virales en comparación a videos del pasado o futuro, lo que podría usarse para analizar cambios en las preferencias y gustos de los usuarios de Youtube.

Este trabajo sin embargo, presenta múltiples limitaciones. En primer lugar, la obtención y preparación de los datos resulta complicada debido a las limitaciones de la API de youtube, se puede buscar fácilmente por separado todos los videos que están actualmente en la categoría de “trending”, pero la obtención de los datos de los videos que no están en la pantalla de “trending” resulta más limitado, debido a que se requiere los links de cada uno de los videos, esta evaluación requiere obtener los datos mediante web scraping lo que puede mediante los algoritmos de youtube sesgar las muestras, ya que va a presentar videos que le interesen a aquel que esté haciendo el scrapping. En segundo lugar, si bien los datos presentados en la base de datos son útiles, se podrían considerar autoevaluativos, en ellos se define ampliamente de que se trata el video pero no especifica características como tiempo de duración del video entre otras, un análisis en profundidad de variables de este estilo podría generar un mejor modelo a posterior. En tercer lugar, la base de datos contiene los videos virales de múltiples países pero debido a limitaciones del lenguaje se decidió trabajar con la base de datos de Estados Unidos, por lo que dificulta generalizar los resultados para otros países, un estudio posterior podría utilizar mediante traducciones o incluso todas las variables juntas un modelo que pueda ser capaz de predecir para todos los países. Por último las tendencias

cambian, por lo que para mantener al modelo actualizado se debería entrenar con nuevos videos cada año para mantener la potencia predictiva del modelo.

En conclusión, este estudio tiene como objetivo encontrar qué categorías de videos son aquellos que tienen la mayor probabilidad de volverse un video viral, con el objetivo de generar un modelo que permita a futuro predecir las chances que tiene un video antes de ser creado.

Bibliografía

Berger, J., & Milkman, K. L. (2012). What makes online content viral? Journal of Marketing Research, 49(2), 192–205. https://doi.org/10.1509/jmr.10.0353

Broxton, T., Interian, Y., Vaver, J., & Wattenhofer, M. (2011). Catching a viral video. Journal of Intelligent Information Systems, 40(2), 241–259. https://doi.org/10.1007/s10844-011-0191-2

Google. (2024). *Documentación de la API de YouTube* [Documentación en línea]. https://developers.google.com/youtube/v3?hl=es-419

Pinto, H., Almeida, J. M., & Gonçalves, M. A. (2013). Using early view patterns to predict the popularity of YouTube videos. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining - WSDM '13*. https://doi.org/10.1145/2433396.2433443

Sharma, R. (2023). *YouTube trending video dataset*. Kaggle. https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/data?select=US_youtube_trending_data.csv