ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Desarrollo de una metodología de análisis de datos para un canal de Youtube

PROYECTO INTEGRADOR

Previo la obtención del Título de:

Ingeniero en Computación

Presentado por:

David Alfredo Santistevan Amat

Edgar Javier Vinueza Herrera

GUAYAQUIL - ECUADOR

Año: 2022

DEDICATORIA

Este trabajo es dedicado a mis padres, David y Claudia, y a mis abuelos Alfredo y Yila, quienes me dieron la motivación y fuerza necesaria para culminar este ciclo de mi vida

David Santistevan

Dedico el presente trabajo a mi familia, que ha sido un pilar fundamental en esta etapa.

Edgar Vinueza

AGRADECIMIENTOS

Agradezco a mis padres su esfuerzo y compañía a lo largo de mi vida, especialmente en esta etapa, a mi hermana Pamela por ayudar con su experiencia a mantener la calidad de este trabajo, y finalmente a todos quienes dieron retroalimentación o información utilizada dentro de este trabajo

David Santistevan

Agradezco a Dios por brindarme la sabiduría, a mi familia por acompañarme en este proceso y cada persona que contribuyó para que esto sea posible.

Edgar Vinueza

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; David Alfredo Santistevan Amat y Edgar Javier Vinueza Herrera damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

DAVID ALFREDO SANTISTEVAN AMAT

Firmado digitalmente por DAVID ALFREDO SANTISTEVAN AMAT
Nombre de reconocimiento (DN): cn=DAVID ALFREDO SANTISTEVAN AMAT, serialNumber=311022084904, ou=ENTIDAD DE CERTIFICACION DE INFORMACION, o=SECURITY DATA S.A. 2, c=EC
Fecha: 2022.11.23 23:20:10 -05'00'

David Alfredo Santistevan Amat

Edgar Javier Vinueza Herrera

EVALUADORES

LUCIA MARISOL VILLACRES FALCONI
Digitally signed by LUCIA MARISOL VILLACRES FALCONI
Date: 2022.11.24 20:43:08 -05'00'

Lucía Marisol Wong-Villacres Falconí

PROFESOR DE LA MATERIA

Firmado electrónicamente por:
EDUARDO SEGUNDO CRUZ RAMIREZ

Eduardo Segundo Cruz Ramirez

PROFESOR TUTOR

RESUMEN

Desde el año 2020 se incrementó el uso de medios digitales como fuente de información en distintas áreas, por ejemplo, el área de la salud. Existe una gran variedad de creadores de contenido que proporcionan información relevante sobre salud, pero muchas veces no logran el alcance o popularidad deseados. También, en muchas ocasiones toma mucho tiempo leer los comentarios para comprender los sentimientos de su audiencia. En este proyecto se presenta una solución para los creadores de contenido con estos problemas. Para esto se realizó un análisis de sentimiento usando BERT Transformers, y luego se usó esa calificación del sentimiento para ofrecerle a los creadores de contenido un puntaje final de popularidad usando distintas variables propias del video como visualizaciones y likes, por ejemplo. Finalmente se realizan visualizaciones para facilitar a los creadores la comprensión de estas métricas.

Palabras clave: Redes sociales, Inteligencia Artificial, Visualización de información, Análisis de Sentimiento

ABSTRACT

The usage of digital media as a source of information has risen since 2020 in different areas, such as the healthcare and wellness area. There is a vast number of creators that provide valuable information above healthcare, but most of the time are unsuccessful at achieving their desired reach and popularity. They also have a problem reading the comments and trying to understand their audiences' feelings, because the long time it takes to read through all comments. This project presents a solution to most content creators with these issues. For this we ran a sentiment analysis using BERT Transformers, and then used the obtained score to provide content creators with a final popularity score using also different variables from the content, such as views and likes, for example. Lastly, multiple visualizations are shown to facilitate content creators the comprehension of these metrics

Palabras clave: Social Media, Artificial Intelligence, Data Visualization, Sentiment Analysis

INDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

1.1. Descripción del problema

En el año 2022, el miedo colectivo alrededor de la pandemia de COVID-19 ha disminuido en gran medida, por lo que existen muchas personas que quieren retomar su vida cotidiana y sus actividades de la forma más saludable posible. En este regreso a la normalidad, las personas buscan adoptar y aprender habilidades que velen por su bienestar y salud. Algunas empiezan con dietas y gimnasio, practican deportes y otras deciden hacer actividades en casa para su bienestar (Iwazaki, Aoki, Kato, & Kimura, 2021), dados los riesgos aún existentes de contagios de COVID-19 al realizar actividades físicas. Para esto buscan contenido de salud y bienestar, con el objetivo de mantenerse saludables, realizando actividad física de forma segura dentro de su hogar.

Un medio que se ha popularizado aún más después del 2020 es YouTube (Rodriguez, 2021). Esta herramienta permite acceder de forma segura a información fácilmente comprensible y con una alta posibilidad de generar mejoras en los hábitos y prácticas de salud y bienestar. Existe una gran cantidad de canales que tienen como objetivo mejorar la calidad de vida de su audiencia a partir de actividades recreativas o buena alimentación. Por otra parte, se ha demostrado que muchos de estos canales en efecto logran enseñar información de valor a su audiencia (Green, et al., 2018).

YouTube ofrece a sus creadores de contenido variables numéricas, como visualizaciones, suscripciones, comentarios, retención de audiencia, entre otras. Estas variables se pueden segmentar tanto por cada video, como del canal, como un todo. Esto se hace con el objetivo de que los creadores de contenido puedan tener una idea de su crecimiento y de la aceptación de sus videos. Sin embargo, las variables numéricas no muestran el verdadero sentimiento de su audiencia (Mathieu, 2015).

Conociendo lo útil que es conocer el sentimiento de su audiencia, los canales de salud tienen la necesidad de conocer la aceptación de su contenido y opinión en la plataforma de manera más eficiente que leer comentario a comentario, debido a que, es importante para los creadores conocer si su canal está generando un impacto positivo a sus viewers. Este conocimiento les permite potenciar su crecimiento y popularidad para que su contenido sea consumido por nueva audiencia. Esto es útil particularmente para los canales de salud y bienestar, pues permiten cuantificar el impacto que tienen en la salud y calidad de vida de sus seguidores.

1.2. Justificación del problema

Si bien Youtube ofrece datos estadísticos generales a los creadores de contenido sobre el número de visualizaciones y *likes,* no ofrece formas de lograr un análisis profundo de la recepción del público, el cual es esencial para determinar el "éxito" real de un video, que es altamente influenciado por la recepción del público (Reynolds, 2020). Existen casos, como el del YouTube Rewind (YouTube, 2018), que se trata de un video que tiene muchas visualizaciones y una alta cantidad de likes, pero cuenta con una recepción negativa por parte del público, apreciable dentro de los comentarios y los dislikes del video de ese año (Reynolds, 2020). Este es un caso que se repite a nivel global, sin embargo, no todos los creadores de contenido logran analizar todo el feedback que reciben de todo su contenido, mucho menos los creadores con una alta cantidad de suscriptores.

Adicionalmente, estas plataformas tienen un algoritmo que muestra primero el contenido que el sistema cree que el usuario desea ver, incluidos los comentarios, y de esta forma los creadores de contenido no notarían efectivamente los comentarios negativos, aparte de que, en muchas ocasiones, le será difícil a un creador de contenido mantener completa objetividad al evaluar la recepción del contenido que publicó.

Los creadores de contenido, especialmente los canales de salud y bienestar, también como aquellos con una alta cantidad de suscriptores quienes, por lo general, tienen altas probabilidades de impacto sobre la población, requieren

apoyo para analizar la enorme cantidad de retroalimentación que reciben, no solo de un video, sino de todo su contenido en forma general.

Un modelo automatizado para análisis de comentarios, posiblemente de inteligencia artificial les brindaría el apoyo necesario. De esta forma estos creadores de contenido tendrán la capacidad de realizar publicaciones conociendo los sentimientos de su audiencia respecto a su contenido. Por ejemplo, los canales de salud y bienestar podrán conocer si los ejercicios que realizan benefician a su audiencia o no.

1.3. Objetivos

1.3.1. Objetivo General

Construir un modelo para medir la popularidad de un canal de bienestar y salud en YouTube, a partir de un análisis de sentimiento de los comentarios que realizan los usuarios en los videos.

1.3.2. Objetivos

- Diseñar un modelo de inteligencia artificial para realizar un análisis de sentimientos de los comentarios en videos de un canal de YouTube.
- Analizar la correlación que puede tener el sentimiento de los comentarios, me gusta, visualizaciones.
- Diseñar un modelo capaz de relacionar múltiples variables o factores para calificar la aceptación de un video frente a su población objetivo.
- Implementar un dashboard en Power BI con los resultados obtenidos del modelo.

1.4. Marco teórico

1.4.1. Herramientas

YouTube es un sitio web dedicado a la publicación y visualización de videos. Es considerada una de las plataformas dedicadas a videos más utilizadas, especialmente para videos de corta y mediana duración. Esta ofrece herramientas a sus creadores de contenido dentro de un dashboard llamado “YouTube Studio”,

en el que principalmente se muestra un análisis del crecimiento y la popularidad del canal y de los videos en función de variables numéricas (Google, 2022). Dentro de esta plataforma se pueden visualizar los datos en gráficos generados por la misma aplicación, lo que le permite al usuario tener una mejor idea del movimiento de su cuenta.

Estas herramientas son importantes para los creadores de contenido, pero no terminan siendo completamente eficaces ya que, solo ofrecen análisis numéricos y estadísticos de su audiencia. A pesar de ser útiles para tener una idea de su impacto con su audiencia y su popularidad, es necesario un análisis cualitativo (Mathieu, 2015). Este análisis cualitativo permitiría a los creadores de contenido a tener un criterio más realista sobre su contenido para así poder llegar a su audiencia deseada.

1.4.2. Popularidad

Hasta ahora existen varios estudios sobre la popularidad en YouTube que se basan netamente en los datos numéricos de los videos. Un estudio utilizó las series de tiempo de número de visitas para estimar el número de visitas en futuros videos (C. Richier, 2015).

Otro estudio se concentró en encontrar las variables numéricas más importantes, además exploraron la relación entre el número de subscriptores junto con el número de visitas concluyendo que solo para canales populares el número de visitas del canal afecta el número de subscriptores (W. Hoiles, 2017).

El video, además de los datos numéricos que tiene, este nos puede dar características del mismo video, un estudio se encargó de extraer características tales como el color, si existen caras presentes o rigidez para elaborar un soporte de regresión vectorial (SVR) para predecir la popularidad (Rokita, 2015).

Como se mencionó anteriormente, las variables cuantitativas relacionadas a un video como la cantidad de usuarios que dieron clic al botón de “Me Gusta” y las reproducciones alcanzadas pueden servir para predecir preliminarmente la popularidad de un video (Welbourne & Grant, 2016), no bastan para determinar el

éxito de un video (Chang, Chen, & Verkholantsev, Revisiting Online Video Popularity: A Sentimental Analysis, 2019).

1.4.3. Análisis de sentimiento

Por esto, se plantea el uso de un análisis de sentimientos, pues se ha probado que se puede predecir el éxito de cierto contenido usando esta herramienta en videos de YouTube (Timani, Shah, & Joshi, 2019). A su vez, se ha mostrado que los sentimientos manifestados en un video pueden influenciar en la popularidad del mismo video (Chang, 2018).

Conociendo que un análisis de sentimiento del contenido de un video puede ayudar a predecir la popularidad de un video (Fontanini, Bertini, & Del Bimbo, 2016), es pertinente evaluar las distintas posibilidades de realizar esta práctica sobre otras características de los videos. Asghar, et. al (2015) por ejemplo, presentan distintas técnicas para realizar análisis de sentimiento sobre los comentarios que emiten los visitantes de un video. Este análisis, sin embargo, puede encontrarse con problemas que inhiben su precisión. Por ejemplo, el sarcasmo en los comentarios, usando palabras positivas para expresar sentimientos negativos, puede confundir un resultado. Pandey (2019) muestra una solución capaz de resolver este problema, pero dicho modelo solo determina si un comentario es sarcástico o no, mas no lo relaciona con un sentimiento positivo o negativo del comentario.

Si bien existen muchas alternativas de análisis de sentimientos, la más conocida determina dos o tres emociones (bueno, malo y neutral) de forma general y unidimensional (Novendri, Callista, Pratama, & Puspita, 2020). Sin embargo, existen distintas alternativas adicionales que Alhujali & Yafooz (Alhujali & Yafooz, 2021) detallan y categorizan los distintos acercamientos a resolver esta problemática que mantienen múltiples emociones y le dan un valor en proporción de la magnitud de la emoción, lo que permite cuantificar mejor cuando un comentario es muy positivo y cuando es levemente positivo (Bhuiyan, Ara, Bardhan, & Islam, 2017).

Todo el análisis de sentimiento de los comentarios y contenido de YouTube mostrado se hace con el objetivo de evaluar la popularidad y sentimiento general

de los videos a evaluar. Para esto existen actualmente varios métodos que permiten evaluar una puntuación final para videos relacionados a la medicina y la salud (Drozd, Couvilllon, & Suarez, 2018). Esto sería útil para cubrir la necesidad de evaluar el valor esperado de un video al momento de reconocer los factores de éxito.

CAPÍTULO 2

2. METODOLOGÍA

2.1. Resumen del capítulo

Este capítulo explica los pasos que se han seguido para resolver el problema de predecir la popularidad de un video existente, usando el data set proporcionado inicialmente.

Para trabajar sobre el data set, es necesario comprenderlo y su estructura, por esto para iniciar este capítulo se presenta una descripción de los 2 data set, el origen en común de ellos y algunas graficas comparativas entre los datos que estos ofrecen.

En la extracción de características se describe la etapa del preprocesamiento, donde se agrega información de otras fuentes, y la postulación de una función “score” para el sentimiento de los comentarios por cada video, así como también se muestra el etiquetado manual de comentarios, que permite entrenar el modelo de inteligencia artificial.

Finalmente, en la sección de evaluación y selección de modelos se describe formalmente los modelos de IA estudiados, y se aclara cuál se va a usar para el correcto análisis de los videos y predicción de la popularidad. Además, se definen los sentimientos usados en el análisis de sentimiento, cómo el análisis de sentimiento lleva a una calificación general de sentimiento, y una fórmula de popularidad por video gracias a la revisión literaria.

A continuación, en la

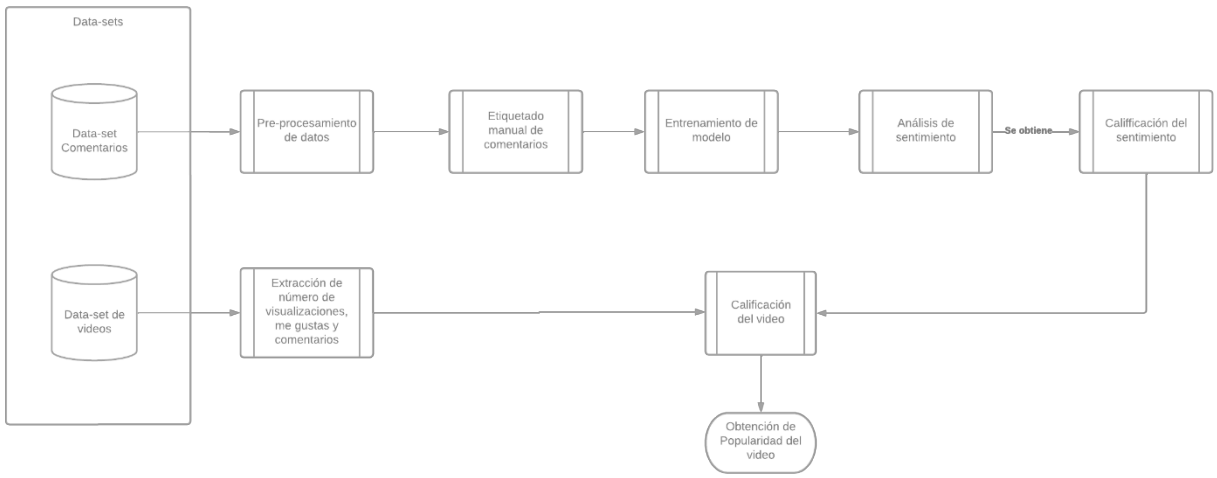

Figura 1 se presenta un resumen ilustrativo con los pasos y procedimientos tomados en la metodología.

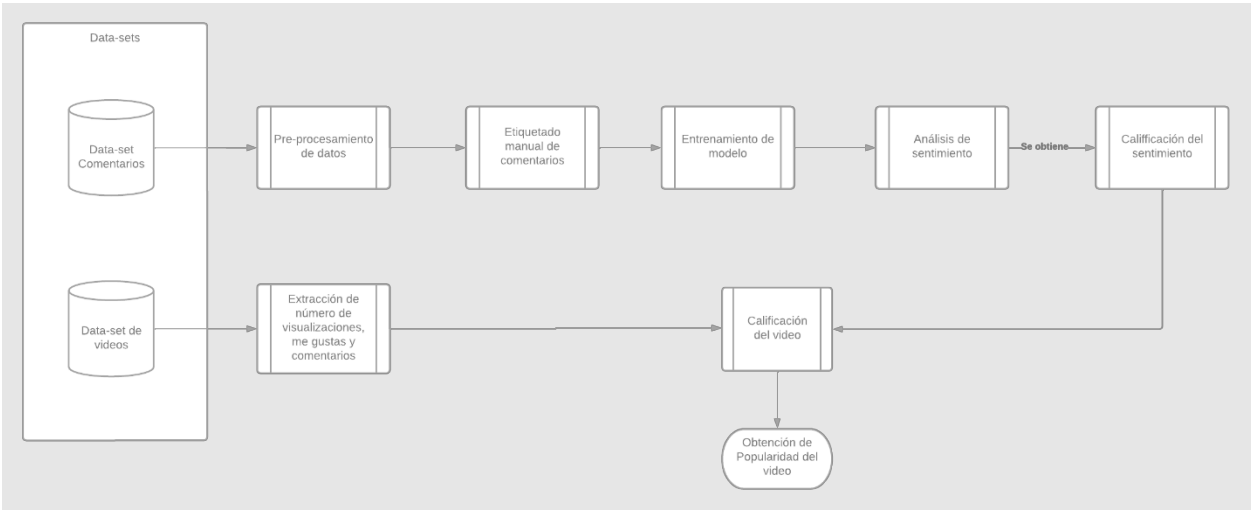

Figura 1 Metodología

2.2. Obtención de los datos

Se tienen dos data sets que contienen información de 252 videos a lo largo de 8 años del canal “Yoga With Adriene”, los cuales forman parte de una serie de videos que este canal lanza cada año como un reto donde se sube 1 video diariamente por 1 mes.

El primer data set son las estadísticas generales de los videos. Este data set se agrupa por año y cada uno consta de aproximadamente 31 videos. Estas estadísticas son las variables numéricas más representativas de los videos
El segundo data set son los comentarios por cada video expuesto en el primer data set.
Este data set está conformado por la información de cada comentario
Los dos data sets fueron entregados por la Universidad de Cardiff.

2.2.1. Descripción de los datos

2.2.1.1. Primer Data Set

Los datos fueron obtenidos de 240 videos desde enero del 2015 hasta enero del 2022. Cada video tiene información relevante, ver Tabla 1

Variable	Valor
Title	TRUE - 30 Day Yoga Journey | Begin!
Description	Kick-off the New Year with 30 Days of Yoga With Adriene! - - - - - - #30daysofyoga #ywatrue #homeyoga
Published	2017-12-22T15:09:26Z
Tag_count	23
View_count	2223866
Like_count	20486
Comment_count	1684
Duration_str	3:29
Title_length	37
rections	23854

Tabla 1 Datos del video 1 del 2015

De este primer data set se lograron múltiples gráficos estadísticos. La cantidad de visualizaciones por cada año a lo largo de 30 días se refleja que obtuvo la mayor cantidad de visitas en los primeros años del data-set, ver Figura 2

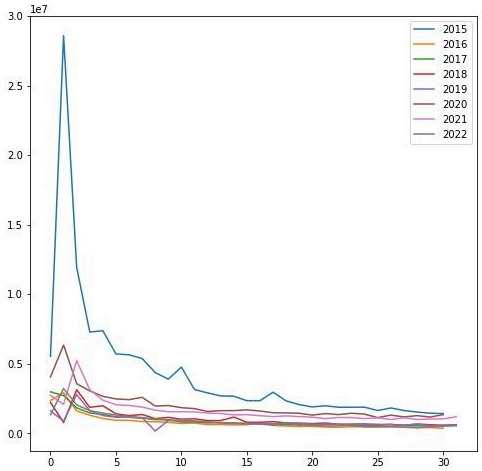

Figura 2 Gráfica estadística de las visitas a lo largo de los 30 días

De la misma forma, la cantidad de Me gusta por cada año a lo largo de 30 días, en este y en la gráfica anterior, se refleja una mayor cantidad de interacciones en el primer año, y en cada año, el primer día de la serie de videos, ver Figura 3.

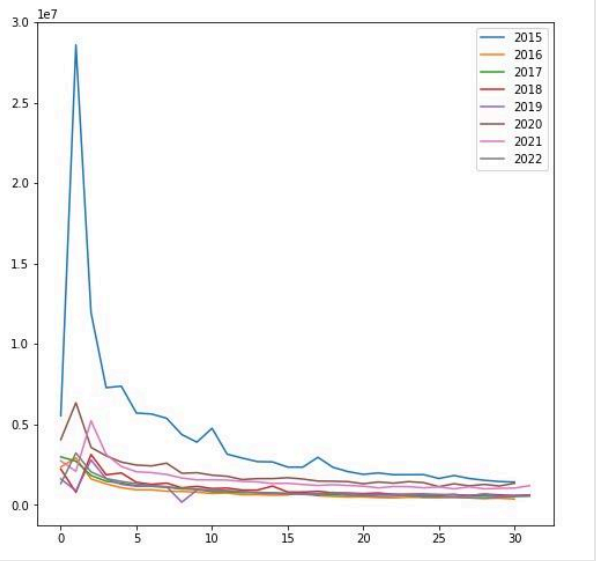

Figura 3 Gráfica estadística de los Me gusta a lo largo de los 30 días

Se relacionó duración vs visitas de los videos a lo largo de los 8 últimos años en una gráfica de dispersión, ver Figura 4.

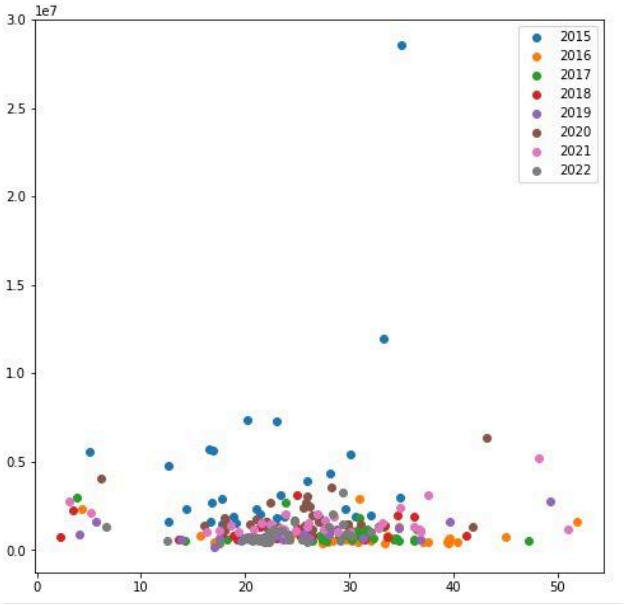

Figura 4 Gráfica estadística de duración vs visitas

Finalmente se obtuvo una gráfica relacionando la proporción de Me gustas sobre visualizaciones en orden cronológico, aquí se puede rescatar una tendencia creciente en el tiempo de esta proporción, ver Figura 5.

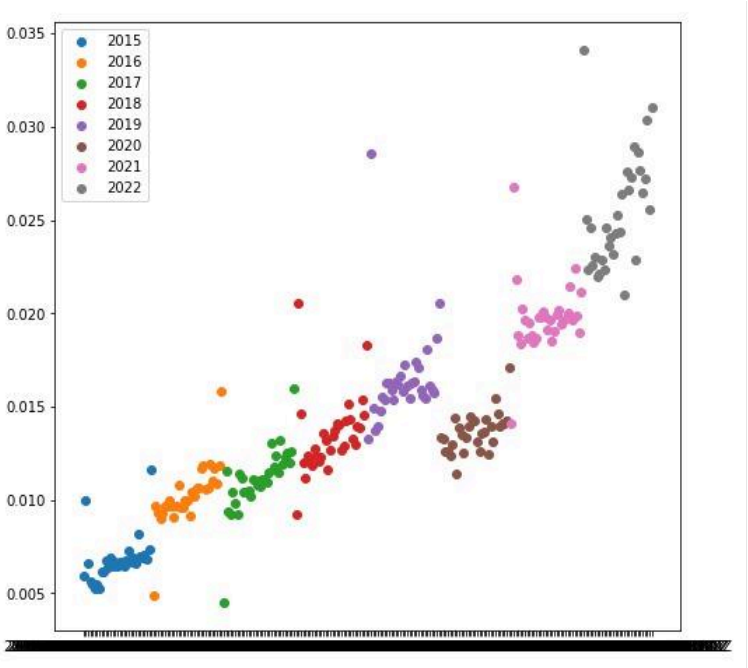

Figura 5 Proporción Me gustas/visualizaciones en orden cronológico

Contrastando con la Figura 2 y la Figura 3, se puede observar que, si bien el volumen de visualizaciones y Me gustas disminuye cada año, la relación de cuántos me gusta hay por cada visita se ha incrementado año a año. También, se puede observar que el 2020 hubo una caída en esta tendencia, posiblemente atribuible al aumento de personas consumiendo este contenido a raíz de la cuarentena que hubo aquel año a raíz del brote de COVID-19 a nivel mundial.

2.2.1.2. Segundo Data Set

Este data set está conformado los comentarios de cada video del primer data set.

Cada comentario de un video contiene el comentario, contador de respuestas, contador de Me gustas y fecha de actualización del comentario, ver Tabla 2

Variable	Valor
comment	Starting May 1st 2021! I just finished the original 30 days of yoga yesterday and i was shocked at how good i felt throughout the month, so im keeping it going with Revolution!
Reply_count	0
Like_count	1
Updated_at	2021-05-02T01:20:49Z

Tabla 2 Datos del comentario 42 del video 1 del 2017

2.3. Extracción de características

2.3.1. Pre-procesamiento de datos

Si bien existe una buena cantidad de información en el data set recopilado inicialmente, se usó el YouTube Data API (Google, 2021) para agregar los

tags que contiene cada video, para determinar si existe alguna relación de estos con los videos.

También, se formateó el valor de la duración para tenerla como un valor numérico, que originalmente era texto plano, con el objetivo de permitir a los modelos cuantificar el tiempo, mas no leerlo como texto.

2.3.2. Etiquetado manual de comentarios

Para el data set de comentarios, se procedió a asignar valores en distintas categorías para cada comentario, con la finalidad de realizar el análisis de sentimiento que tiene como objetivo, obtener una puntuación de los comentarios procesados para poder usarla en la fórmula de popularidad.

Para esto, se usó la segmentación dada por Madden, Ruthven, & McMenemy (2013) para comentarios de YouTube, y se definieron las siguientes categorías:

- Información
    - Si da información de cualquier tipo, sea sobre si mismo o su entorno, si la información se considera negativa, se asigna un valor negativo, si se considera positiva, un valor positivo.
- Opinión
    - Si el usuario ejerce su opinión en el comentario.
- Consejo
    - Si da o pide consejos/recomendaciones.
- Impresión
    - Impresión o sorpresa en el comentario, sea que se sorprende que le salió bien, o algo que salió en el video.

Adicionalmente, se encontraron las siguientes categorías de comentarios:

- Progreso
    - Cómo se siente sobre cómo ha progresado física o mentalmente
- Especifico con el ejercicio
    - Cómo se siente sobre el ejercicio mostrado en el video

- Estado del cuerpo
    - Cómo se siente respecto a su cuerpo, si le duele algo o tiene alguna enfermedad es negativo, si se siente bien es positivo
- Especifico con el contenido del video como tal
    - Si habla sobre si le gustó el video, o si habla de algo específico del video

Estas categorías se escogieron al leer los comentarios en los videos, y se reconocieron como los temas más comunes. Esto servirá para dar información adicional al cliente sobre la interacción de los usuarios con el canal que se va a utilizar como referencia.

2.4. Evaluación y selección de arquitectura

Con los data sets preprocesados y con la obtención de las funciones de "score" por comentarios y popularidad del video, se eligen los modelos para el análisis de sentimiento de los comentarios y predicción de la popularidad de los videos.

Para el análisis de sentimiento se eligió la arquitectura de procesamiento de lenguaje natural, BERT, sobre la cual se realizará una. Mientras que para la predicción se va a usar las distintas variables cuantitativas de cada video, y el resultado del análisis de sentimiento

2.4.1. Arquitectura BERT

Bidirectional Encoder Representations from Transformers (BERT), ver Figura 6 es un modelo para procesamiento del lenguaje natural (PLN) desarrollada por Google.

Esta técnica basada en Deep learning, cuyo factor diferenciador es aplicar entrenamiento bidireccional de Transformers, es una de las arquitecturas más populares del estado del arte para el procesamiento del lenguaje natural.

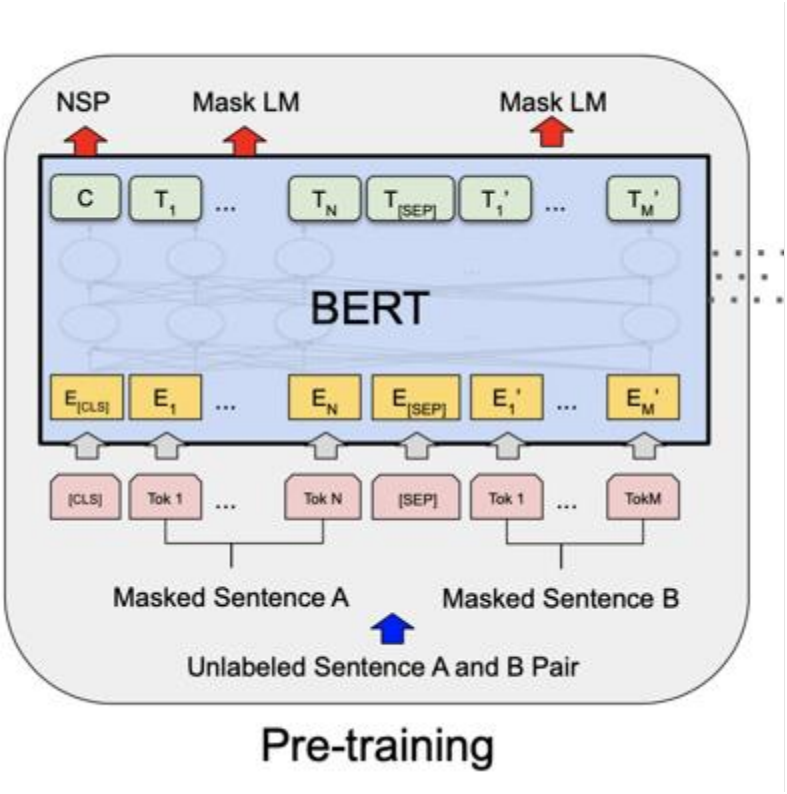

Figura 6 Arquitectura Bert

Bert tiene dos etapas de entrenamiento (Sudarshan, 2021), que son las siguientes:

2.4.1.1. Masked Language Model

Antes de introducir secuencias de palabras en BERT, Un 15% de las palabras de cada secuencia se reemplazan con un token. Luego, el modelo intenta predecir el valor original de las palabras enmascaradas, según el contexto proporcionado por las otras palabras no enmascaradas en la secuencia (Sudarshan, 2021).

2.4.1.2. Next Sentence Prediction (NSP)

La segunda etapa es Next Sentence Prediction (NSP), el cual se basa en relacionar oraciones ya que por esta relación con es capturada directamente por el modelamiento del lenguaje. En esta etapa el modelo se pre entrena para una tarea binarizada de predicción de la siguiente oración

(Devlin, Chang, Lee, & Toutanova, 2018). Este modelo resulta ideal para el procesamiento de lenguaje natural y para clasificar los comentarios por sentimiento, además se puede hacer Transfer learning por lo que facilitaría la etapa de entrenamiento del idioma.

Adicionalmente al modelo, se agregará una capa linear con cada característica, y se tendrán 8 modelos, que equivalen a las 8 categorías nombradas anteriormente. Esta última capa de salida determinará el valor de cada sentimiento.

Para este modelo se va a usar un modelo pre entrenado (Devlin, Chang, Lee, & Toutanova, 2018), que permita solo entrenar el clasificador final, y este cuantifique los distintos sentimientos entre $[-1,1]$, de forma que se tenga una idea de los sentimientos existentes en todos los comentarios.

2.4.2. Calificación del sentimiento

Una vez obtenidos los sentimientos de los comentarios, hay que llegar a una calificación final para el video. Para esto se aplican las siguientes fórmulas:

$$suma(x) = \begin{cases} 0; & |x| < 0.1 \\ x; & 0.1 \leq |x| \leq 1 \end{cases}$$

Aquí, se anulan los comentarios que tienen un sentimiento cercano a 0, porque se consideran neutrales ante ese sentimiento, y si se los llega a considerar no se tendría una respuesta apropiada al sentimiento mostrado por los usuarios.

$$Conteo(X, k) = \sum_{i=1}^{n} suma(X_{ki})$$

Aquí, para cada sentimiento k, se realiza la sumatoria de los comentarios de dicho sentimiento X_k, para cuantificar el sentimiento general.

$$Sentimiento(X) = \frac{\sum_{i=1}^{4} Conteo(X, i)}{n}$$

Finalmente, el sentimiento general del video es la media ponderada de los sentimientos, donde los pesos son la cantidad de comentarios que tienen dicho sentimiento, y dividido para la cantidad total de comentarios. Este cálculo se va a realizar únicamente con los sentimientos generales de los videos, para que sea replicable para cualquier otro tipo de canal.

2.4.3. Calificación de video

Para calificar la popularidad de un video, se consideró el valor propuesto en la literatura (Mekouar, Zrira, & Bouyakhf, 2017), y con esto se va a utilizar la siguiente fórmula:

$$P = \frac{\text{sentimiento} \cdot (\alpha \cdot \text{views} + \beta \cdot \text{cantcomentarios} + \gamma \cdot \text{likes})}{dias_{publicado}}$$

Los parámetros son $\alpha = 0.1$, $\beta = 60$ $\gamma = 3$. Con esta popularidad, se puede cumplir el objetivo de predecir el éxito de contenido futuro al tener una cuantificación proporcional a las visualizaciones de los videos. De esta forma no existe un sesgo por los videos con más tiempo publicados.

CAPÍTULO 3

3. RESULTADOS

3.1. Resumen del capítulo

En este capítulo se muestran los resultados del entrenamiento del modelo BERT usado dentro del contexto de los comentarios del canal Yoga With Adriene. Posteriormente se muestra el uso de este modelo de lenguaje para calificar los sentimientos existentes en los comentarios agregando una capa de calificación tal y como se detalló en el capítulo 2.

Para concluir se muestra cómo se usaron los modelos para calificar los comentarios de los 252 videos, y se usó esta calificación para la puntuación de popularidad de los videos. Finalmente, se muestran las visualizaciones realizadas en Power BI de los resultados obtenidos.

3.2. Entrenamiento del modelo

3.2.1. Modelo de lenguaje

Usando el modelo BERT base (Devlin, Chang, Lee, & Toutanova, 2018), se lo entrenó para entender de mejor manera los comentarios dentro del contexto del canal de Yoga With Adriene. Dentro del contexto, el modelo antes del fine-tuning obtuvo el resultado expuesto en la Figura 7, al aplicar Masked Language Model.

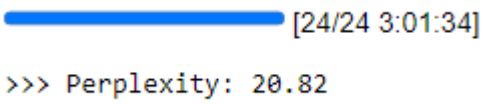

Figura 77. Perplejidad del modelo BERT pre entrenado.

Esto significa que para adivinar una palabra dentro del Masked Language Model el modelo duda entre 21 palabras aproximadamente como la opción correcta. Una vez se aplicó el entrenamiento con todos los comentarios, se llegó al siguiente resultado en la Figura 8.

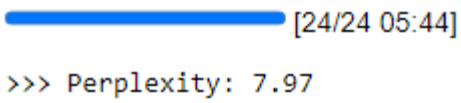

Figura 88. Perplejidad del modelo una vez entrenado con los comentarios del canal Yoga With Adriene.

Se puede apreciar que sí se logró que el modelo aprenda del lenguaje de los comentarios, moviendo la perplejidad del modelo a aproximadamente 8, a diferencia de 21 como era antes del entrenamiento.

3.2.2. Resultados del entrenamiento

Se usaron 8 modelos con la misma arquitectura, mostrada en la Figura 9.

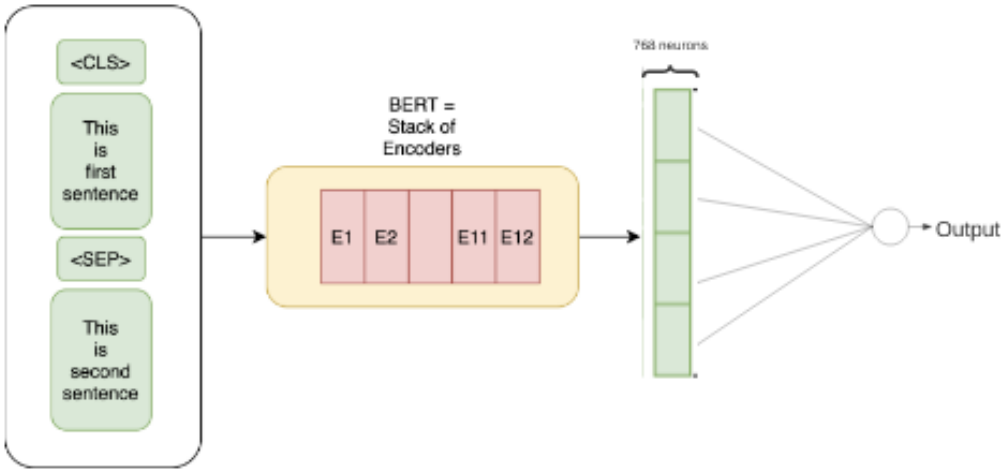

Figura 99. Arquitectura del modelo calificador de comentarios usando una capa BERT.

Existe una función de activación ReLu salida de la capa BERT, y un dropout de 0.3 en dicha capa. La capa BERT es la misma para los 8 sentimientos, el modelo entrenado en la sección anterior, y solo cambian los pesos de la capa lineal para calificar cada uno de los sentimientos. Un ejemplo de los 8 entrenamientos se muestra en la Figura 10 y la Figura 11, para el sentimiento de progreso.

```
Epoch 38/40
----------
Train loss 0.21248566802768482 accuracy 9.685258964143426
Val   loss 0.24812102876603603 accuracy 9.119521912350598

Epoch 39/40
----------
Train loss 0.21356012045391023 accuracy 9.685258964143426
Val   loss 0.2479484905488789 accuracy 9.119521912350598

Epoch 40/40
----------
Train loss 0.21320170081324047 accuracy 9.685258964143426
Val   loss 0.24814549693837762 accuracy 9.119521912350598

CPU times: user 26min 46s, sys: 7.2 s, total: 26min 53s
Wall time: 27min 7s
```

Figura 1010. Salidas del proceso de entrenamiento del modelo de regresión del sentimiento progreso.

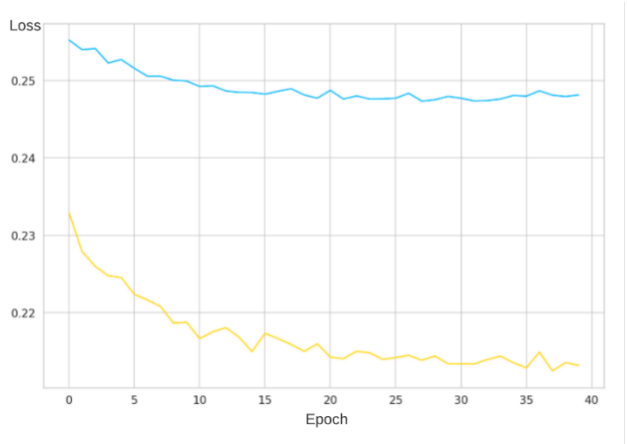

Figura 1111. Pérdida de entrenamiento (amarillo) y de validación (azul) del modelo en cada época de entrenamiento.

Sin embargo, este entrenamiento dio los siguientes resultados en la Figura 12.

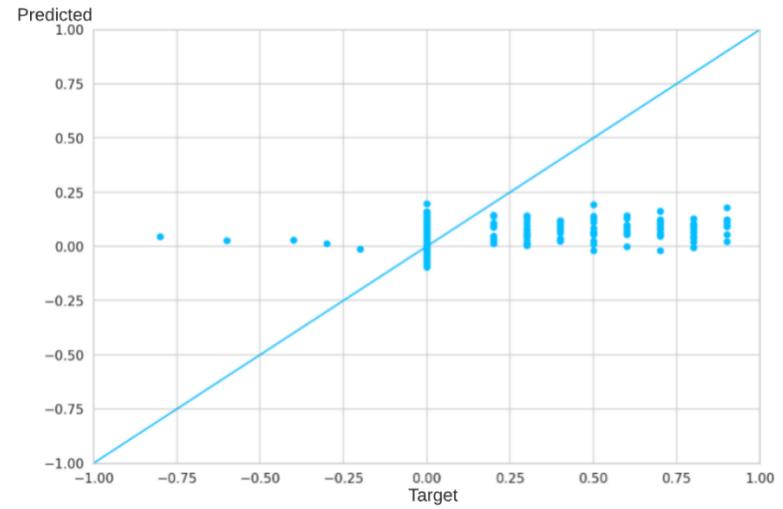

Figura 1212. Salidas del modelo para los comentarios (eje Y), y el valor del sentimiento en el comentario (eje X).

Esta es una tendencia entre todos los sentimientos, por lo que decidimos aplicar una normalización de los datos de salida y obtuvimos lo mostrado en la Figura 13:

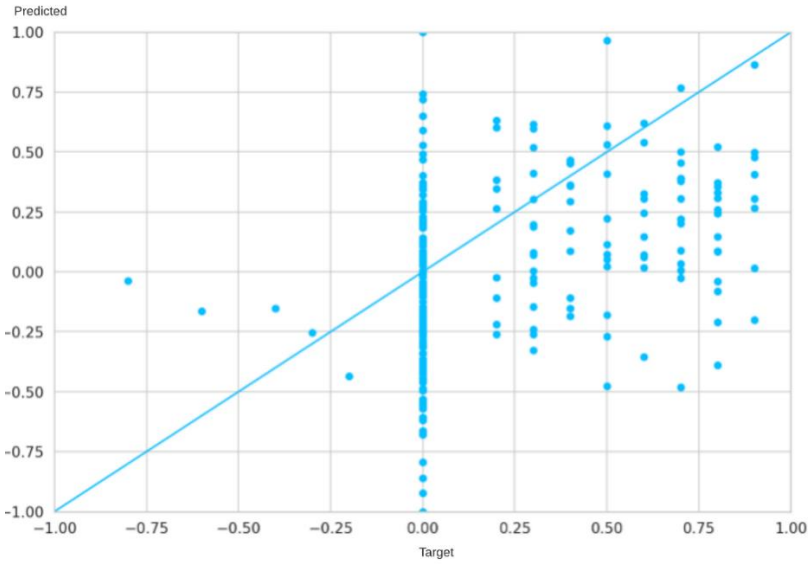

Figura 1313. Salida del modelo normalizada vs valor objetivo del sentimiento.

De esta forma apreciamos que el modelo sí reconoce que los comentarios positivos deben tener un valor mayor que los comentarios con el sentimiento negativo, pero tienen un problema cuando es un comentario neutral. Con esta normalización se llega a un error de 0.3497 unidades en el sentimiento de impresión. Si bien el error aumenta en el data set de

testing, esto nos da un mejor resultado en comparación con el modelo que da valores cercanos a la media del sentimiento en los datos de entrenamiento. Finalmente se llegó a los siguientes errores de entrenamiento en la Tabla 3.

Sentimiento	Error modelo	Error normalizado
Información	0.2469	0.3588
Opinión	0.2265	0.3749
Consejo	0.1056	0.2783
Impresión	0.2554	0.4312
Progreso	0.2273	0.3497
Ejercicio	0.1737	0.2931
Cuerpo	0.1702	0.3389
Video	0.2527	0.3329

Tabla 3. Errores finales de cada sentimiento.

3.3. Modelo final

Con los 8 modelos ya entrenados, se procedió a usarlos para calificar los 374 mil comentarios. Este proceso se realizó para cada sentimiento, tomando alrededor de 2 horas por cada uno, ver Figura 14.

```
CPU times: user 1h 58min 54s, sys: 10.2 s, total: 1h 59min 4s
Wall time: 2h 1min 25s
```

Figura 1414. Tiempo de calificación de un sentimiento para los 374 mil comentarios.

Con un tiempo total de aproximadamente 16 horas, se logró calificar los comentarios de todos los videos. Se llegó a una distribución de los sentimientos como la mostrada en la Figura 15.

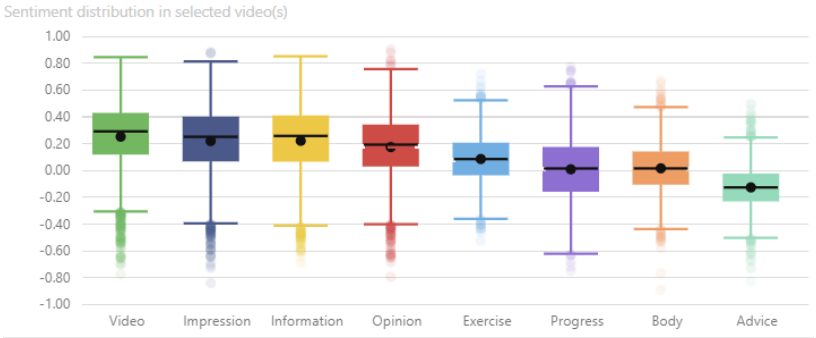

Figura 1515. Distribución de sentimientos en los comentarios.

Algo importante a destacar es que los sentimientos que tuvieron el error más bajo una vez normalizado tienen un rango intercuartil más pequeño que los de mayor error, lo que puede implicar un sesgo al cero, pues también estos sentimientos son los que tuvieron más comentarios donde su valor fue cero.

3.4. Puntuación de popularidad

Una vez se tuvo el valor de cada sentimiento en todos los comentarios, se pudo aplicar nuestra fórmula de popularidad para cada video planteada en la sección 2.4.3, de esta forma se calificó el score de popularidad de los videos, llegando a la distribución por años en la Figura 16.

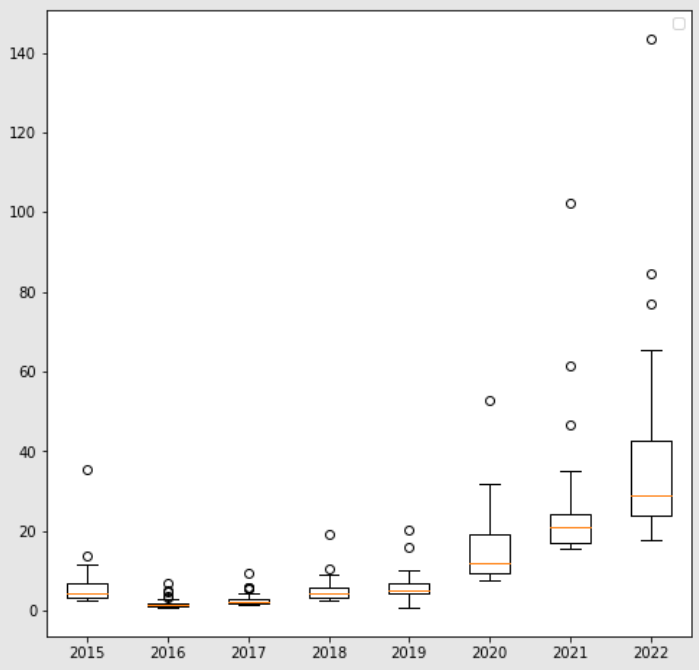

Figura 1616. Distribución de la puntuación de popularidad por cada año.

Sin embargo, es muy difícil saber si un video es exitoso o no simplemente con el valor, por lo que analizamos los valores de popularidad en los distintos percentiles.

Percentil	Valor de popularidad
10	1.639
20	2.641
30	3.439
40	4.507
50	5.743

60	9.009
70	15.608
80	20.999
90	28.863
100	143.550

Tabla 4. Valores de popularidad en los percentiles múltiplos de 10 entre todos los videos.

De esta forma determinamos, basados en los percentiles de la Tabla 4, que un valor de popularidad de 4 sería considerado como satisfactorio, pues se encuentra entre el percentil 30 y 40, y de la misma forma, se determinó 20 como un éxito total, que es un valor muy cercano al percentil 80.

3.5. Dashboard funcional

Se realizó el dashboard en Power BI, usando los datos actualizados con los sentimientos en los comentarios y la calificación de popularidad

3.5.1. Sección de introducción

Se incluyó una sección que sirva de introducción a los usuarios que comiencen a usar el dashboard. Esta incluye una breve explicación de este y una sección de navegación. Ver Figura 17

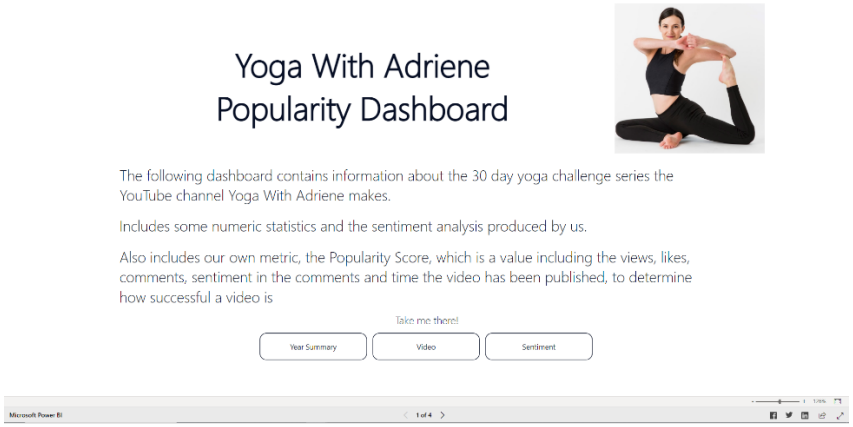

Figura 1717. Sección de introducción del dashboard realizado en Power BI.

3.5.2. Resumen anual

Se realizó una sección que otorga información separada por años, para que se pueda juzgar los números de cada serie completa del canal Yoga With Adriene como un todo. Se visualizan los totales por año de cada una de las

variables numéricas utilizadas (visualizaciones, me gustas y comentarios), además de la puntuación promedio de popularidad de cada año y el porcentaje de visualizaciones que dejaron un me gusta en los videos de cada año. Se da una opción de filtrar por año para solo ver los valores de un año específico, o de varios años. Ver Figura 18.

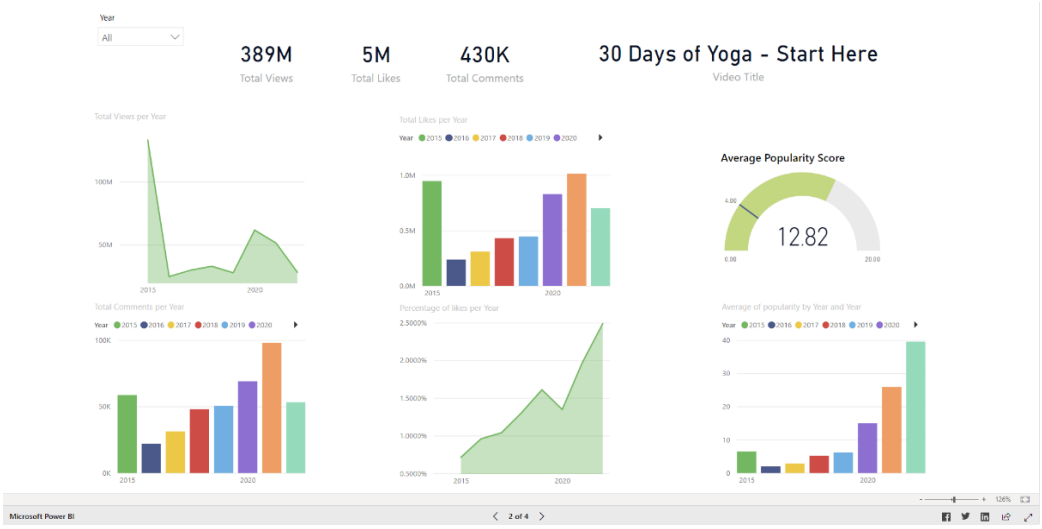

Figura 1818. Vista del resumen anual en el dashboard.

3.5.3. Resumen por videos

Aquí se incluyen las mismas variables, sin embargo, no en forma de resumen, si no separadas por cada año y cada día, en otras palabras, por cada video. Se puede filtrar por año y día. Adicionalmente, como se puede seleccionar cada video individualmente, se incluye el título y la descripción del video, para darle la capacidad al usuario de identificar de qué video específicamente se está mostrando información. Ver Figura 19.

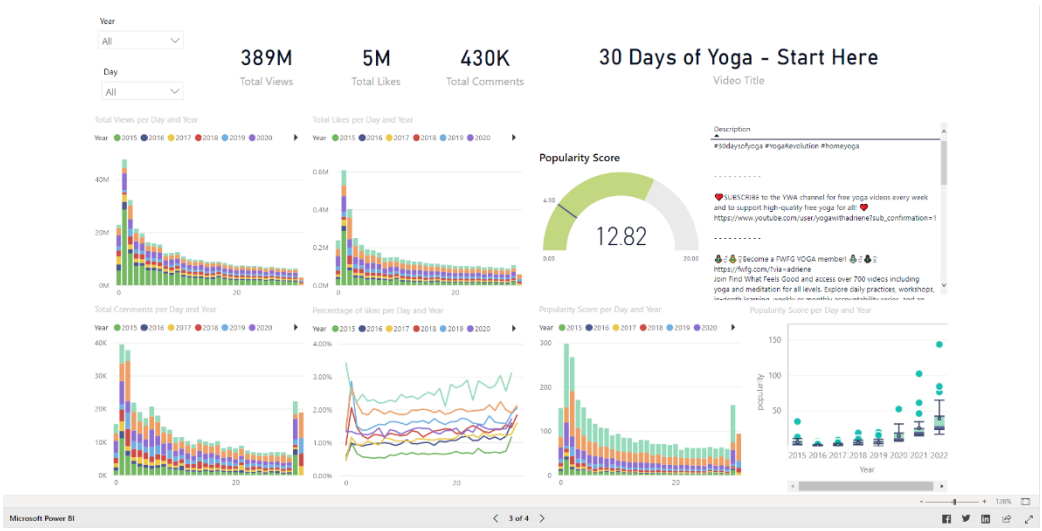

Figura 1919. Vista del resumen por videos del dashboard.

3.5.4. Resumen de sentimientos

En la sección de sentimientos, se pueden aplicar los mismos filtros que en la sección de videos, y esta muestra información relacionada a los sentimientos de los comentarios. Muestra el título del video de la misma forma que la sección anterior para identificarlo, la puntuación de sentimiento general obtenida a través de la metodología explicada en la sección 2.4.2, y otros gráficos como la distribución de los sentimientos en los videos seleccionados, así como la distribución de los sentimientos por día y la serie de tiempo de la evolución del sentimiento a través de los días para cada año. También se incluye una vista de tabla de los comentarios con la puntuación de los 8 sentimientos que tienen, para dar la posibilidad de ver qué comentarios son los más significativos en algún sentimiento. Ver Figura 20.

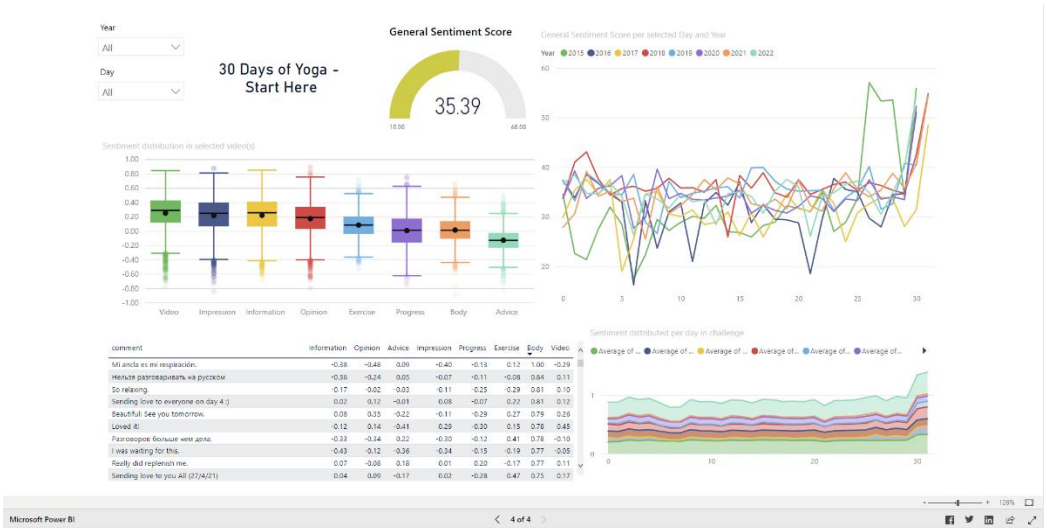

Figura 2020. Vista del resumen por sentimientos en el dashboard.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

4.1. Resumen del capítulo

En este capítulo se presentan las conclusiones en base a los resultados obtenidos y los objetivos específicos dispuestos al principio del proyecto, además un listado de recomendaciones que pueden ser consideradas para futuros proyectos en la misma área de estudio.

4.2. Conclusiones

- En base a los resultados obtenidos por el modelo de análisis de sentimiento a los comentarios de cada video y el puntaje de popularidad calculado, se confirmó que existe una fuerte relación entre el éxito de un video con el sentimiento de la audiencia en los comentarios.
- La metodología desarrollada puede ser usada para cualquier contenido en redes sociales, replicando los distintos pasos especificados.
- Se puede agregar sentimientos específicos al replicar la metodología en otro caso.
- La precisión de la metodología se ve afectada directamente por el tamaño del dataset de entrenamiento, pues esto podría eliminar la necesidad de aplicar la normalización de los datos de salida del modelo calificador.
- Dado que el entrenamiento del modelo con 8 categorías de comentarios duró 6 horas, si aumentaran las categorías específicas, el tiempo de entrenamiento también, por lo que sería necesario más recursos computacionales.
- Se implementó con éxito un dashboard intuitivo en PowerBI, este consta de 3 páginas de gráficos estadísticos con filtros de tiempo para visualizar la información de forma segmentada.

4.3. Recomendaciones

- Se debe aumentar el dataset para mayor precisión en la calificación de comentarios

- Se debe aumentar el número de personas para la precalificación de los comentarios por el tiempo que consume.
- Se recomienda distribuir los comentarios negativos y positivos equitativamente para que aumentar la precisión.
- Se recomienda incluir una capa de preprocesamiento para eliminar comentarios en otros idiomas, o usar algún modelo de lenguaje de varios idiomas.

5. Referencias

C. Richier, R. E. (2015). Forecasting online contents' popularity.

W. Hoiles, A. A. (2017). Engagement and Popularity Dynamics of YouTube Videos and Sensitivity to Meta-Data. *IEEE Transactions on Knowledge and Data Engineering*.

Rokita, T. T. (2015). Predicting popularity of online videos using support vector regression.

Green, J. C., Aziz, T., Joseph, J., Ravanam, A., Shahab, S., & Straus, L. (2018, Marzo). YouTube Enhanced Case Teaching in Health Management and Policy. *Health Professions Education, 4*(1), 48-58. doi:https://doi.org/10.1016/j.hpe.2017.02.006

Drozd, B., Couvilllon, E., & Suarez, A. (2018). Medical YouTube Videos and Methods of Evaluation: Literature Review. *JMIR Med Educ, 4*(1). doi:10.2196/mededu.8527

Timani, H., Shah, P., & Joshi, M. (2019). Predicting Success of a Movie from Youtube Trailer Comments using Sentiment Analysis. *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 584-586). Nueva Delhi: IEEE.

Alhujali, R. F., & Yafooz, W. M. (2021). Sentiment Analysis for Youtube Videos with User Comments: Review. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, (pp. 814-820). Coimbatore. doi:https://doi.org/10.1109/ICAIS50930.2021.9396049

Pandey, P. (2019, Agosto 6). *Sentiment Analysis is difficult, but AI may have an answer. | by Parul Pandey* . Retrieved from Towards Data Science: https://towardsdatascience.com/sentiment-analysis-is-difficult-but-ai-may-have-an-answer-a8c447110357

Bhuiyan, H., Ara, J., Bardhan, R., & Islam, M. R. (2017). Retrieving YouTube video by sentiment analysis on user comment. *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* (pp. 474-478). IEEE. doi:10.1109/ICSIPA.2017.8120658.

Welbourne, D. J., & Grant, W. J. (2016, Agosto 25). Science communication on YouTube: Factors that affect channel and video popularity. *Public Underst Sci.*, 706-718. doi:10.1177/0963662515572068

Asghar, M. Z., Ahmad, S., Marwat, A., & Kundi, F. M. (2015). Sentiment Analysis on YouTube: A Brief Survey. *MAGNT Research Report, 3*(1), 1250-1257. doi:10.48550/arXiv.1511.09142

Novendri, R., Callista, A. S., Pratama, D. N., & Puspita, C. E. (2020). Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes. *Bulletin of Computer Science and Electrical Engineering, 1*(1), 26-32. doi:10.25008/bcsee.v1i1.5

Chang, W.-L. (2018). Will Sentiments in Comments Influence Online Video Popularity? *2018 IEEE International Conference on Big Data (Big Data)* (pp. 3644-3646). Seattle, WA, USA: IEEE. doi:10.1109/BigData.2018.8621938

Chang, W.-L., Chen, L.-M., & Verkholantsev, A. (2019). Revisiting Online Video Popularity: A Sentimental Analysis. *Cybernetics and Systems, 50*(6), 563-577. doi:10.1080/01969722.2019.1646012

Fontanini, G., Bertini, M., & Del Bimbo, A. (2016). Web Video Popularity Prediction using Sentiment and Content Visual Features. *ICMR '16: Proceedings of the 2016 ACM on*

International Conference on Multimedia Retrieval, (pp. 289-292). doi:10.1145/2911996.2912053

Iwazaki, J., Aoki, K., Kato, C., & Kimura, K. (2021). COVID-19 Pandemic and the Verification of Effects of Yoga Intervention Using YouTube on Mental Health and Subjective Happiness of Workers. *In Psychology, 12*(12), 2083-2096. doi:https://doi.org/10.4236/psych.2021.1212126

Trougakos, J. P., Chawla, N., & McCarthy, J. M. (2020). Working in a Pandemic: Exploring the Impact of COVID-19 Health Anxiety on Work, Family, and Health Outcomes. *Journal of Applied Psychology, 105*(11), 1234-1245. doi:10.1037/apl0000739

YouTube. (2018, Diciembre 6). YouTube Rewind 2018: Everyone Controls Rewind | #YouTubeRewind. Retrieved from https://www.youtube.com/watch?v=YbJOTdZBX1g

Reynolds, C. (2020). IT'S REWIND TIME, EVERYBODY: THE CONTESTATION OF PLATFORM CULTURE IN YOUTUBE'S YEARLY REVIEW. *AoIR Selected Papers of Internet Research*. doi:10.5210/spir.v2020i0.11314

Blandford, A., Wesson, J., Amalberti, R., AlHazme, R., & Allwihan, R. (2020). Opportunities and challenges for telehealth within, and beyond, a pandemic. *The Lancet Global Health, 8*(11), e1364–e1365. doi:https://doi.org/10.1016/s2214-109x(20)30362-4

Jardines Méndez, J. B. (2007). Acceso a la información y equidad en salud. . *Revista Cubana de Salud Pública*.

Jensen, K. B. (1987). Qualitative audience research: Toward an integrative approach to reception, Critical Studies in Mass Communication.

Google. (2022). *Ayuda de Youtube*. Retrieved from Support Google: https://support.google.com/youtube/answer/9002587

Madden, A., Ruthven, I., & McMenemy, D. (2013). A classification scheme for content analyses of YouTube video comments. *Journal of Documentation, 69*(5), 693-714. doi:10.1108/jd-06-2012-0078

Google. (2021, Julio 2). *API Reference | YouTube API | Google Developers*. Retrieved from Sitio web de API de YouTube: https://developers.google.com/youtube/v3/docs

Sudarshan, R. (2021). *Getting Started with Google Bert.*

Jacob Devlin, M.-W. C. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. doi:10.48550/ARXIV.1810.04805

Mekouar, S., Zrira, N., & Bouyakhf, E.-H. (2017). Popularity Prediction of Videos in YouTube as Case Study: A Regression Analysis Study. *BDCA'17: Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*, (pp. 1-6). Nueva York.