Vórtice de Odio: Análisis de Comentarios en YouTube

Leonela Rivera Leinecker

CONSIDERACIONES INICIALES

Las relaciones sociales en las últimas décadas han ido ampliando sus modalidades en plataformas virtuales. Las ventajas que ofrecen el anonimato y la interacción a través de pantallas parecen favorecer la afloración de todo tipo de emociones intensas. En este proyecto nos vamos a centrar en aquellas interacciones de contenido negativo, principalmente en la plataforma de YouTube y en español.

> *Pregunta de Investigación*: ¿Los tópicos relacionados a la política generan gran contenido de comentarios de odio?
>
> *Hipótesis Principal*: Se relaciona la mayor cantidad de comentarios de odio con una línea política en particular.
>
> *Hipótesis Alternativa*: Los comentarios y/o publicaciones de contenido negativo generan mayor interacción que los positivos.

OBJETIVO GENERAL

- Identificar y clasificar el contenido de los comentarios de odio

OBJETIVOS ESPECÍFICOS

- Obtener publicaciones y comentarios asociados en relación a los diferentes tópicos planteados.
- Explorar el contenido de los comentarios.
- Identificar la frecuencia e intensidad de odio de los comentarios.
- Relacionar los comentarios con los usuarios.
- Predecir la ocurrencia por categoría de los comentarios negativos
- Compara la prevalencia y naturaleza de los comentarios de odio en diferentes plataformas sociales en relación con la política.

METODOLOGÍA

Mediante el Procesamiento de Lenguaje Natural (NLP) utilizando la librería Pysentimiento de Análisis de Sentimientos analizaremos la valoración del contenido de las publicaciones, esto devolverá una valoración de *negativo, neutro o positivo* de cada comentario. Centraremos el análisis en aquellos que sean de contenido negativo para analizar la intensidad de odio con la librería Speech Haters que devuelve un score de *hateful* y *aggressive*. A partir de los resultados anteriores, generamos una función de categorización para identificar la temática del contenido del mensaje. Así, y basados en la frecuencia de comentarios relacionados con la temática, se establecieron tres categorías: *ideología política, racismo* y *machismo.*
En base a los resultados de procesamiento de los comentarios de la etapa anterior hemos explorado diferentes modelos de Machine Learning para análisis multicategoría. Utilizando el modelo de *Support Vector Machine,* especialmente *SVC* buscaremos categorizar los comentarios, tal que genere una etiqueta para indicar el contenido del mensaje.
En un enfoque no conservador utilizaremos BERT (Bidirectional Encoder Representations from Transformers) de técnica basada en redes neuronales para el pre-entrenamiento del procesamiento del lenguaje natural (PLN), para la clasificación de los comentarios.

RECOLECCION Y ANALISIS DATOS

A través de la utilización de API, extraemos información de las páginas de YouTube. El criterio de búsqueda estará basado en ciertos tópicos de la agenda política escogidos por valoración interna, estos son:

- Inmigración (desagregar cada punto para mejorar el filtro de búsqueda)
- Economía
- Educación
- Salud
- Vivienda
- Género
- Violencia de Género
- Territorio
- Amnistía
- Feminismo
- Monarquía

A través de la utilización de librerías como Pandas, Numpy, Matplotlib, Seaborn, entre otras generamos bases datos para realizar análisis estadísticos básicos. Implementaremos diferentes técnicas de procesamientos de datos para la

normalización y tokenización de los comentarios para mejor análisis de texto. También elaboramos diversos gráficos para facilitar la visualización y exploración de los datos

CONCLUSIONES

Con todo lo expuesto, podemos considerar que nuestra pregunta de investigación *"¿los tópicos relacionados a la política generan gran contenido de comentarios de odio?",* puede responderse parcialmente, puesto que de la totalidad de comentarios (N= 110000) un 60% corresponden a comentarios negativos, y de ese grupo un 48% contiene mayores niveles de *hateful* y *aggressive*. En cuanto a nuestras hipótesis planteadas, no podemos aseverar que se produzcan más interacciones dependiendo del tópico de que se trate el video, ni resulta fácil identificar a qué línea política pertenecen los usuarios que emiten esos comentarios; pero sí se puede indicar que de las tres categorías utilizadas para clasificar los comentarios, aquellos que pertenecen a la categoría *ideología política* son los que mayor frecuencia presentan al tiempo que son los comentarios de mayor valores de *hateful* y *aggressive.*
En referencia al objetivo general de este proyecto (*Identificar y clasificar el contenido de los comentarios de odio*), siguiendo un enfoque más conservador hemos implementado el modelo de *SVC,* que nos ha arrojado una precisión de 0.67 (accuracy) a la vez que valores de precisión en torno a los 0.90 y un valor de recall mayor a 0.85 (Tabla 1).

Tabla 1: predicciones en el conjunto de prueba SVC

	precisión	recall	f1-score	support
ideología política	0.88	0.89	0.89	65
machismo	0.90	0.97	0.93	71
racismo	0.95	0.85	0.90	66
accuracy			0.91	202
macro avg	0.91	0.90	0.90	202
weighted avg	0.91	0.91	0.91	202

Lo anterior nos indica que si bien el valor de predicción no es elevado, al considerar los valores de *presicion* y *recall* demuestran que el modelo se ajusta bien a las

categorías.
Partiendo de un enfoque menos conservador, hemos implementado el modelo pre entrenado de BERT para predecir las mismas categorías. Por razones de coste computacional hemos definido *epochs = 3,* y *batch_size=32,* lo que nos devolvió valores de precisión de 0.63 (accuracy), *precision* en torno a los 0.60 y *recall* cercanos a .70 (tabla 2).

Tabla 1: predicciones en el conjunto de prueba BERT

	precisión	recall	f1-score	support
ideología política	0.65	0.42	0.51	74
machismo	0.59	0.74	0.66	57
racismo	0.66	0.77	0.71	71
accuracy			0.63	202
macro avg	0.63	0.64	0.63	202
weighted avg	0.64	0.63	0.62	202

En observancia con estos datos, podemos considerar que para esta problemática es mejor utilizar un enfoque convencional por su beneficio en tiempo y coste computacional.

CONSIDERACIONES FINALES Y OPORTUNIDADES DE MEJORA.

Con todo lo anterior, podemos pensar que a la hora de analizar el impacto de los comentarios negativos, agresivos y de odio en las redes sociales, identificar y/o etiquetar el contenido del comentario favorece el entendimiento del mismo. A su vez, puede pensarse como una herramienta de higiene en el tipo de consumo de esas redes sociales, si se advierte al usuario el posible contenido de esos comentarios.
Como posibilidades de mejora entendemos que mejorar la calidad del dataset para el entrenamiento de los modelos permitirá ajustar los parámetros y a su vez las métricas. Otras opciones de mejora en torno a los datos sería Investigar si las medidas tomadas por las plataformas (eliminación de contenido, suspensión de cuentas, etc.) tienen un impacto real en la reducción del odio; geolocalización de los comentarios; identificar usuarios con mayor actividad y comparar sus interacciones con diferentes tópicos; comparar temáticas similares en diferentes plataformas virtuales.