UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

Maestría en Ciencia de Datos

Universidad Autónoma de Nuevo León

Tarea 1 - Análisis estadístico de comentarios del primer debate presidencial

Por

Juan Jesús Torres Solano

Profesor: José Alberto Benavides Vázquez

Índice general

1. INTRODUCCIÓN

En el presente trabajo se realiza el análisis estadístico de texto, empleando como fuente a los comentarios de la transmisión en vivo del primer debate presidencial, realizado el día 7 de abril del 2024; los comentarios presentan una estructura lingüística mal estructurada, de igual forma, se identifican Emojis entre apalabras o comentarios completos a partir de los mismos.

1.1. PRIMER OBJETIVO

Generar información estadística que nos permita visualizar los elementos característicos, que forman parte de los comentarios, como son, la frecuencia palabras, votos a los comentarios (Top 10 de votos), frecuencias de participación en los comentarios (Top 10 de Usuarios) y la relación entre los votos y las replicas.

1.2. SEGUNDO OBJETIVO

Realizar una comparativa entre el numero de menciones y la estadística de intención de voto previo a este debate, a modo de comparación estadística entre la expresión del nombre de los candidatos en los comentarios y la intención de voto documentada hasta ese momento.

2. DESCRIPCIÓN DE LOS DATOS

Del total de datos unicamente consideraremos el siguiente conjunto de columnas:

- ID: identificador de comentario
- text : Comentario del Usuario.
- author : Nombre de usuario que realiza el comentario.
- votes : numero de votos recibidos
- replies : número de replicas en los comentarios.

Se identifican un total de 4307 comentarios publicados en el video de Youtube (URL: https://www.youtube.com/watch?v=kZaucITWv00t=18s), de los cuales 2352 representan algun comentario sin estructura textual o verbal, correspondiendo a signos puntuales o emojis.

En total tenemos un conjunto de 1955 participaciones con algún tipo de comentario textuales ademas de Emojis, en el cuadro 2.1 podemos observar una muestra representativa del conjunto de datos.

ID	text	author	votes	replies
0	El reloj debió verse en todo momento, para ten...	@celinaramirez6363	394	16.0
1	Me uno a exigir que pongan los relojes de los ...	@puellacodicum8569	555	19.0
2	Me parece muy bien cómo contestó Maynes, fue d...	@alondrareal8518	79	NaN
...	...	...	...	...
1953	Maynes, palero de Claudia #CrimesAgainstHumani...	@yamerosolitario2	0	NaN
1954	Xóchitl Presidente	@marcoantonioarroyo952	0	NaN
1955	Cuánto dinero les pagaste por esos reconocimie...	@maru641	0	NaN

Cuadro 2.1

A partir de los datos podemos generar algunos elementos estadísticos interesantes del conjunto, por ejemplo el top 10 de mensajes por usuario, lo que nos indica la participación dentro de los comentarios, identificando los usuarios mas activos e identificar posteriormente su inclinación política con respecto el debate (ver figura 2.1).

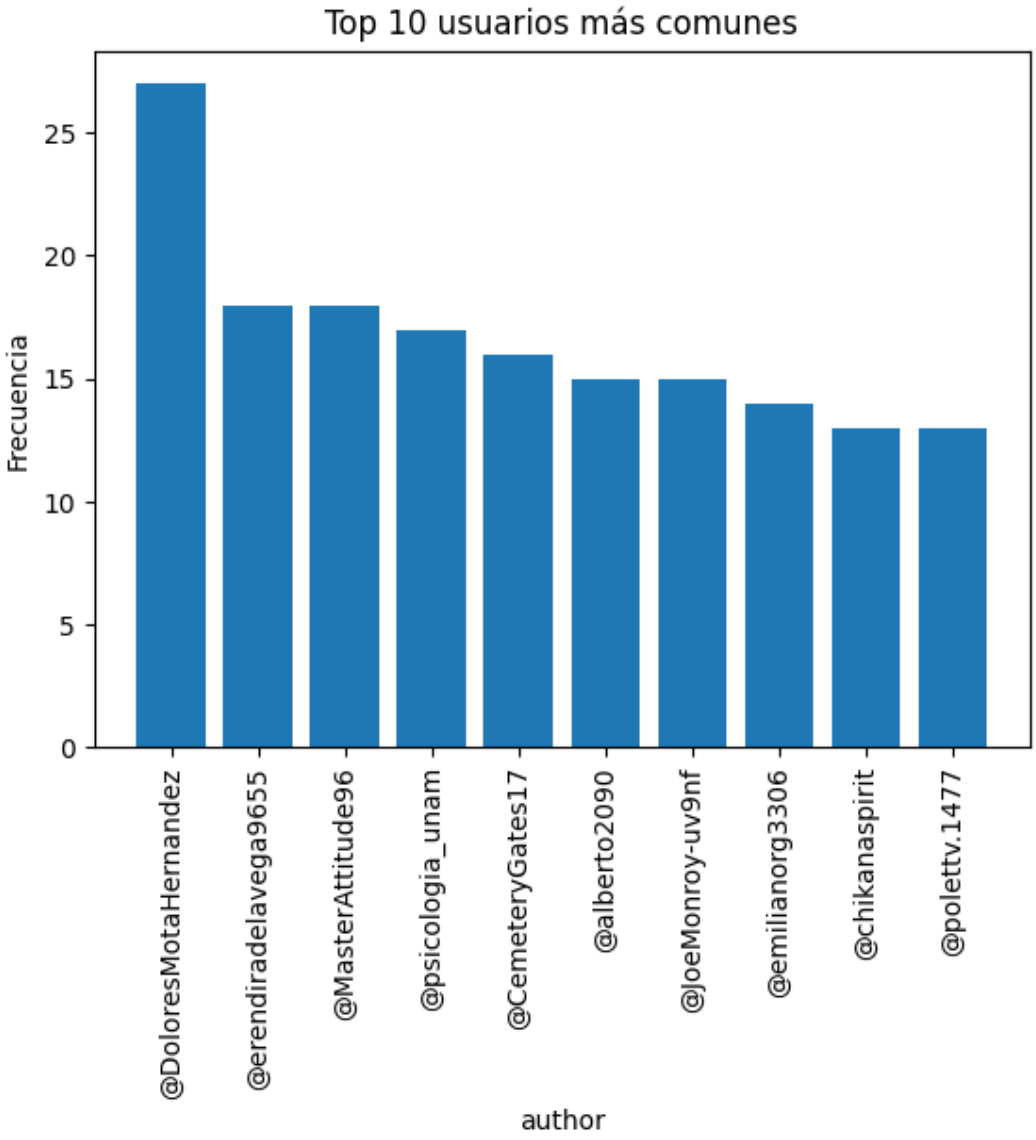

Figura 2.1: Top 10 de usuarios activos con mas comentarios durante el debate.

En cuanto a la frecuencia de numero de votos, que un porcenpodemos observar que un porcentaje alto de los comentarios no presentan votos, mientras que un porcentaje alto de numero de votos entre 1 y 22 acumulan un alto porcentaje, pero no el mayor, en la figura 2.4 podemos identificar comentarios individuales con un alto numero de votos, los cuales son poco frecuentes pero muy significativos.

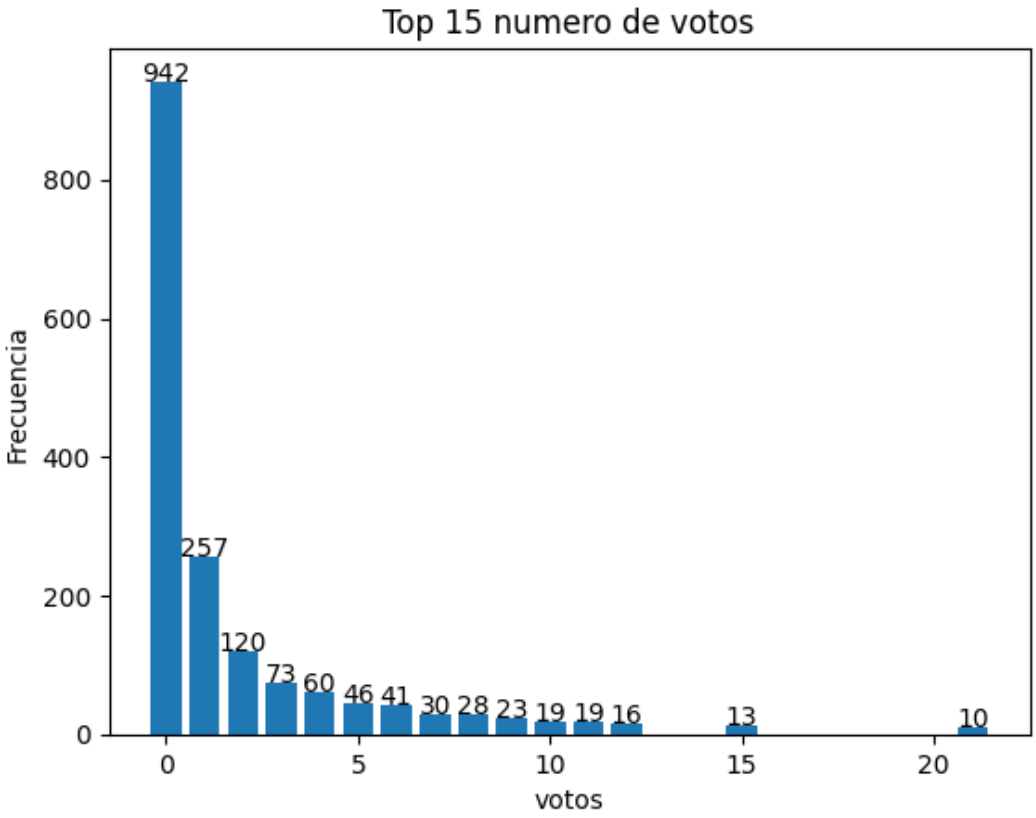

Figura 2.2: Top 10 del mayor numero frecuencia de votos en los comentarios.

En el comportamiento de las replicas podemos identificar una mayor frecuencia en un menor numero de replicas por comentario, es decir la gran mayoria de pocas replicas ocurren mas frecuentemente, al igual que la grafica anterior, podemos observar un comportamiento mas general en la figura 2.4, donde se identifican la poca frecuencia de replicas continuas en en comentarios, es decir solo algunos comentarios son significativos y acumulan un alto numero de replicas.

Al observar la distribución de votos Vs replicas de los comentarios, se puede identificar una relacion entre los comentarios con mayor numero de votos y los comentarios con mayor numero de replicas, correspondiendo entre si (ver figura 2.4).

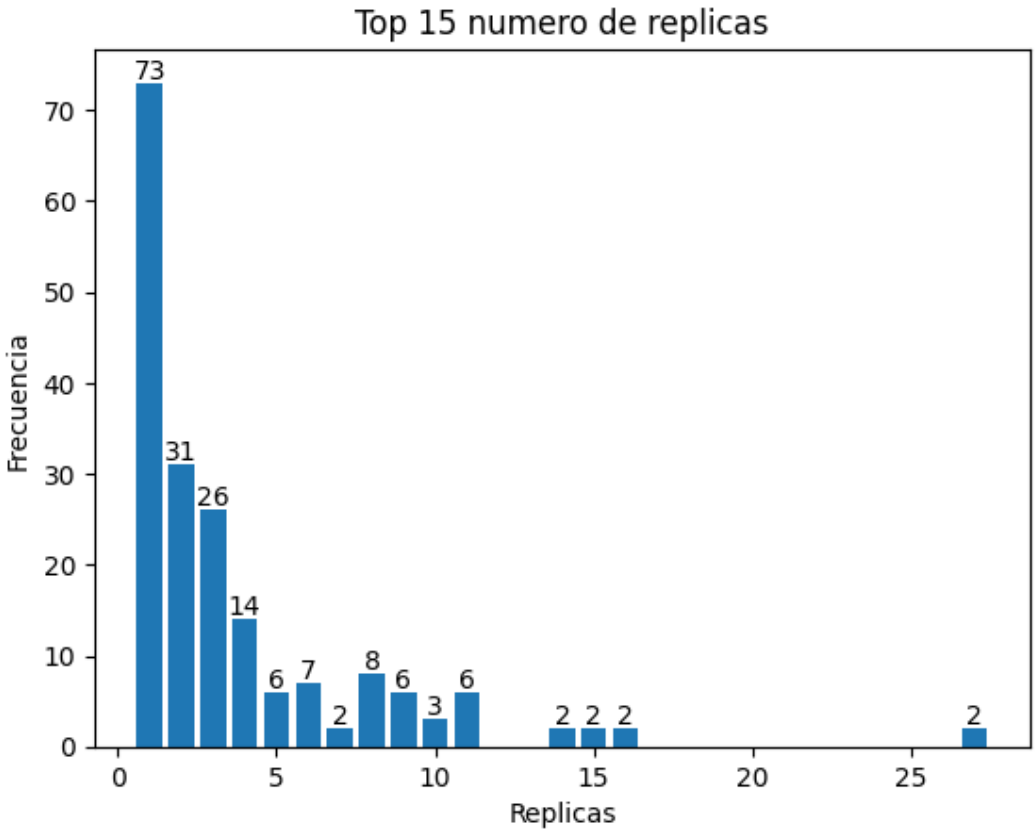

Figura 2.3: Top 10 del mayor numero de frecuencias de replicas en los comentarios.

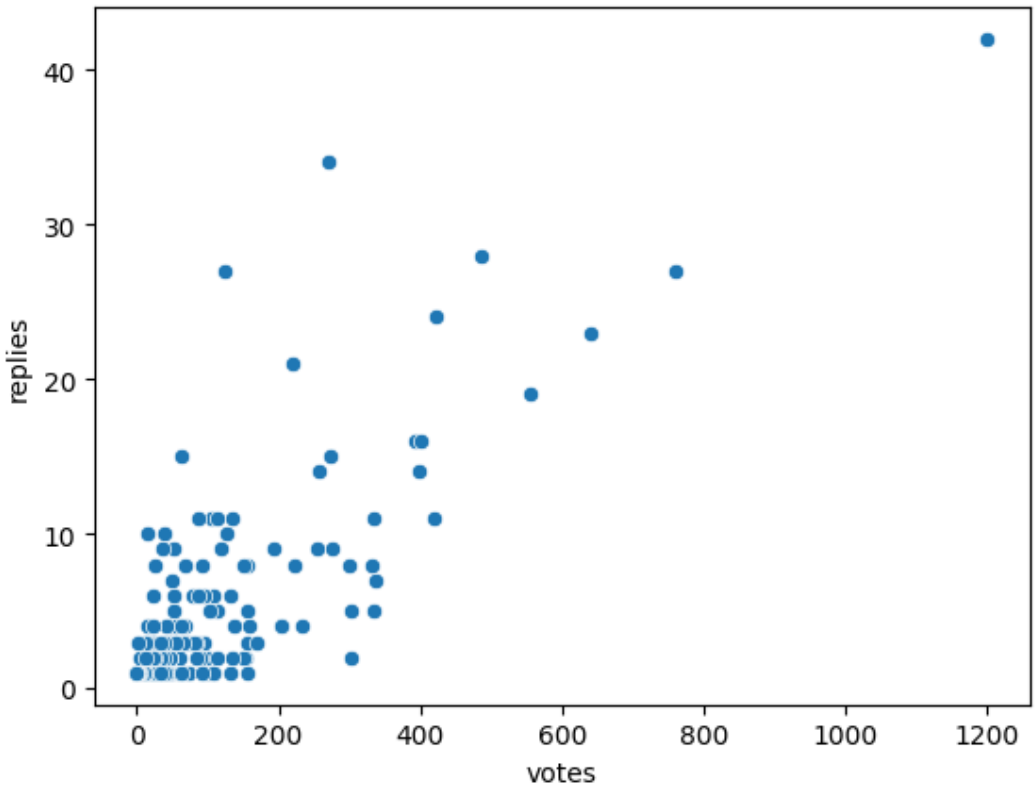

Figura 2.4: grafico de distribución de replicas contra votos

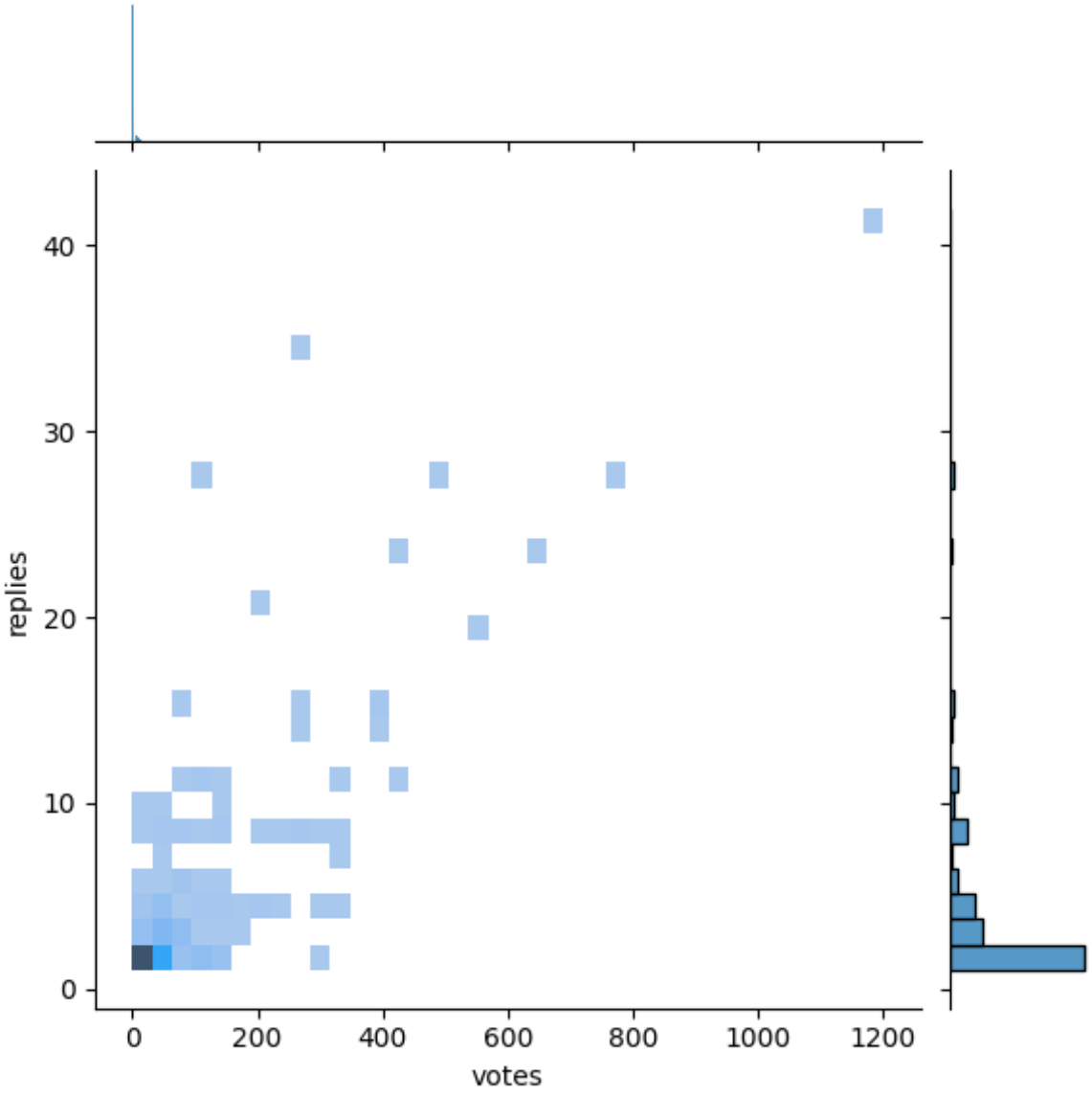

Figura 2.5: grafico de frecuencias y distribución

3. PROCESADO DE LOS COMENTARIOS

4. CONCLUSIONES

Según las encuestas y análisis publicados antes del primer debate presidencial en México 2024, las intenciones de voto se distribuían de la siguiente manera:

Claudia Sheinbaum (Morena, PT y PVEM): 49Xóchitl Gálvez (PRI-PAN-PRD): 26Jorge Álvarez Máynez (Movimiento Ciudadano): 18

La encuesta de El Financiero del 1 de abril de 2024 mostraba que Sheinbaum lideraba con un 35

La encuesta de FactoMétrica y Reporte Índigo publicada el 9 de abril de 2024 mostraba que Sheinbaum lideraba con un 69

La encuesta “flash” publicada el 7 de abril de 2024 mostraba que Sheinbaum lideraba con un 49

BIBLIOGRAFÍA

[1] A. Savitzky, M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures., Analytical Chemistry 36 (8) (1964) 1627–1639. doi:10.1021/ac60214a047.

[2] C. J. Harvey, J. M. Pilcher, R. J. Eckersley, M. J. Blomley, D. O. Cosgrove, Advances in ultrasound, Clinical Radiology 57 (3) (2002) 157–177. doi:10.1053/crad.2001.0918.