**PREDICIENDO TWEETS**

Problem Set 4 – GRUPO 6

Víctor Dulio Chique

Víctor Iván Sánchez

Natalia Castro Alarcon

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1. INTRODUCCIÓN:**

Las opiniones de personalidades políticas, empresariales o que tengan algún tipo de influencia importan. En EEUU por ejemplo, algunos estudios muestran que algunque los mercados financieros tienden a reaccionar sobre los tweets de personalidades políticas en particular de Donald Trump. [[1]](#footnote-1) Principalmente los políticos usan Tweeter para defender sus posiciones y precisar sus ideologías. Particularmente aquellos que tienen posiciones contundentes, pueden producir empatía o reacciones adversas de contradictores y por tanto sus posiciones y comentarios pueden identificado, por los patrones inmersos en su lenguaje. Las personas pueden predecir cuál va a ser su punto de vista o no les sorprende que cierta frase pertenezca a un personaje en específico. En ese sentido ¿Pueden los algoritmos de aprendizaje de máquinas predecir correctamente a quién pertenecen los *tweets* que se publican?

En este trabajo buscamos responder a esta pregunta a través de modelos de clasificación que utilizaron los tweets de tres políticos colombianos muy reconocidos: Claudia López, Álvaro Uribe y Gustavo Petro. El objetivo se centró en predecir acertadamente a quién corresponde cada *tweet* identificando el conjunto de palabras que usualmente hacen parte de sus mensajes y buscando el modelo que arrojara la mejor medida de “*accuracy”*, es decir el mayor número de predicciones acertadas sobre el total de predicciones.

**2. DATOS**

**2.1 Descripción de los Datos**

Tenemos una base de datos con 9.347 Tweets de 3 de los políticos mas prominintes de Colombia, Gustavo Petro, Alvaro Uribe y Claudia Lopez, la base solo cuenta con un identificador del tweet, el nombre del usuario que lo realizo y el texto del tweet, de igual manera contamos con una base de datos de test con 500 tweets de cada usuario, esta base es la que debemos predecir.

**2.1.1 Limpieza de datos**

Dado que tenemos una base de datos que contiene texto crudo, en este caso tweets de políticos, usamos Procesamiento de Lenguaje Natural (NPL), el cual resulta útil para tareas especificas como la estandarización, eliminación de stopwrods o palabras vacías, la tokenizacion y en particular, eliminar del texto todo aquello que no aporte información relevante para la predicción. En las redes sociales, por ejemplo, los usuarios escriben de la forma que quieren, lo que podría resultar en un elevado uso de abreviaturas, signos de puntuación, palabras redundantes, entre otros. Así mismo usamos la técnica de tokenizar los textos, que consiste en dividir el texto en las unidades que lo conforman. En general, el proceso de limpiar y tokenizar el texto consiste, principalmente en convertir el texto en minúscula, eliminar los signos de puntuación, eliminar los números, eliminar espacios en blanco y tokenizar por palabras individuales.

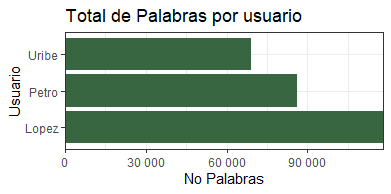
**2.2 Descripción de los datos**

En esta sección se realiza un análisis exploratorio de los datos. La gráfica 1 muestra que Claudia López fue quien más palabras utilizó en los tweets. En el cuadro 1 se puede apreciar que Álvaro Uribe utilizó 41.4% menos palabras que ella y Gustavo Petro 30% menos. De igual manera, Claudia López escribió la mayor cantidad de palabras diferentes pero la diferencia frente a Gustavo Petro y Álvaro Uribe fue solamente de 6.36% y 15% menos respectivamente. En general se pude inferir que Claudia López habla sobre una mayor variedad de temas y/o utiliza mayor número de palabras para expresar una idea o informar sobre un acontecimiento. Álvaro Uribe pareciera centrase en ciertos contenidos y los escribe de forma más concisa y Gustavo Petro se encuentra en un punto medio.

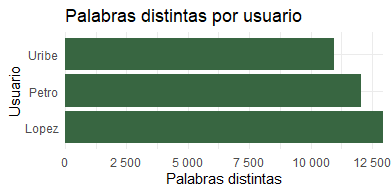
|  |  |  |
| --- | --- | --- |
| Usuario | Total palabras utilizadas en tweets | Palabras diferentes |
| Claudia López | 118.104 | 12876 |
| Gustavo Petro | 86.218 | 12056 |
| Álvaro Uribe | 69.175 | 10942 |

Cuadro 1: Total palabras utilizadas y palabras diferentes

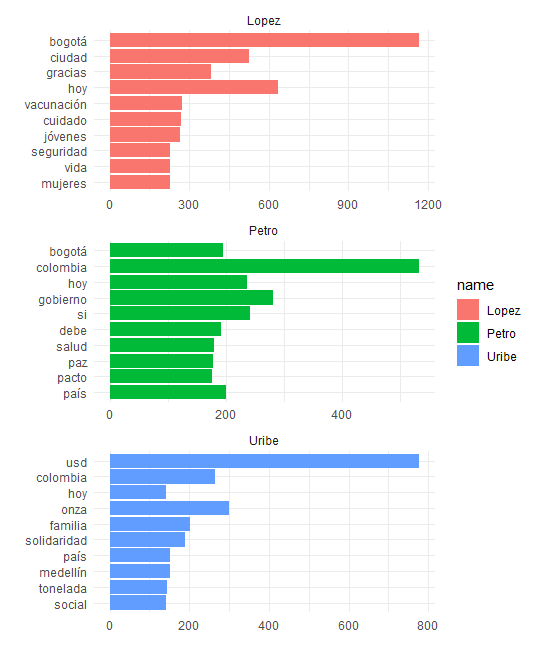
El cuadro 5 muestra cuáles son las palabras que con mayor frecuencia utilizan López, Uribe y Petro. La palabra más utilizada por Claudia López es “Bogotá”, la de Gustavo Petro es “Colombia” y la de Álvaro Uribe “USD”. En general López se centra en lo relacionado con la ciudad, las mujeres, el reconocimiento a quienes ayudan a logra un objetivo, la vacunación (los tweets se tomaron durante el periodo final de la pandemia), los jóvenes y la seguridad. Gustavo Petro se refiere a temas que tienen que ver con el país, el manejo del gobierno, la salud, la paz y hace frecuentemente mención de su movimiento: Pacto Histórico. Finalmente, Álvaro Uribe escribe sobre el dólar lo cual pareciera mostrar especial preocupación por el tema de la devaluación del peso colombiano, menciona frecuentemente a Colombia y la palabra país, y opina sobre la familia, la solidaridad y la ciudad de Medellín.



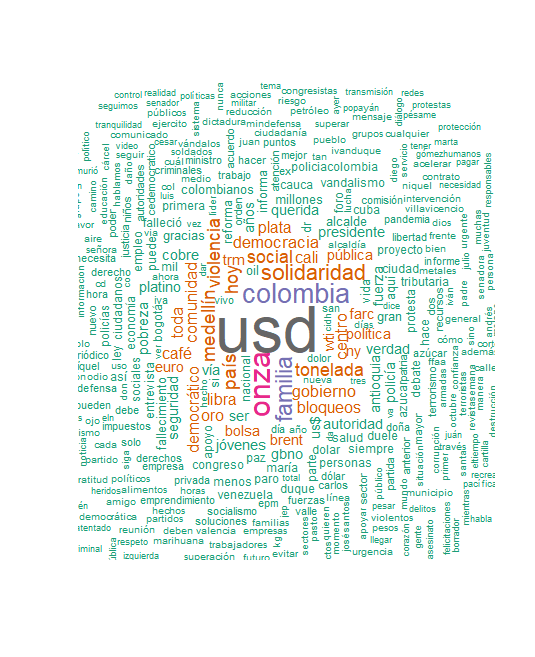
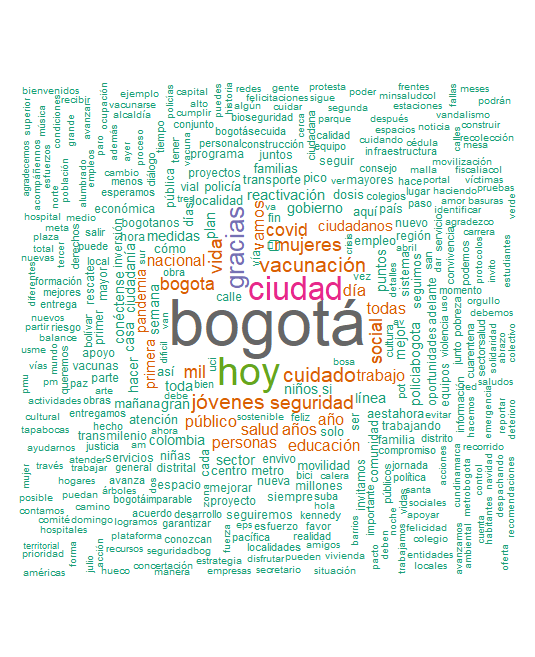
*Gráfico 1*

**

*Gráfico 2*



Cuadro 5: Palabras más frecuentes después de Stopwords



*Imagen 1. Wordcloud* ***Claudia Lopez*** *Imagen 2. Wordcloud* ***Alvaro Uribe***

**3. Modelos y Resultados**

**3. Conclusiones y Recomendaciones**

**4. Bibliografía**

Anexos

**Modelo Lasso con variable precio en logaritmos**

train <- Bogota\_train %>%

select(lprice, property\_type, rooms\_tot, bathrooms, distancia\_avenida\_principal,

distancia\_comercial, distancia\_parque, distancia\_universidad, rooms\_bath,

dist\_av2, dist\_com2, dist\_parq2, dist\_univ2)

Lambda optimo: 9.69804377996242e-05"

"Error (mse) de entrenamiento: 0.0874021221908704"Chart

Description automatically generated

"Error (mae) de entrenamiento: 0.239060247990877"

Chart, histogram

Description automatically generated

**Modelo Lasso con polinomios**

OLS RIDGE LASSO ELASTIC\_NET

(Intercept) 591372140.6 591372140.6 591372140.59 591372140.6

property\_typeCasa -30761762.9 -27671041.9 -30468958.28 -30353812.9

`poly(Superficie, 2)1` 91779606.0 59134294.2 89308622.98 88606128.6

`poly(Superficie, 2)2` -85156464.6 -47340141.7 -78480095.63 -76386494.9

rooms\_tot -29503545.3 -17639518.3 -28165170.91 -27587415.6

bathrooms 54106319.6 49748845.1 53703186.13 53654590.6

`poly(Superficie, 2)1:rooms\_tot` 50964122.5 50675645.2 52044514.67 52994058.6

`poly(Superficie, 2)2:rooms\_tot` 19389049.8 2427501.8 18100666.15 16456599.7

`poly(Superficie, 2)1:bathrooms` 10180131.7 36947714.1 11116079.36 10612037.0

`poly(Superficie, 2)2:bathrooms` 5628467.7 -13004017.4 373812.81 0.0

`bathrooms:poly(distancia\_parque, 2)1` 143694.6 615552.4 11128.64 0.0

`bathrooms:poly(distancia\_parque, 2)2` -1261591.7 -1399073.8 -989937.24 -873658.9

`bathrooms:poly(distancia\_comercial, 2)1` -7522750.9 -7529736.1 -7337572.18 -7237881.6

`bathrooms:poly(distancia\_comercial, 2)2` -8690855.3 -8449249.3 -8515198.76 -8446796.2

`bathrooms:poly(distancia\_avenida\_principal, 2)1` 9010529.7 8554239.6 8684796.39 8557105.4

`bathrooms:poly(distancia\_avenida\_principal, 2)2` -2948503.2 -3008093.2 -2752174.13 -2678022.8

`bathrooms:poly(distancia\_universidad, 2)1` -1167006.3 -1371331.4 -942099.92 -846661.4

`bathrooms:poly(distancia\_universidad, 2)2` -4901361.6 -5165554.3 -4726439.11 -4658935.7

OLS RIDGE LASSO EN

RMSE 151007264 151577050 150950062 151591374

Chart, histogram

Description automatically generated

1. Llorente, Juan. (2020). El impacto de los tweets de Donald Trump sobre los mercados financieros. [↑](#footnote-ref-1)