Grupo 6. Problem Set 4



**Natalia Castro Alarcon.**

**Victor Dulio Chique. 200319157**

**Victor Ivan Sanchez. 201510287**

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

1. INTRODUCCCIÓN

Los *tweets* no solo son una herramienta importante para opinar, informar e influir sobre la opinión pública sino que permiten conocer los perfiles de las personas que los publican. Quienes hacen parte del gobierno, la oposición o se encuentran haciendo algún tipo de campaña, utilizan Twitter para defender sus posiciones y precisar sus ideologías. Particularmente aquellos que tienen posiciones contundentes generan no sólo empatía entre quienes concuerda con ellos sino que comienzan a ser identificados con cierta facilidad. 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.

¿Pueden los algoritmos de aprendizaje de máquinas predecir correctamente a quién pertenecen los *tweets* que se publican? En este trabajo se buscó 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.

FALTA BREVE SINTESIS DE LOS MODELOS QUE SE UTILIZARON Y RESULTADOS.

1. 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.

|  |  |  |
| --- | --- | --- |
|  | Total palabras utilizadas en tweets | Palabras diferentes |
| Claudia López | 118104 | 12876 |
| Gustavo Petro | 86218 | 12056 |
| Álvaro Uribe | 69175 | 10942 |

Cuadro 1: total palabras utilizadas y palabras diferentes

|  |  |
| --- | --- |
| Chart, bar chart, treemap chart  Description automatically generated  *Gráfica 1: Total de palabras* | Chart  Description automatically generated  *Gráfica 2: Palabras diferentes* |

Train



Test



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.

Graphical user interface

Description automatically generated with medium confidence

Cuadro 5: Palabras más frecuentes.

Graphical user interface

Description automatically generated with medium confidence