**Taller 4.**

**Predicción de Tweets**

**Integrantes:** Isabella Mendez Pedraza. Cód.: 201814239

Manuela Ojeda Ojeda. Cód.: 201814476

Juan Sebastian Tellez Melo. Cód.: 201513710

Andres Mauricio Palacio Lugo. Cód.: 201618843

**Link del repositorio:** <https://github.com/AndresMPL/Repositorio_PS4.git>

1. **Introducción.**

La política colombiana actual se ha representado por tener líderes en el Gobierno nacional con opiniones bastante diferentes entre sí, principalmente en temas importantes para el país como la educación, la salud, la política monetaria, entre otros.

Recientemente, tres líderes políticos se han destacado por tener opiniones muy controversiales: Álvaro Uribe, quien fue presidente por 8 años (2002 a 2010), pertenece al partido Centro Democrático y tiene una ideología conservadora. Claudia López, quien es la actual alcaldesa de Bogotá, pertenece al partido Alianza Verde y con ideología de centroizquierda. Y, por último, Gustavo Petro, quien es el actual presidente, pertenece al partido Colombia Humana y su ideología es más de izquierda.

Con el auge de la digitalización, se ha vuelto común que las personas usen las redes sociales para comunicar sus opiniones acerca de cualquier tema, así como para educarse o informarse de temas del día a día. Los lideres políticos se han sumergido en esta cultura digital y suelen dar sus opiniones, comentar y debatir acerca de los temas de coyuntura actual del país por medio redes sociales, principalmente por Twitter.

Dada la gran cantidad de información que se maneja en la web los tweets se han convertido en una fuente de información de gran interés debido a que permite detectar tendencias de opinión de los usuarios. Por medio del análisis de contenido es posible identificar patrones de comportamiento entre los usuarios y puntos de inflexión en las corrientes de opinión (Baviera,2017).

Para el ejercicio que se presentará en este documento, contamos con una base de datos que contiene tweets de Álvaro Uribe, Claudia López y Gustavo Petro y vamos a abordar un problema de clasificación para lograr predecir de que político colombiano es el tweet.

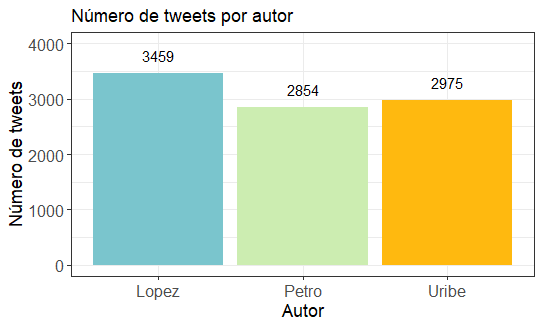
Para este ejercicio se usó el texto de cada tweet como información para generar variables, a partir de la cuál construimos una red neuronal y generamos un modelo para la predicción esperada. El modelo con el mejor puntaje de predicción utilizó una capa de entrada con 512 neuronas, dos capas ocultas y una capa de salida con 4 nodos.

1. **Datos.**

Los datos utilizados para este ejercicio se obtuvieron de dos bases de datos, Train y Test, que contenían 9.349 y 1.500 registros, respectivamente, correspondientes a tweets escritos por Claudia López, Álvaro Uribe y Gustavo Petro. Cada base de datos contenía el texto completo del tweet; en caso de Train se tenía el autor del tweet, por el contrario, en la base de datos de Test no se tenía el autor, por lo tanto, el ejercicio consistió en generar un modelo que lograra predecir qué político colombiano había escrito cada tweet de Test.

Los datos de entrenamiento contenían 3.459 (37%) tweets escritos por Claudia López, 2.854 (31%) escritos por Gustavo Petro y 2.975 (32%) escritos por Álvaro Uribe, tal como se muestra en la gráfica 1.

**Gráfica 1. Número de tweets por autor**

****

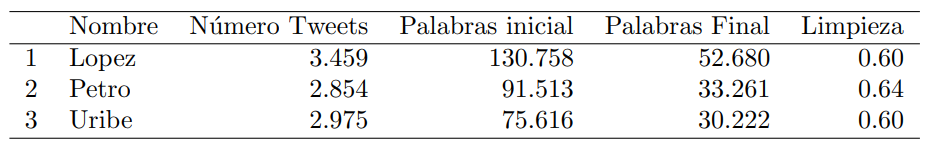
Las dos bases de datos, Train y Test, requirieron un proceso de limpieza del texto que consistió en la eliminación de acentos, la conversión del texto a minúscula, la eliminación de símbolos, signos de puntuación, emojis, direcciones de internet y espacios sobrantes.

Posteriormente, se realiza la separación del texto en palabras individuales, creando una nueva fila para cada palabra (tokenización) y se procede a realizar la eliminación de palabras vacías (*stopwords*).

Se realiza la lematización (*stemming*) del texto con el fin de extraer las variantes morfológicas de las palabras y convertirlas a raíces comunes, lo que se describe en el script denominado “*1\_Cleaning*” del repositorio señalado al inicio de este documento.

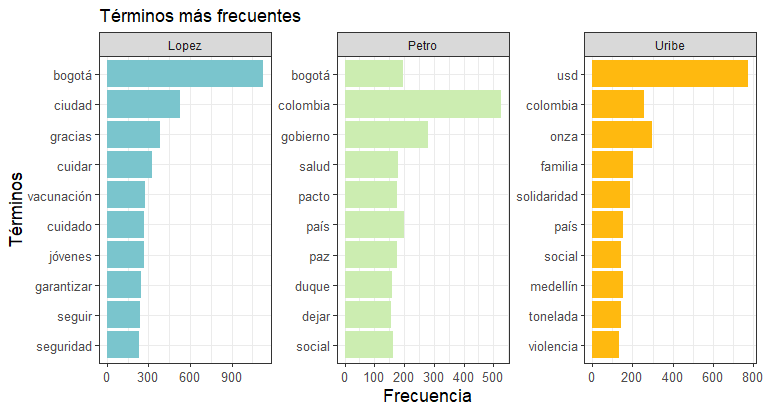
Para mostrar el impacto del proceso de limpieza, se tiene que en la base de datos Train se produjo una disminución de las palabras del 61%, tal como se muestra en la tabla 1, lo que significó que se pudieran agrupar temas comunes y eliminar palabras que generarían ruido para el modelo.

**Tabla 1. Resultados del proceso de limpieza**



Así, al terminar el proceso de limpieza de los textos, se identificaron los temas más frecuentes por cada autor, en el caso de la BD Train; se observa que Claudia López escribió principalmente sobre Bogotá, Gustavo Petro sobre Colombia y Álvaro Uribe acerca del dólar, como se observa en el gráfico 2. Adicionalmente, se evidencia que Claudia López toca temas relacionados con el cuidado, vacunación, jóvenes y seguridad principalmente; Gustavo Petro se refiere a temas relacionados con el gobierno, la salud, la paz, Duque y social; y Álvaro Uribe habla principalmente de familia, solidaridad, país, social, Medellín y violencia.

**Gráfico 2. Términos más frecuentes por autor**

****

Posteriormente, haciendo uso de las herramientas de R se generaron nubes de palabras, en total y por cada autor de los tweets, en las cuales se ven en orden los principales temas mencionados en los tweets, mostrados en las imágenes 1 a la 4. Este proceso se realizó mediante el script “*2\_Cleaning*”.

|  |  |
| --- | --- |
| **Imagen 1. Nube de palabras de la base de datos Train - Total** | **Imagen 2. Nube de palabras base de datos Train – Claudia López** |
|  |  |

|  |  |
| --- | --- |
| **Imagen 3. Nube de palabras base de datos Train – Álvaro Uribe** | **Imagen 4. Nube de palabras base de datos Train – Gustavo Petro** |
|  |  |

1. **Modelo y resultados.**

Utilizamos el texto de cada tweet como información y construimos una red neuronal que nos permitirá realizar la predicción esperada.

El modelo tiene combinaciones lineales de entradas que pasan a través de funciones de activación no lineales llamadas nodos. Para poder superar algunas de las limitaciones de los modelos lineales y manejar una clase más general de funciones se utilizan las capas ocultas.

Una red neuronal toma un vector de entrada de p variables

Y construye una función no lineal para predecir la respuesta , en este caso a quien político colombiano pertenece el tweet.

El modelo de la red neuronal tiene la forma:

Número de modelos cargados en kaggle – Descripción de los parámetros que se cambiaron

Métricas del mejor modelo

Número de capas

Número de parámetros

Velocidad de aprendizaje

Optimizador utilizado

Función de pérdida

Épocas de entrenamiento del modelo

Gráfica de la historia de entrenamiento del modelo

Valor del accuracy

Puntaje en Kaggle

1. **Conclusiones.**
2. **Bibliografía.**

Baviera, T. (2017). Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. Dígitos. 1(3):33-50. https://riunet.upv.es/handle/10251/153230

1. **Anexos.**