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), y en este sentido se encuentra, por ejemplo, el análisis hecho por Hakan et al. (2022), acerca de las opiniones de los miembros del Congreso de los Estados Unidos sobre Turquía con el fin de investigar si sus percepciones en Twitter reflejaban algunas de las tendencias en las relaciones entre Estados Unidos y Turquía y para ello utilizaron diferentes metodologías estadísticas, análisis de texto computacional y herramientas de modelado de

temas, a partir de los datos extraídos de Twitter y llegaron a concluir que, efectivamente, los datos de esta red social fueron útiles para acercarse a la percepción de Turquía entre los miembros del Congreso de los Estados Unidos y en tal sentido señalaron que "*(...) el uso de datos de redes sociales para analizar la política exterior y las relaciones internacionales es una iniciativa novedosa y valiosa*", lo que permite ver el potencial del análisis de estos datos a partir de herramientas como Big Data y el Machine Learning para diferentes situaciones como el análisis de percepciones, el relacionamiento de temas a diferentes sectores, líderes o políticos o la predicción de tendencias, entre otras.

Teniendo en cuenta lo anterior, el ejercicio que se presentará en este documento consiste en generar un modelo que permita predecir el autor de un tweet específico y para ello contamos con una base de datos que contiene tweets de Álvaro Uribe, Claudia López y Gustavo Petro y entre estos tres autores se realizará la clasificación de los tweets.

En este ejercicio se usó el texto de cada tweet como información para generar variables, a partir de las cuales 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.

2. Datos.

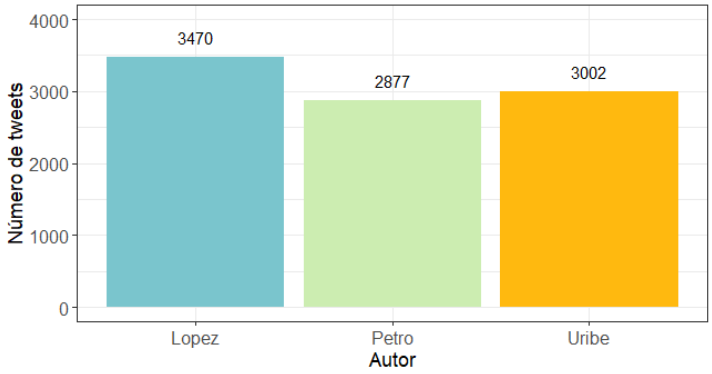
Los datos utilizados para este ejercicio se obtuvieron de dos bases de datos, Train y Test, que contenían 9.349 y 1.500 registros, respectivamente, correspondientes a tweets escritos por Claudia López, Álvaro Uribe y Gustavo Petro. Cada base de datos contenía el texto completo del tweet; en caso de Train 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 3.470 (37%) tweets escritos por Claudia López, 2.877 (31%) escritos por Gustavo Petro y 3.002 (32%) escritos por Álvaro Uribe, tal como se muestra en la gráfica 1.

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), se procedió a realizar la eliminación de palabras vacías (*stopwords*) y de las palabras que tenían una frecuencia inferior a 10 en toda la base de datos.

Gráfica 1. Número de tweets por autor



Finalmente, se realizó 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.

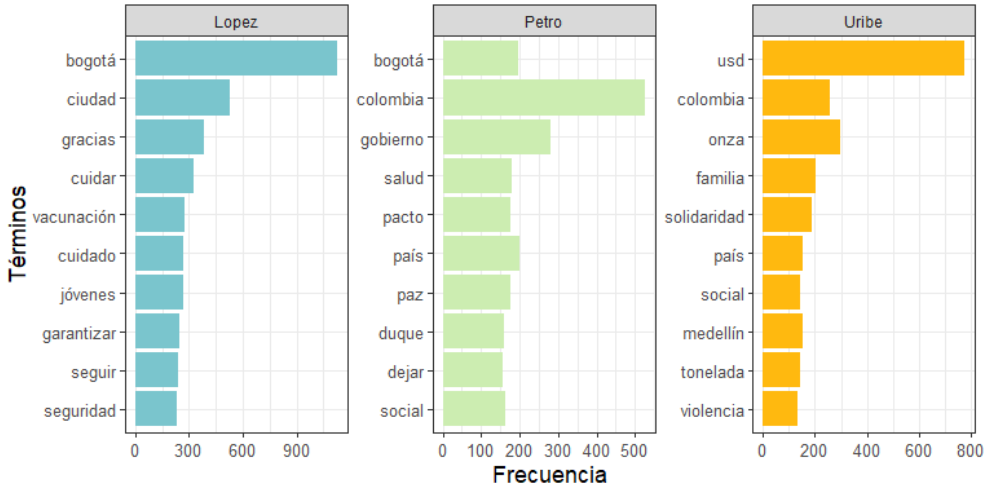
Para mostrar el impacto del proceso de limpieza, se tiene que en la base de datos Train se produjo una disminución de las palabras del 50%, tal como se muestra en la tabla 1, lo que significó que se pudieran agrupar temas comunes y eliminar palabras que generarían ruido para el modelo, sin perjuicio de que los errores en la escritura de algunas palabras causaran términos diferentes o que la lematización no identificara por completo las raíces comunes de las palabras.

Tabla 1. Resultados del proceso de limpieza

	Nombre	N Tweets Inicial	N Tweets Final	N Palabras Inicial	N Palabras Final	Limpieza
1	Lopez	3.470	3.459	130.758	52.680	0.60
2	Petro	2.877	2.854	91.513	33.261	0.64
3	Uribe	3.002	2.975	75.616	30.222	0.60

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 lo social; y Álvaro Uribe habla principalmente de familia, solidaridad, país, Medellín y la 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 denominado "2_*Cleaning*".

3. Modelo y resultados.

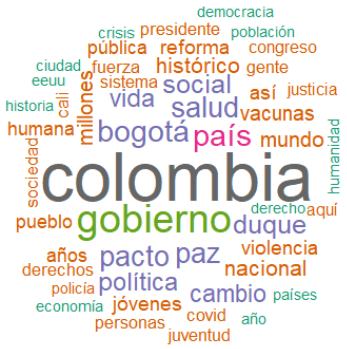
Utilizamos el texto de cada tweet de la BD Train como información para el entrenamiento del modelo que consistió en una red neuronal para realizar la predicción esperada. Ahora bien, para la generación de la matriz de términos que se utilizaría para el entrenamiento del modelo, realizamos la búsqueda de los términos comunes entre las bases de Train y Test, bajo el supuesto de que esto aumentaría la precisión del modelo; este procedimiento se encuentra en el script "*1_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



El modelo de red neuronal 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 utilizamos capas ocultas.

Una red neuronal toma un vector de entrada de p variables:

$$X = (X_1, X_2, \dots, X_P)$$

Y construye una función no lineal para predecir la respuesta, que en nuestro caso corresponde al nombre del autor de un tweet.

$$y = f(X) + u$$

Para nuestro ejercicio se probaron en kaggle 21 modelos diferentes, cambiando para cada modelo los siguientes parámetros:

- Frecuencia de las palabras a eliminar en la limpieza del texto, probando entre no eliminar y eliminar palabras con frecuencia inferior a 10.
- Número de columnas de las matrices de términos, lo que correspondería a variables del modelo, probando entre dejar todos los términos comúnes o combinaciones entre 500, 1.000 y 2.000 términos.
- Capas del modelo
- Neuronas en cada capa
- Velocidad de aprendizaje del modelo
- Optimizadores
- Funciones de activación y pérdida
- Número de epochs

Luego de realizar combinaciones de estos parámetros para los modelos generados, la mejor predicción alcanzó un puntaje de 0.83 en kaggle, esto corresponde al nivel de *Accuracy* del modelo, para el cual se estructuró una red neuronal de la siguiente manera:

- Frecuencia (n) de las palabras a eliminar en la limpieza del texto: n < 10
- Número de columnas de las matrices de términos: 2.296, columnas comunes entre las BD de Train y Test,
- Capas del modelo: 4 capas
    - Capa 1: capa de entrada, 512 neuronas, función de activación "relu", dropout rate = 0.5.
    - Capa 2: capa oculta, 256 neuronas, función de activación "relu", dropout rate = 0.5.
    - Capa 3: capa oculta, 128 neuronas, función de activación "softmax", dropout rate = 0.5.
    - Capa 4: capa de salida, 4 neuronas, función de activación "softmax", dropout rate = 0.5.
- Optimizador: "adam" con learning_rate = 0.001.
- Función de pérdida: "categorical_crossentropy".
- Métrica: "CategoricalAccuracy".
- Epochs: 200
- Batch size = 2^8
- Validation split = 0.3
- Callbacks =
    - callback_reduce_lr_on_plateau: factor = 0.5, patience = 5, este parámetro se utilizó para reducir la tasa de aprendizaje del modelo si éste no mejoraba durante 5 épocas seguidas.
    - callback_early_stopping: patience = 10, este factor se utilizó para detener el entrenamiento del modelo si éste no mejoraba durante 10 épocas seguidas.

En la tabla 2 se realiza una comparación de las métricas y los resultados de 4 de los modelos generados y evaluados en kaggle. Los datos de esta tabla muestran que cómo se fueron ajustando los parámetros para poder lograr cada vez una mejor predicción llegando a obtener un *Accuracy* de 0.83 del modelo antes especificado, que en este caso corresponde al Modelo 9.

Tabla 2. Comparación modelos

Parámetro	Modelo 9	Modelo 3	Modelo 2	Modelo 1
Accuracy Kaggle	0.83000	0.80666	0.78666	0.74333
Número de columnas de las matrices de términos	2.296	2.296	2.296	1.500
Capas	4 Capas: -512 neuronas	2 Capas: -10 neuronas	2 Capas: -2 neuronas	2 Capas: -2 neuronas

Parámetro	Modelo 9	Modelo 3	Modelo 2	Modelo 1
	-256 neuronas -128 neuronas -4 neuronas	-4 neuronas	-4 neuronas	-4 neuronas
Optimizador	adam con learning_rate = 0.001	adam	adam	adam
Función de pérdida	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy
Métrica	Categorical Accuracy	Categorical Accuracy	Categorical Accuracy	Categorical Accuracy
Epochs	200	100	100	100
Batch size	2^8	2^8	2^8	2^8
Validation split	0.3	0.2	0.2	0.2
callback_reduce_lr_on_plateau	factor = 0.5, patience = 5			
callback_early_stopping: patience	10			

4. Conclusiones.

La política colombiana actual tiene lideres con ideologías diferentes que utilizan Twitter para comunicar y debatir sobre temas coyunturales, convirtiendo los tweets en una fuente de información interesante para identificar patrones de comportamiento y tendencias de opinión. En el ejercicio de Machine Learning descrito en este documento, basado en la utilización de la información de texto de los tweets de tres políticos colombianos, se buscaba caracterizar parte de las tendencias políticas que definen las posturas de estas tres líneas de opinión y de esta manera hacer un ejercicio de predicción del autor.

En el trabajo de ajuste y especificación de los modelos se cambiaron continuamente los parámetros, aumentando la complejidad de los modelos cada vez. De esta manera, se identificó el modelo con mayor capacidad predictiva, el cual logro un accuracy de 83%. Este resultado se obtuvo aumentando las capas ocultas y el número de neuronas utilizadas en cada capa en relación con los primeros modelos que se calcularon, los cuales usaban una capa y menos neuronas. También fue relevante aumentar el número de veces que los datos pasaban por la red neuronal para alcanzar el mejor resultado descrito.

Como resultado, el uso de información de los tweets de los políticos colombianos nos permitió entrenar un modelo de redes neuronales para predecir quien lo escribió. Este es un acercamiento que proporciona nuevas formas de identificar patrones en los discursos políticos, diferencias en las líneas de política pública, y evidenciar los temas más relevantes dependiente de estas tendencias políticas. Esta información puede ser usada para futuros

análisis sobre tendencia política en el país y para evidenciar la agenda política en las diferentes líneas de ideológicas de la política colombiana.

5. Bibliografía.

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

Hakan Mehmetcik, Melih koluk and Galip Yuksel. (2022). *Perceptions of Turkey in the US Congress: A Twitter Data Analysis.* Uluslararası İlişkiler / International Relations. Vol. 19, No. 76 (2022), pp. 6989. URL: https://www.jstor.org/stable/10.2307/27195129.