Universidad de Los Andes
Facultad de Economía
Big Data

Grupo No. 2 - Integrantes:
Alexandra Rizo
Id. 202210094

Danna Camila Bolaños
Id. 201911675

Héctor David Taticuán
Id. 202225884

Carlos David Vergara Díaz
Id. 201414896

Fecha de entrega: 20 de marzo de 2023.

Taller Grupal No. 4

Problem Set No. 4 – Predicting Tweets

1. Introducción:

Por medio de este documento, se presentan los resultados y análisis frente al Problem Set No. 4 *Predicting Tweets* del curso Big Data de la Facultad de Economía de la Universidad de Los Andes.

1.1. Twitter y la Política: Álvaro Uribe, Gustavo Petro y Claudia López en el contexto político colombiano.

1.1.1 Twitter y la política:

El uso de redes sociales en el contexto político es un asunto ampliamente debatido, dado el impacto que tiene la información compartida en dichas plataformas en el resultado de la vida política de un Estado. Para el caso particular de Twitter, dada la facilidad en su uso, y el alcance que tiene, junto al hecho de que permite entregar información rápida desde los actores políticos a la sociedad, es posible que se generen dudas en la forma como los trinos (tweets), influencian el pensamiento o comportamiento de sus seguidores[1].

Twitter, es una red social creada en 2006 por varios estudiantes de NYU como una nueva forma de comunicaciones para que sus usuarios entregaran mensajes cortos hacia grupos de personas[2]. Para 2008, Twitter contaba con 300.000 tweets por día, y en un lapso de 2 años alcanzó a casi 50.000.000 de tweets por día, llegando a ser una de las redes sociales más populares, no solo para que las personas compartieran sus ideas, sino también para entornos comerciales y políticos que buscaban hacer llagar mensajes a grupos de personas[3]. En palabras de Antonie Boutet et al, citando a Tumasjan y Cha:

> "*Social media such as Facebook and Twitter have revolutionised the way people communicate with each other. Users generate a constant stream of online messages through social media to share and discuss their activities, status, opinions, ideas and interesting news stories; social media might be an effective means to examine trends and popularity in topics ranging from economic, social, environmental to political issues.*
>
> *In modern politics, political parties must have an online presence to reach the users they want to influence. They also monitor social media to measure the success of their political campaigns and then refine their strategies (e.g. to help their*

[1] Frente a este punto ver: John H. Parmelee y Shannon L. Bichard. *Politics and the Tweeter Revolution: How Tweets Influence the Relationship between Political Leader and the Public.* Ed. Lexington Book. (2012).
[2] Pennington Creative. *The Originis of Twitter.* (s.f.). Disponible en: https://penningtoncreative.com/the-origins-of-twitter/
[3] Marketing Zone Icesi. *Esta es la historia de Twitter, la app que revolucionó la comunicación en 140 caracteres.* (2023). Disponbile en: https://www.icesi.edu.co/marketingzone/esta-es-la-historia-de-twitter-la-app-que-revoluciono-la-comunicacion-en-140-caracteres/

> *candidates win in elections). This phenomenon creates a new opportunity for users to participate to democratic process and trends to increase civic participation from users and changes how people are engaging in politic debates. The promise of social media starts to be exploited by government agencies from different countries as vehicles for consultation and for enhancing civic engagement*"[4]

No obstante, y pese a los eventuales beneficios que puede representar el uso de redes sociales en el contexto político, para muchos autores y sectores de la opinión pública, las redes sociales constituyen un peligro para la democracia, esto, dado al nivel de influencia que pueden tener sobre el electorado y la falta de controles en cuanto al tipo de información que es transmitida. Un caso paradigmático es el estudio de la presidencia de Donald J. Trump, (2017-2021), y la influencia de twitter en diferentes procesos democráticos adelantados en Estados Unidos en aquel entonces. Frente a este punto, indican Gladiere et al:

> *"Every few decades, presidents and candidates for president upend American politics by turning new technology into a potent political weapon. Consider how Franklin Roosevelt used the radio to hold "fireside chats" with a Depression-wracked American public, or how John F. Kennedy and Ronald Reagan used their superior understanding of the visual component of television to build support for themselves and to stymie their political opponents. (…)*
>
> *To this list we must now add Donald Trump, whose use of Twitter—the social networking platform through which users can broadcast their thoughts to the world in 140-character morsels known as "tweets"—was crucial to his political rise and his unlikely victory in the 2016 presidential election. Trump is not the first politician to use Twitter since it was launched in 2006; having a Twitter presence has become as necessary for a presidential candidate as having a web site, television ads, and campaign volunteers. But Trump's use of the platform was—and continues to be—unique. Before becoming a candidate, Trump used Twitter to turn himself from a real estate mogul turned reality show host into a political gadfly, propagating absurd, racist conspiracy theories about President Obama's birthplace and citizenship and weighing in on other political developments. As a candidate in 2016, some of Trump's late-night (or early-morning) tweets set the terms of a day's morning news cycle. Other tweets took aim at his opponents in the Republican primary and in the general election, or at reporters, news organizations, and even private citizens who criticized him or his campaign."*[5]

Todo ha llevado a que los analistas políticos estudien el impacto de twitter en la política, llegando a generar modelos para medir, por ejemplo, las palabras más usadas por los candidatos dentro de sus trinos, o las tendencias de re twitteo de los trinos de cada candidato, tendencias y otros aspectos[6].

[4] Antonie Boutet et al. *What´s in Twitter, I know what parties are popular and who you are supporting now!* Soctial Network Analysis and Mining, no. 3. (2013).
[5] Christopher J. Galdierei, Jeniffer C. Lucas y Taua S. Sisco. *Introduction: Politics in 140 Characters or Less* en "The Role of Tweeter in the 2016 US Election. Ed. Palgrave Pivot. (2017).
[6] Kevin Macnish y Jai Galliot. *Big Data and Democracy.* Ed. Edinburg Press. (2020).

1.1.2 El contexto político nacional: Gustavo Petro, Álvaro Uribe y Claudia López como usuarios activos en la plataforma.

Las redes sociales, sin duda, han modificado la manera en cómo se comunica la sociedad y ha transformado de manera exponencial la comunicación política. Nuestro país no ha sido ajeno a esa transformación y actualmente los políticos utilizan las redes sociales para darse a conocer y para informar y a veces, desinformar sobre lo que sucede en el territorio. Gustavo Petro, Álvaro Uribe y Claudia López lo han entendido muy bien y han utilizado Twitter como una plataforma de comunicación importante para informar, comunicar y hasta rendir cuentas.

Gustavo Petro es el primer presidente de izquierda de Colombia y luego de muchas décadas de hacer oposición y activismo en las calles. Es economista de profesión y ex guerrillero del M-19 en los años 90. Fue congresista cuatro periodos y también alcanzó la Alcaldía de Bogotá en el año 2012.

Para el actual presidente, la red social Twitter es la mejor manera de comunicar; él asegura que, entre más comunicación, menos violencia. Dicho esto, sus mayores y más importantes anuncios los hace a través de Twitter. Tanto así que sus alocuciones desde la Casa de Nariño se han reducido en comparación con los anteriores presidentes y sus ruedas de prensa las lidera el vocero del Gobierno Nacional, el ministro Alfonso Prada, esto porque se siente más cómodo argumentando y fortaleciendo posición desde los 140 caracteres de Twitter. Petro ha entendido muy bien los tiempos en los que gobierna y que sus trinos serán replicados por los principales medios de comunicación del país, dándole así un alcance Nacional, y en ocasiones global, a sus anuncios más importantes.

Actualmente el presidente cuenta con 6.6 millones de seguidores[7] y sus trinos alcanzan las dos mil interacciones. La cuenta de Twitter de Gustavo Petro se caracteriza por tener anuncios presidenciales, respuestas polémicas a otros tuiteros y posiciones ideológicas claras.

En la otra orilla del espectro tenemos al expresidente Álvaro Uribe, político colombiano desde la década de los 80 cuando fue nombrado alcalde de Medellín, posteriormente fue senador y a inicios de los años 2000 se posesionó como presidente de Colombia. Es un hombre de trayectoria política muy amplia y ha sabido consagrarse como el actual líder de la oposición en cabeza de su partido Centro Democrático. Uribe también ha entendido muy bien la dinámica de las redes sociales y en Twitter es muy activo para enviar mensajes a sus seguidores y copartidarios.

Actualmente el expresidente suma más de cinco millones de seguidores[8] en Twitter y sus trinos están entre los dos mil y las siete mil interacciones. Su cuenta se caracteriza por tener espacios de opinión y oposición.

Finalmente, en lo que al parecer se considera la mitad del espectro político, está la alcaldesa de Bogotá D.C., Claudia López. Fue activista y denunció la parapolítica en el año 2006, posteriormente fue Senadora de la República y para el año 2020 ganó la Alcaldía de Bogotá.

[7] Consultado de la cuenta de Gustavo Petro el 18 de marzo de 2023. https://twitter.com/petrogustavo
[8] Consultado de la cuenta de Álvaro Uribe Vélez el 18 de marzo de 2023. https://twitter.com/AlvaroUribeVel

Claudia López cuenta con 2.8 millones de seguidores en Twitter[9] y sus trinos están entre los 500 y mil interacciones. Su cuenta se caracteriza por hacer anuncios de política pública en Bogotá, oponerse a anuncios presidenciales como las medidas que pretenden tomarse desde el nivel central frente al Metro de Bogotá y finalmente utiliza esta plataforma para contestarle a sus seguidores y detractores los comentarios expuestos allí.

En este marco, y considerando la información suministrada, es que se realizará un análisis desde modelos neuronales a los trinos (Tweets) de estos tres (3) actores políticos.

1.2. Estructura del documento:

Para llegar a lo anterior, y considerando lo expuesto, el documento se dividirá en los siguientes acápites, a saber: Primero, enlace al repositorio de GitHub; Segundo, proceso de limpieza de datos y estadísticas descriptivas; tercero, análisis de los modelos y; cuarto, conclusiones.

2. Enlace a repositorio de GitHub:

El enlace de Github donde podrá encontrarse el repositorio con las respuestas del taller es el siguiente:

Enlace al repositorio en GitHub
https://github.com/Carlosvergara1995/Problem_Set_4_Predicting_Tweets.git

3. Descripción de los datos:

A continuación, se presenta información sobre los datos utilizados dentro del ejercicio, así como las acciones tomadas para su limpieza y procesamiento.

3.1. Origen de los datos:

Los datos para realizar el entrenamiento del modelo son en total 9349 registros de tweets, repartidos de la siguiente forma:

- Claudia López: 3470
- Gustavo Petro: 2877
- Álvaro Uribe: 3002

Se observa que es una muestra balanceada, por lo tanto, no se realiza ningún tipo de balanceo.

Respecto al tiempo, no se tiene información en qué fecha se realizó cada uno de los trinos. Para test se tiene 1500 trinos, los cuales solo cuenta con una columna de id y no con la persona que realizó el trino, lo cual es el objetivo de este taller.

[9] Consultado de la cuenta de Claudia López el 18 de marzo de 2023. https://twitter.com/ClaudiaLopez

3.2. Limpieza y pre-prosesamiento de datos:

Para la limpieza de datos se realizaron los siguientes pasos:

Primero, todos los datos (palabras) se transforman a minúscula; segundo, se quitan características especiales como las comas, entre otros; tercero, se elimina todas las características especiales que no sean letras o números, el objetivo es eliminar signos de preguntas, signos de interrogación, de puntación entre otros; cuarto, se elimina las palabras recurrentes, dado que estas palabras generan ruido en nuestro análisis. La eliminación se realizó por medio del paquete "NLTK" en Python, el cual se utiliza para la limpieza de datos. La lista de stopwords de este paquete cuenta con 315 palabras predeterminadas, pero dado que no se encuentra "Usted" y "Ustedes" se agrega manualmente y quinto, finalmente se aplica lematización con el paquete WordNetLemmatizer.

A manera de ejemplo se presenta una de las frases antes de la limpieza y el preprocesamiento:

- Frase antes de procesar: Esto significa el proyecto de acuerdo de rescate social y económico que estamos tramitando ante el Concejo de Bogotá D.C.

- Frase después del preprocesamiento: significa proyecto acuerdo rescate social económico tramitando Concejo de Bogotá D.C.

3.3. Análisis descriptivo:

Con fundamento en las palabras encontradas en las bases de datos objeto de estudio, se encontró que los actores políticos utilizaron de manera más reiterada las siguientes palabras u expresiones. Se advierte que entre mayor sea el tamaño de la palabra o expresión en la gráfica, más veces ha sido utilizada por el usuario (o la usuaria) de Twitter:

Claudia López **Cargo actual:** Alcaldesa mayor de Bogotá D.C.

Fuente: Elaboración propia

Gustavo Petro Urrego **Cargo actual:** Presidente de la República

Fuente: Elaboración propia

Álvaro Uribe Vélez **Cargo actual:** Cabeza del principal partido de oposición al Gobierno Nacional

Fuente: Elaboración propia

De los datos se puede ver lo siguiente:

Principales palabras usadas en los tweets de Claudia López **Cargo actual:** alcaldesa mayor de Bogotá D.C.	
Palabra/expresión	**Número**
Bogotá	1273
Hoy	602
Ciudad	487
Gracias	341
Año	332
Día	292
Toda	285
Vacunación	268
Jóvenes	261
Cuidado	257

Principales palabras usadas en los tweets de Gustavo Petro Urrego **Cargo actual:** Presidente de la República	
Palabra/expresión	**Número**
Colombia	474
Gobierno	273
Si	239
Hoy	216
País	197
Debe	190
Bogotá	181
Solo	181
Pacto	177
Ser	176

Principales palabras usadas en los tweets de Álvaro Uribe Vélez **Cargo actual:** Cabeza del principal partido de oposición al Gobierno Nacional	
Palabra/expresión	**Número**
Policia	696
Onza	299
Colombia	251
Familia	201
Solidaridad	187
Medellin	145
País	144
Tonelada	138
Hoy	135
Toda	135

3.4. Creación de la matriz TF-IDF

Para la creación de la matriz de pesos se usa el paquete TfidfVectorizer de sklearn en Python. Una matriz TF-IDF (Term Frequency-Inverse Document Frequency) es una representación numérica de un corpus de documentos que se utiliza comúnmente en la minería de texto y el procesamiento del lenguaje natural. Es una matriz que refleja la frecuencia con que aparecen las palabras en un conjunto de documentos, y que tan importante o reiterada es cada palabra en el contexto del corpus completo.

La función TfidfVectorizer se entrena con los datos de train, dado que al agregar los datos de test se estaría dando información adicional al train, por lo tanto, se entrena con train y con el modelo entrenado se realiza la transformación tanto en test como en train. Adicionalmente se agregarón dos parámetros, para no seleccionar palabras menores que no tengan al menos dos caracteres y que la muestra debe estar en al menos el 0.04 % del documento, esto, con el propósito de no tener en cuenta toda la matriz, dado que, si se seleccionan todas las palabras, se superarían las 26.000 columnas, y seleccionando 0.04% se seleccionó alrededor de 6000 columnas, las cuales son las que tienen mayor peso, dado su repetición.

4. Modelos:

Este es un problema de clasificación, por lo tanto, se aplicaron modelos de clasificación como son:

- Regresión Logística.
- Naïve-Bayes
- Random forest

- Redes neuronales

4.1. Entrenamiento y selección de hyper-parámetros:

Para la selección en todos los casos se utilizó validación cruzada con 15 folds, esto dado que no se dispone las etiquetas de test para realizar la verificación del modelo. Con esto se buscó que el accuracy de los datos de train sea lo más parecido a los datos de test. Finalmente, el score de selección de modelos fue el accuracy y para cada modelo se utilizó gridsearch con hiperparametros diferentes para cada modelo.

4.1.1 Regresión logística

Dado que es una matriz dispersa se decidió utilizar regularización como uno de los hiperparametros, con los siguientes valores posibles [0, 0.001, 0.01, 0.1, 1]. Se habría podido utilizar más valores, con la finalidad de obtener una mayor precisión, pero dado que es una matriz de tamaño considerable se decidió acotarlos. Adicionalmente, al elegir estos valores el modelo puede elegir qué tipo de penalización utilizar, ya se Lasso (cuando toma el valor de cero), Ridge (cuando toma el valor de 1) o elastic net.

El segundo hiperparametro, se da en relación con los valores de lambda, los cuales inicialmente variaban entre 1 y 30, con saltos de 5, pero en ninguna de las pruebas, estos superaron el valor de 10. Por ello, para una mejor precisión, se cambiaron los intervalos de 1 a 10 con saltos de 1. Así, se estipula un valor superior a 1 debido a que es una matriz dispersa, por lo tanto, la regularización esperada es fuerte.

4.1.2 Random forest

Para la realización de este modelo, se ajustan varios hiperparametros los cuales son los siguientes: (i). El número de árboles en el bosque. Se prueban tres opciones diferentes: 10, 100 y 1000, en la primera estimación descubrimos que con 10 y 100 árboles no convergía en algunos casos, por lo tanto, se cambió a 500 y 100, para disminuir el número de modelos probados; (ii). La profundidad máxima de cada árbol en el bosque. Se prueban seis opciones diferentes: None (ninguna restricción en la profundidad), 10, 20, 30, 40 y 50; (iii). El número mínimo de observaciones necesarias para dividir un nodo interno. Se prueban tres opciones diferentes: 2, 3 y 4, número mínimo de observaciones necesarias para ser una hoja. Se prueban tres opciones diferentes: 1, 2 y 4. Estos valores se fueron ajustando dado que los hiperparametros no tomaron valores extremos, esto con la finalidad de reducir el tiempo de computación.

Estos hiperparametros se combinan en todas las posibles formas, generando múltiples modelos. Se crea un nuevo clasificador de Random Forest con los mejores parámetros encontrados por GridSearchCV y se ajusta a los datos de entrenamiento y se utiliza el modelo entrenado para realizar las predicciones.

4.1.3 Naive Bayes Multinomial

Este modelo considera el número de apariciones de cada termino para evaluar la contribución de su probabilidad condicional dada la clase del documento. Es similar al modelo simple, pero el modelo simple determina si algo está o no, dado que es binomial, por eso para este análisis

de texto se usa el multinomial; además que su implementación en sencilla. Solo utilizaremos un hiperparametro de suavizado aditivo (alpha). Este se usa para realizar una estimación de las probabilidades condicionales en Naive Bayes y se usa fundamentalmente para evitar multiplicar por cero, los valores elegidos para probar el modelo son 0.01, 0.095,0.1,0.12, 0.2, 0.3, 0.15, 0.5, 0.9, 1, 2, 5. Si bien se analiza un número más grande de hiperparametros, el tiempo de cómputo es corto.

Este modelo de clasificación exige que los términos sean independientes, lo cual aplica para este tipo de problemas.

4.1.4 Redes Neuronales

Para este modelo se utilizó un clasificador MLP. El MLP es un tipo de red neuronal artificial que se compone de múltiples capas de neuronas interconectadas, donde cada capa procesa y transforma la información que recibe. Se selecciono este tipo de neurona con base en la literatura revisada donde se encontró que este modelo es una opción ampliamente utilizada para clasificar debido a su capacidad para aprender datos no lineales.

Para entrenar el MLP, se ajustaron los hiperparámetros del modelo, utilizando la técnica de GridSearch con validación cruzada con 5 folds, este valor es menor al utilizado en los modelos anteriores, dado que el tiempo de cómputo es mayor. Los hiperparametros que se ajustaron incluyen:

- hidden_layer_sizes: El número de neuronas en la capa oculta del MLP. Se probaron tres opciones: (10,), (50,), y (100,).

- alpha: Un parámetro de regularización que controla la complejidad del modelo. Se probaron tres opciones: 0.0001, 0.001, y 0.01.

- max_iter: El número máximo de iteraciones que el MLP puede realizar durante el entrenamiento. Se probaron dos opciones: 500 y 1000.

- solver: El optimizador utilizado para minimizar la función de pérdida durante el entrenamiento. Se probaron dos opciones: 'lbfgs' y 'adam'. Existen otros métodos de solución, pero se espera que se seleccione Adam, dado que es un optimizador popularmente utilizado en el entrenamiento de redes neuronales para clasificación de texto debido a su eficacia en la adaptación de la tasa de aprendizaje y en la actualización de los pesos de la red.

4.2. Comentarios frente al modelo seleccionado:

Para la selección de modelos se considera la siguiente tabla:

Selección de modelos			
Modelo	**Hiperparametros seleccionados**	**Accuracy con CV**	**Tiempo de computación**
Regresión logística	Alpha: 0 (Ridge) - Lambda: 3	0.8468260827811388	44 m 26.3 s

Selección de modelos			
Modelo	**Hiperparametros seleccionados**	**Accuracy con CV**	**Tiempo de computación**
Random forest	max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 1000	0.781262089969955	45m 0.1s
Naive Bayes Multinomial	Alpha : 0.2	0.8369833038372365	0.8s
Redes neuronales	alpha: 0.0001, hidden_layer_sizes: (10,), max_iter: 1000, solver: 'adam'	0.8139897511609343	1433m 47.9s

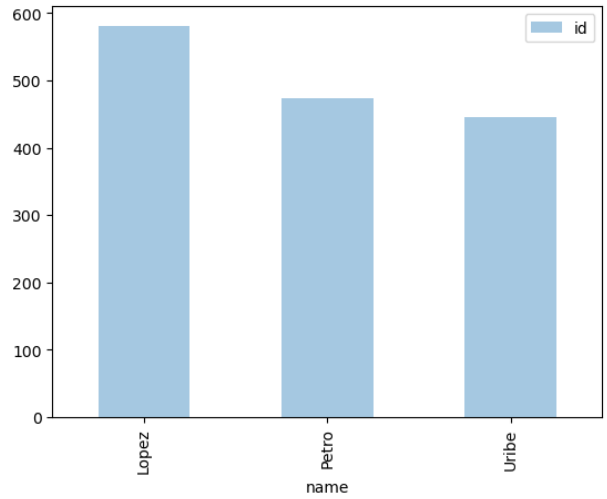
Como se esperaba, la regularización usada fue ridge, con una lambda mayor a tres, pero destacan los tiempos de cómputo. Por ejemplo para redes neuronales fue superior a 24 horas, mientras que para Naive Bayes fue menor a un segundo, además el accurracy de Naive Bayes es cercano al de la regresión logística. Por esta razón se decide selección el modelo de Naive Bayes y mejorar el acurracy por medio de una modificación a la matriz tfi-df.

- Se elimina el parámetro de aparición mínima de 0.04%. El numero de variables cambio a 28895 y el accurracy aumento a 0.853671749598716, igualmente el tiempo de cómputo aumenta 5.7 s.

- A la modificación anterior se adiciona que agregue frases de 2 palabras, aumentando el número de variables a 154 081, el accurracy a 0.8696092453114925 y el tiempo de cómputo a 2.4 s.

El modelo seleccionado es Naive Bayes Multinomial con Alpha de 0.1, para una matriz rfi-df con todas las palabras y frases de 2 palabras, principalmente por tener un buen accuracy y tiempos de cómputos muy bajos. Los resultados son los siguientes:

Resultados del modelo Naive Bayes Multinomial con Alpha de 0.1



Fuente: Elaboración propia

Se observa que la respuesta es balanceada y las mayores predicciones son para Claudia López con 581 predicciones, Petro con 474 y finalmente Uribe con 445.

5. Conclusiones:

El modelo de clasificación de Naive Bayes multinodal es el clasificador más sencillo de implementar y fue capaces de producir soluciones muy eficientes en comparación con aquellas que se han obtenido a través de métodos más sofisticados, adicionalmente el tiempo de cómputo de este modelo es significantemente menor al que presentaron los demás.

A raíz de la sencilles del modelo se encontró un mejor ajuste tanto dentro como fuera de muestra,