**Problem Set 4: Predicting Tweets**

1. **Introducción**

En los últimos años, se ha generado un especial interés por parte de la academia para realizar análisis técnicos sobre las redes sociales que conduzcan a generar evidencia, principalmente en época de comicios, para la toma de decisiones de los partidos políticos, candidatos y ciudadanos. Caso puntual objeto de análisis es la red social Twitter, la cual se ha convertido en una de las más empleadas en el mundo, al ser un espacio que permite a las personas comunicar y estar en contacto a través de mensajes rápidos y frecuentes (Eskibel, 2022), y especialmente por las siguientes características (Gomila, 2020):

1). Es una plataforma que favorece la viralidad de la información publicada en comparación con otras redes sociales, dado su diseño, interfaz y disponibilidad de recursos (hashtags, retuits, etiquetas o trendic topics); 2). Es la red social más utilizada por el sistema político y mediático, es decir, a pesar de no ser la plataforma más empleada, es la que tiene el público más influyente; 3). Twitter se ha convertido en una herramienta fundamental para el diseño y la ejecución de la estrategia comunicativa en campaña electoral de los partidos políticos y; 4). Su estructura facilita a los investigadores un acceso más directo a los datos, simplificando su extracción, procesamiento y análisis (Gomila, 2020).

Adicionalmente, esta social media cuenta con características que la hacen única a la hora de compararlas con otras. Entre las cuales se destacan: 1). Impacto: Están los líderes de opinión, políticos, gobernantes, candidatos, dirigentes, periodistas, medios de comunicación, referentes sociales y culturales, formadores de opinión, gente activa e influyente; 2). Noticia: Suele ser una [vía más directa para salir en las noticias](https://danieleskibel.com/noticias/) que las tradicionales ruedas de prensa; 3). Contacto: es una forma rápida de contactar con personas relevantes para el político; 4). Brevedad: un tweet son 140 caracteres. No es una columna, ni un discurso, ni unas largas declaraciones; 5). Velocidad: es una herramienta ideal para difundir o seguir en tiempo real las novedades de un evento que está desarrollándose; 6). Interacción: permite intercambiar ideas, dialogar, discutir, defender posiciones, responder; 7). Síntesis: facilita a los políticos sintetizar su pensamiento en frases breves que sirvan para un título periodístico y que sean fácilmente recordadas; 8). Receptividad. El público de Twitter recibe con mucha mayor naturalidad los mensajes vinculados a la política; 9). Movilidad: Twitter es perfecto para usar desde los smartphones (Eskibel, 2022).

En este sentido, Twitter se ha convertido en un espacio para la difusión de ideas, información y sentimientos por parte de los usuarios. No es la más utilizada a nivel mundial, pero sí es la más influyente en el discurso público. Por esta razón, en las últimas décadas se ha convertido en la plataforma para que muchos políticos de todo el mundo se comuniquen con su audiencia y puedan dar a conocer sus iniciativas de política pública y de mandato de gobierno. Ejemplo de ello, se ha visto en las elecciones de Estados Unidos, pues la campaña electoral de Obama en 2008 marcó el camino. Twitter fue una de sus herramientas más novedosas, y el día de la elección ya tenía 100 mil seguidores. En 2012, apenas 4 años después, Obama llegó al día de la elección con 20 millones de seguidores (Eskibel, 2022).

Sin embargo, por la cantidad de cifras que se manejan en las redes sociales, y en especial esta, resulta complejo hacer predicciones sobre sus tweets. Este estudio se basa en un modelo predictivo que contribuye a predecir a quién pertenece cada tuit. El conjunto de datos de capacitación contiene tuits de las cuentas de tres destacados políticos colombianos: Claudia López, Gustavo Petro y Álvaro Uribe.

En el presente documento se consideran cuatro (4) modelos predictivos y se profundiza en el que presenta mejor desempeño. Se evalúan una regresión lineal simple, regresiones lineales regularizadas de Lasso y Ridge y, finalmente, un modelo Random Forest. Como resultado del ejercicio, se obtiene que el modelo Random Forest presenta el mejor desempeño en la predicción de tweets, con un RMSE (Root Mean Squared Error) de COP $ 239.012.001,98.

**Nota:** La base de datos usada, al igual que el script de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace: <https://github.com/Yilap/Repositorio_Taller4.git>

**Contexto**

En Colombia, el 81% de la población es usuaria activa de las redes sociales, lo que equivale a 41,8 millones de personas. De este porcentaje, el 52% de los usuarios son mujeres y el 48% son hombres. El 37% de las mujeres y el 35% de los hombres se encuentran entre los 18 y 44 años. Por otra parte, en promedio, los colombianos destinan aproximadamente 3 horas y 46 minutos del día a conectarse a redes sociales (8 redes diferentes). Para el 2020, este porcentaje fue de 69%, lo que demuestra un crecimiento de 12 puntos porcentuales y en lo transcurrido entre 2014 y 2022 se estima que la cantidad de usuarios se ha incrementado en 109% (Rosgaby, 2022).

En la siguiente gráfica se observa que la plataforma más usada por los colombianos es Whatsapp. Para el 2020, el 95% de los usuarios usaba Facebook y el 77% Instagram, por lo que para el 2022 Facebook tuvo una pérdida de -3.3 puntos porcentuales e Instagram una ganancia de +7.4 puntos porcentuales (Rosgaby, 2022). En el caso de Twitter, esta red tiene una audiencia de 4.3 millones de usuarios, es decir, el 8.4% de la población de nuestro país y el 12% de los usuarios de Internet (Rosgaby, 2022).

**Tabla No. 1** Uso de redes sociales en Colombia



**Fuente:** Branch (2022)

En cuanto al uso de Twitter por parte de Claudia López, Gustavo Petro y Álvaro Uribe, se puede afirmar lo siguiente:

1. **Datos**
2. ***Descripción de las fuentes de datos***

Para el desarrollo de este Problem Set se utilizará un conjunto de datos de prueba que contiene 500 tweets sin etiquetar, el cual es extraído de a través de acceso programático a los datos de Twitter mediante las API (interfaces de programación de aplicaciones). De esta manera, Twitter permite acceder a partes del servicio mediante las API para permitir la creación de software que se integre con Twitter.

Los datos de Twitter son únicos y se extraen a partir de datos de la mayoría de las redes sociales. API ofrece acceso amplio a estos datos que los usuarios han decidido compartir de manera pública, en este caso los políticos colombianos objeto de estudio. Estos datos están conformados por dos bases de datos: la base test y la base train.

1. ***Análisis descriptivo de los datos (estadísticas descriptivas)***

Teniendo en cuenta que Twitter es una de las redes sociales mas utilizada por los políticos, en este caso por Gustavo Petro, Alvaro Uribe y Claudia lopez, se realizo un análisis a fin de conocer cuál es el que más Tweets tiene y a su vez, cual es la palabra que más escriben en sus publicaciones. En este sentido, y como se mencionó anteriormente, para el análisis de los modelos de predicción se utiliza una muestra tomada por las API de Twitter, sobre cuantos Tweets tiene cada uno de los representantes objeto de estudio, esto por medio de una base datos de test de 1500 y una base de datos train de 9349 observaciones respectivamente.

Seguido de esto, se usó el paquete “tidytext”, el cual tiene la función llamada unnest\_tokens, la cual permitió separar los Tweets por tokens (este caso son las palabras), para cada a uno de los documentos, luego se construyó un objeto matriz, utilizando el criterio TFIDF el cual permite saber, cual es la palabra más representativa de cada uno de los personajes objeto de estudio. Esto se construyó uniendo los datos train y test, con el objetivo de tener el inventario completo de las palabras que eventualmente pueden salir. Finalmente, se procedió a montar la información en cada uno de los modelos citados.

Por otro lado, se puede evidenciar que en las estadísticas descriptivas, Claudia lopez es la que más Tweets tiene, con un total de 3470 Tweets, seguida de Alvaro Uribe con 30002 Tweets y finalmente Gustavo Petro con 2877 Tweets. Respecto al número de caracteres y teniendo en cuenta que en esta red social solo esta permitido un máximo de 180 caracteres, se observa que en promedio Claudia Lopez es la que más caracteres coloca en sus publicaciones con un total de 242 caracteres, seguida de Petro( 193 caracteres) y Uribe( 160 caracteres). Finalmente, se observa que Lopez es la que mas categoriza sus Tweets por medio del numeral, con un porcentaje del 46%, seguida de Uribe(12%)y Petro(8%). (Ver gráfica No. 1).

**Tabla No. 1.** Estadísticas descriptivas variables numéricas

| **Descripción de los tweets** | | | |
| --- | --- | --- | --- |
| Base de datos de entrenamiento | | | |
| Autor | No. | Núm. Caracteres | Pct. que contiene numerales |
| López | 3.470 | 242,62 | 0,46 |
| Petro | 2.877 | 193,58 | 0,08 |
| Uribe | 3.002 | 160,05 | 0,12 |

**Grafica No. 1.** Nube de palabras: Álvaro Uribe



**Fuente:** R Studio

**Grafica No. 2.** Nube de palabras: Claudia López



**Fuente:** R Studio

**Grafica No. 3** Nube de palabras: Gustavo Petro

****

**Fuente:** R Studio

1. **Modelos y resultados**
2. **Conclusiones y recomendaciones**
3. **Bibliografía**

Eskibel, D., (2022). 10 razones por las que los políticos prefieren Twitter. Recuperado de: <https://danieleskibel.com/twitter10/>.

Gomila, G., (2020). ¿Para qué usan Twitter los partidos en campaña?. Recuperado de: <https://agendapublica.elpais.com/noticia/13745/qu-usan-twitter-partidos-campana>.

Rosgaby, M., (2022). Estadísticas de la situación digital de Colombia en el 2021-2022. Branch – Marketing digital. Recuperado de: <https://branch.com.co/marketing-digital/estadisticas-de-la-situacion-digital-de-colombia-en-el-2021-2022/>

Twitter., (s.f.). Información sobre las API de Twitter. Recuperado de: <https://help.twitter.com/es/rules-and-policies/twitter-api>

Twitter., (s.f.). Preguntas frecuentes para usuarios nuevos. Centro de ayuda. Recuperado de: <https://help.twitter.com/es/resources/new-user-faq>