**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 elecciones, 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. 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. Se ha convertido en una herramienta fundamental para el diseño y la ejecución de la estrategia comunicativa en campaña electoral de los partidos políticos
4. Su estructura facilita a los investigadores un acceso más directo a los datos, simplificando su extracción, procesamiento y análisis

En los últimos años, Twitter se ha convertido en un espacio fundamental para la difusión de ideas, información y sentimientos por parte de los usuarios. Aunque no es la plataforma más utilizada a nivel mundial, sí es la más influyente en el discurso público. Por esta razón, muchos políticos de todo el mundo la utilizan para comunicarse con su audiencia y dar a conocer sus iniciativas de política pública y de gobierno. Un ejemplo destacado de esta situación son las elecciones presidenciales de Estados Unidos, en las cuales la campaña electoral de Barack Obama en 2008 marcó un antes y un después. En esa ocasión, Twitter fue una de las herramientas más novedosas utilizadas por Obama, y para el día de la elección ya contaba con 100.000 seguidores. Cuatro años después, en 2012, Obama llegó al día de la elección con 20 millones de seguidores en Twitter, lo que demuestra la importancia que esta red social adquirió para la comunicación política en un corto período de tiempo (Eskibel, 2022).

Con esto en mente, el presente documento tiene como objetivo presentar un modelo predictivo que contribuye a determinar a quién pertenece cada tuit basado en el contenido de este, considerando que el lenguaje refleja los valores e ideales de las personas que publican. La principal motivación de este estudio surge con la expresión "A rose by any other name would smell as sweet" (en español, "Una rosa con otro nombre olería igual de dulce") de la obra de William Shakespeare "Romeo y Julieta".

El conjunto de datos de entrenamiento contiene tuits de las cuentas de tres destacados políticos colombianos: Claudia López, Gustavo Petro y Álvaro Uribe. 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 accuracy de XXX%.

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**

Twitter se destaca por sus características únicas en comparación con otras redes sociales. Entre ellas, se encuentran (Eskibel, 2022):

1. Impacto: líderes de opinión, políticos, gobernantes, candidatos, periodistas, medios de comunicación, referentes sociales y culturales, formadores de opinión, entre otros
2. Noticia: una vía más directa para aparecer en las noticias que las ruedas de prensa tradicionales
3. Contacto: una forma rápida de contactar con personas relevantes para el político
4. Brevedad: un tweet son solo 140 caracteres, lo que permite una síntesis efectiva del pensamiento
5. Velocidad: 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 y responder
7. Síntesis: se publican 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

En Colombia, el uso de las redes sociales ha experimentado un crecimiento significativo en los últimos años. Según Rosgaby (2022), el 81% de la población colombiana es usuaria activa de las redes sociales, lo que equivale a 41,8 millones de personas. El 52% de los usuarios son mujeres y el 48% son hombres, y el 37% de las mujeres y el 35% de los hombres se encuentran entre los 18 y 44 años. En promedio, los colombianos dedican alrededor de 3 horas y 46 minutos al día a conectarse a redes sociales, utilizando alrededor de 8 redes diferentes. Para el 2020, el 95% de los usuarios colombianos usaba Facebook, mientras que el 77% utilizaba Instagram. En el caso de Twitter, esta red social cuenta con una audiencia de 4,3 millones de usuarios, lo que representa el 8,4% de la población del país y el 12% de los usuarios de Internet en Colombia.

Como se muestra en la **Figura No. 1,** en la actualidad WhatsApp es la plataforma más utilizada por los colombianos, sin embargo, Twitter sigue siendo una herramienta importante para políticos, líderes de opinión y medios de comunicación debido a sus características únicas y su capacidad para difundir noticias y opiniones de forma rápida y directa. En este caso, es fundamental analizar el comportamiento de tres de los políticos más relevantes hoy en día: Claudia López, Gustavo Petro y Álvaro Uribe.

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



**Fuente:** Branch (2022)

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: test (1.500 observaciones) y la base train (9.349 observaciones).

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

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 280 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 tabla No. 1).

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

| **Descripción de los tweets** | | | |
| --- | --- | --- | --- |
| Base de datos de entrenamiento | | | |
| **Autor** | **No.** | **Núm. Caracteres** | **% que usa hashtags** |
| López | 3.470 | 242,62 | 46% |
| Petro | 2.877 | 193,58 | 8% |
| Uribe | 3.002 | 160,05 | 12% |

Fuente: Elaboración propia

Para profundizar en el análisis, se crean nubes de palabras (wordclouds) basadas en el conjunto de datos de entrenamiento (train dataset) para cada uno de los tres políticos considerados en el estudio: Claudia López, Gustavo Petro y Álvaro Uribe. Las nubes de palabras permiten visualizar los términos más utilizados por cada uno de los políticos en su cuenta de Twitter. Esto puede ser útil para identificar patrones de lenguaje y estilo de comunicación que puedan ayudar a predecir quién escribió cada tuit, además de identificar temas y problemas comunes que son importantes para cada político. El análisis de nubes de palabras es una herramienta clave para identificar patrones y temas en grandes conjuntos de datos de texto.

****En el caso de **Claudia López**, las 10 palabras más relevantes son: Bogotá, hoy, ciudad, gracias, vacunación, jóvenes, cuidado, seguridad, vida y mujeres. Estas palabras están relacionadas con su cargo actual como alcaldesa de la capital del país. La palabra "Bogotá" es la más frecuente, lo cual es de esperarse, ya que se trata de la ciudad que gobierna. La frecuencia de la palabra "hoy" sugiere que la alcaldesa está comunicando acciones y decisiones en tiempo real. La presencia de palabras como "ciudad", "seguridad" y "vida" indica que López está intentando comunicar que está enfocando su gestión en mejorar la calidad de vida de los ciudadanos y en garantizar su seguridad. La aparición de la palabra "vacunación" en la tabla sugiere que la alcaldesa implementó medidas significativas para combatir la pandemia del COVID-19 y proteger a la población. La palabra "jóvenes" indica que la alcaldesa está prestando atención a la juventud y está trabajando en programas y políticas para su bienestar, al igual que "mujeres" sugiere que la alcaldesa está prestando atención a temas de equidad de género y derechos de las mujeres en su gestión. De esta forma, la frecuencia de estas palabras indica que la alcaldesa está comunicando a la población su trabajo en mejorar la calidad de vida de los ciudadanos y en promover la equidad y el bienestar social.

**Gráfica 1 - Nube de palabras: Claudia López  
Fuente: R Studio**

Por otra parte, las palabras más relevantes en los tweets del actual presidente de Colombia, **Gustavo Petro**, son: Colombia, Gobierno, hoy, país, Bogotá, debe, salud, pacto, ser y paz. Si bien se esperaban palabras como Colombia, Gobierno o país, resalta la presencia de “Bogotá" en la lista, aunque es coherente con el hecho de que Petro ha sido alcalde de la capital colombiana y que su carrera política ha estado ligada a esa ciudad. "Salud" es otra palabra importante en el wordcloud, considerando que ha sido una de las principales reformas que ha impulsado dentro de su proyecto político y que en los últimos años hemos estado inmersos en el contexto del COVID-19. Igualmente, la frecuencia de la palabra "pacto" tiene sentido gracias a su pertenencia al partido del Pacto Histórico, así como la palabra "paz”, pues su proyecto de la paz total ha sido una de las banderas de su agenda política y se trata de comunicar que busca solucionar los conflictos armados que han afectado al país durante décadas. A grandes rasgos, se puede notar que su propuesta política se enfoca en la transformación social y económica del país.

**Gráfica 2 - Nube de palabras: Gustavo Petro  
Fuente: R Studio**

Finalmente, en el caso del ex-presidente Álvaro Uribe, las palabras con mayor protagonismo son: usd, onza, Colombia, familia, solidaridad, país, Medellín, tonelada, social, hoy, violencia, democracia y centro. Las palabras USD, onza y tonelada aparecen con gran frecuencia porque es común que el mandatario comparta en su cuenta de twitter información sobre actualidad macroeconómica y comercio internacional con indicadores con la tasa de cambio, el precio del petróleo, del café, entre otros. Por otra parte, destacan los conceptos de familia y solidaridad son propios del discurso conservador que promueve valores tradicionales, al igual que la palabra violencia, pues Uribe ha sido históricamente uno de los principales persecutores de grupos al margen de la ley como las FARC.

Adicionalmente, aparecen palabras como Medellín, teniendo en cuenta que la mayoría de sus simpatizantes son del departamento de Antioquia, su ciudad natal y donde tiene más fuerza su ideología política. Para terminar, en cuanto a las palabras relacionadas con la política, es importante notar que democracia y centro son términos asociados al partido Centro Democrático, del cual Uribe es líder. En general, el análisis de las palabras más frecuentes en los tweets del ex-presidente parece indicar una mayor preocupación por temas económicos y financieros, y menos atención a los problemas políticos y sociales del país. Esto es coherente con su ideología conservadora y su enfoque en mantener la estabilidad del sistema político y económico del país.

**Gráfica 3 - Nube de palabras: Álvaro Uribe   
Fuente: R Studio**

En conclusión, se evidencia que en el caso de Claudia López, se puede ver una gran frecuencia de palabras relacionadas con Bogotá, lo que sugiere que la alcaldesa se enfoca en asuntos de su ciudad. Para Álvaro Uribe, se observan palabras como "violencia" y "democracia", lo que sugiere que su enfoque político está en temas relacionados con la seguridad y la democracia. En el caso de Gustavo Petro, se observan palabras como "paz" y "salud", lo que sugiere que se enfoca en temas sociales y de justicia.

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>