**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.

**Figura 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: la base test y la base train.

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

El precio de una vivienda puede estar determinado por diversos factores tanto económicos como sociales. Para el caso de estudio, se analizan factores internos (propios de las viviendas) como externos (elementos geoespaciales como distancia a hospitales, parques y colegios, etc.). En este sentido, es necesario tener información de variables asociadas al precio, con el propósito de obtener un modelo robusto. Para el análisis de los modelos de predicción se utiliza una muestra tomada de Properati sobre los precios de venta y características de inmuebles ubicados en la localidad de Chapinero en Bogotá, lo que arroja una muestra representativa de entrenamiento y testeo del precio de venta de las viviendas en esta zona, así como de los principales atributos que definen el precio de mercado de esta.

**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. 2.** Nube de palabras: Álvaro Uribe



**Fuente:** R Studio

**Grafica No. 3.** Nube de palabras: Claudia López



**Fuente:** R Studio

**Grafica No. 4** 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>