Problem Set 4: Predicting tweets?

Sofia Charry Tobar
s.charry@uniandes.edu.co

Laura Manuela Rodriguez Morales
lm.rodriguezm@uniandes.edu.co

Nicol Valeria Rodríguez Rodríguez
nv.rodriguezr1@uniandes.edu.co

Brahyan Alexander Vargas Rojas
ba.vargas@uniandes.edu.co

El repositorio del ejercicio es: https://github.com/Laura5513/Taller-4-BDML

1. Introducción

Actualmente, la cantidad de datos e información obtenidos de las redes sociales sigue en constante aumento, lo que permite generar una variedad de patrones de datos útiles para distintos tipos de investigación, tales como el comportamiento social humano, la seguridad del sistema, la criminología, entre otros. En este trabajo se busca estimar un modelo confiable y preciso que permita predecir los tweets de políticos en Colombia, con el fin de analizar y comprender sus posiciones, tendencias, resultados y sentimientos. Con el fin de llevar a cabo este estudio, se creó una base de datos que incluye tweets de tres políticos destacados en Colombia: la actual alcaldesa de la ciudad de Bogotá (Claudia López), el actual presidente de Colombia (Gustavo Petro) y un ex-presidente de Colombia influyente en la escena política (Álvaro Uribe). Para iniciar el análisis, se utilizó un conjunto de datos sin etiquetar, el cual fue transformado para proponer diferentes modelos de predicción.

Se propone seguir la estrategia descrita en la literatura (Riofrio et al., 2022; Arcila-Calderón et al., 2017) para llevar a cabo una limpieza detallada y precisa de las palabras más relevantes y frecuentes en los tweets publicados por políticos. Dado que los políticos abordan una amplia variedad de temas, se identifican estas palabras clave en cada publicación que realizan. Posteriormente, toda la información recopilada se agrupa en una base de datos que permitirá predecir los tweets de los políticos. Es importante tener en cuenta que la

fecha de la información con la que se cuenta no se especifica, por lo que es posible que las publicaciones estén vinculadas a eventos específicos.

Después de haber definido la base de datos, se procedió a evaluar varios modelos y finalmente se seleccionó el modelo Lasso debido a su alta precisión en el análisis. En conclusión, las palabras clave identificadas en la base de datos resultan de gran utilidad para predecir el contenido de los tweets publicados por políticos, lo que permite comprender de manera más precisa sus opiniones y descubrir sus verdaderas intenciones.

2. Data

2.1. Base de datos

La base de datos combina información de tweets publicados por las cuentas de tres políticos colombianos: Gustavo Petro, Claudia López y Álvaro Uribe. Se realizó un pre-procesamiento de los tweets para realizar el análisis de lenguaje natural. Luego de este proceso, se apartó un 30 % para realizar la validación del modelo (2.804 observaciones) y un 70 % para para realizar el entrenamiento del modelo (6.545 observaciones) de la base de entrenamiento proveída por el equipo pedagógico. La base de testeo es la que nos interesa predecir y tiene 1.500 observaciones.

La base de datos es útil para predecir la cuenta que publicó cada tweet puesto que los predictores dan cuenta de las palabras en los tweets y cada político se caracterizará por utilizar con mayor frecuencia ciertas palabras. Específicamente, los predictores premian la frecuencia de palabras poco repetidas en el texto pues se utilizó el método TD-IDF en el preprocesamiento, como se verá a continuación, por lo que las palabras que caractericen más las publicaciones de cada político resaltan más relevantes, por encima de palabras que comúnmente utilicen en común todos los 3 políticos. Sin embargo, como se trabaja con lenguaje natural y en el procesamiento se pueden perder términos de palabras conjuntas, es posible que algunas palabras no tengan mucho sentido por si solas o que se pierda el contexto del mensaje del tweet por este problema. De todas formas, este es un problema que siempre se tiene al utilizar textos como datos.

Con el objetivo de clasificar los tweets a partir de variables numéricas, el procesamiento de la base de datos se realizo en 4 etapas:

1. **Limpieza**: En esta primera etapa, se elimina los caracteres, números o componentes especiales de los tweets que no son relevantes en la tarea de clasificación. De esta manera, se eliminan:
    - Tildes y caracteres especiales del español.

- URLs.
- @ (@usuario).
- Espacios en blanco adicionales.
- Emojis.
- Puntuación.
- Números.
- Stopwords.

2. **Lematizar (Stemmizar)**: En esta segunda etapa, se extraen las raíces de cada palabra, lo cual permite (i) simplificar el análisis al reducir las diferentes versiones de una misma palabra a su raíz y (ii) refinar el análisis de frecuencia de la siguiente etapa.

3. **Frecuencia de palabras**: En tercer lugar, los datos se transforman a una matriz TF-IDF (Term Frequency-Inverse Document Frequency), con el objetivo de obtener la frecuencia con que aparecen las palabras en los tweets y, por lo tanto, su importancia. Lo anterior permite la obtención de variables numéricas a partir de los caracteres.

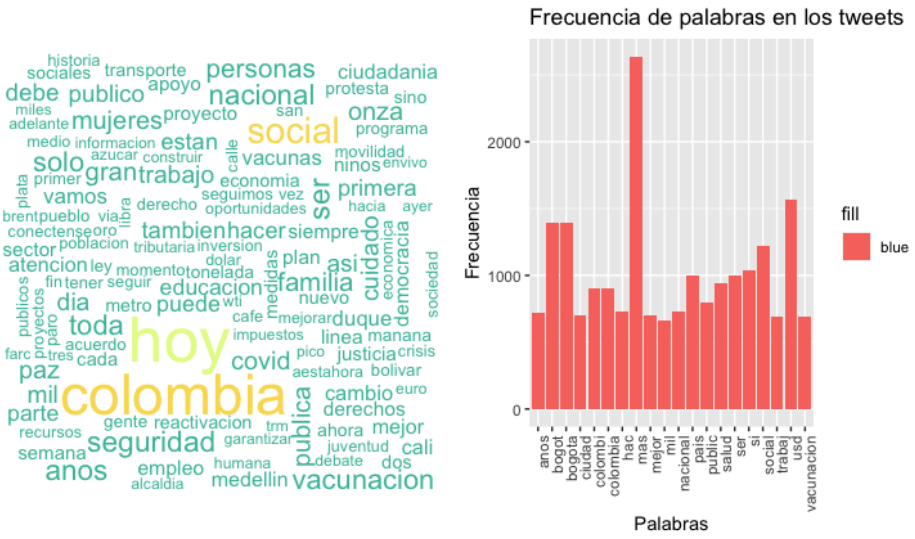
4. **Filtracion de variables menos frecuentes**: Finalmente, se remueven las palabras de acuerdo con la presencia de ceros en su frecuencia. En particular, se remueven las palabras que son más *sparse* a 0,99. Este paso es especialmente importante, debido a que entre más se eliminen variables con mayor presencia de ceros, se afecta significativamente el desempeño de los modelos.

Este proceso se aplica tanto en la base de entrenamiento como de testeo.

2.2. Análisis descriptivo

Una vez realizada la separación adecuada de las palabras, procedimos a analizar su comportamiento en el gráfico 1. Esta técnica nos permite visualizar las palabras más frecuentes en un conjunto de textos, presentándolas en un tamaño proporcional a su frecuencia. Esto nos facilita la identificación de las palabras más comunes. En el primer gráfico (a), se muestran las palabras más utilizadas en las publicaciones de los políticos. Entre ellas se encuentran Çolombiaz "social", lo que indica que estos políticos hablan con frecuencia sobre el país y su parte social. Por otro lado, en el gráfico (b) se observa la frecuencia de las palabras, entre las que destacan çomponente social y económico". Los políticos también hablan sobre el país en general y las principales ciudades del mismo. Además, se puede evidenciar que los tweets hacen referencia a la época de la pandemia, ya que una de las palabras más recurrentes es "vacunaciónz "salud"

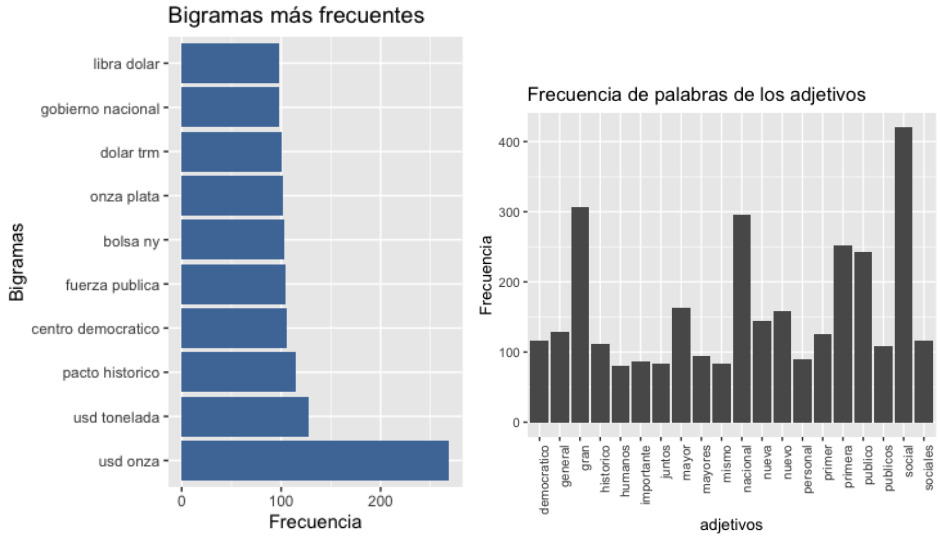
Gráfico 1



(a) (b)

En el Gráfico 2 se presentan dos gráficos relevantes. En el primer gráfico (a), se muestran los biogramas más comunes de los tweets seleccionados, lo cual revela que los políticos mencionan frecuentemente la divisa dólar y una unidad de peso. Además, se puede observar que se refieren a los distintos partidos políticos, lo que sugiere que existe una importante diferencia ideológica en constante debate. Por otra parte, en el segundo gráfico (b), se presenta un análisis de sentimiento basado en los adjetivos utilizados en los tweets. Este análisis es esencial para determinar la actitud o el sentimiento de los gobernantes hacia temas específicos. El gráfico de barras muestra los adjetivos más utilizados por los políticos, entre los que se destacan "social", "nacionalz "gran", los cuales se repiten con mayor frecuencia.

Gráfico 2



(a) (b)

Después de realizar el pertinente análisis, se creó una base de datos para desarrollar un modelo de predicción de tweets. En este caso, se analizaron alrededor de 500 tweets, desglosados en palabras clave para evaluar las características de los políticos, lo que resultó en aproximadamente 9349 observaciones. Se obtuvieron alrededor de 2258 variables que muestran la frecuencia de la palabra utilizada en el tweet del político, y una variable adicional que indica si se trata de Üribe"(1), "López"(2) o "Petro"(3). En la Tabla 1 se presentan algunas estadísticas descriptivas de las variables utilizadas en el estudio. Por ejemplo, la variable .ªlcaldía"tiene una media de 0.0048, lo que indica que esta palabra aparece el 0.0048 de las veces. En cambio, el máximo de la variable "fiscalía.^{es} de 3.3493, lo que sugiere que esta palabra se utilizó con más frecuencia en algunos tweets en específico. Además, el mínimo de la variable "diferencia.^{es} 0, lo que significa que esta palabra no se usó en algunas ocasiones. Aunque las otras variables del modelo no se distribuyen de la misma manera, su interpretación es similar.

Tabla 1: Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
name	9349	1.9866	0.7929	1	3
alcaldia	9349	0.0048	0.0.0646	0	3.3493
diferencia	9349	0.0015	0.0386	0	2.1995
fiscalia	9349	0.0020	0.0.0333	0	1.2940

3. Modelo y resultados

3.1. Variables utilizadas

La variable Y a predecir es *name*, variable categórica que toma tres diferentes valores: Lopez, Petro y Uribe, para así clasificar qué político/a colombiano/a pertenece cada tweet. Lo anterior se logra a partir de 2259 indicadores sobre la frecuencia de las palabras en los tweets, los cuales conforman el conjunto de variables explicativas X.

3.2. Entrenamiento del modelo

La evaluación del modelo se realizó de la siguiente manera:

1. La muestra de entrenamiento **train** se dividió en dos conjuntos adicionales, de tal manera que el 70 % se utilizó en el entrenamiento del modelo, y el 30 % restante en su evaluación. Este procedimiento se debe a que la base de **test** no contiene la variable de *name*.

2. Una vez se entrenaron los modelos y se verificó que no hubiese problemas de *overfitting*, se clasificó a qué político/a colombiano/a pertecene cada tweet a partir de la base de **test**, para luego introducir los resultados en Kaggle y obtener la medida de evaluación *Accuracy*.

El modelo que evidencia tener mejor desempeño es un Lasso. Con respecto a la selección de los hiperparámetros, el modelo se entrenó a partir de los siguientes valores: (i) ***alpha***: 1 y (ii) ***lambda***: seq(0.001,1,by = 0.001). De esta manera, el mejor modelo Lasso seleccionado de acuerdo con los hiperparámetros es lambda 0.004. Por otro lado, de acuerdo con la Tabla 2, las variables que resultan más útiles en la clasificación de variables son: (i) En el caso de Uribe, "farc", "gbno", "cafe, "pte", (ii) para Petro, "paronacionalj", "subterraneo", "hidrotuango", "cuerpos" y (iii) para Lopez, "vial", "cuidado", "bogota", "concentracion". Como es de esperarse, las palabras más importantes de cada tweetero están relacionadas con su agenda política.

Tabla 2: 40 palabras más importantes para cada tuitero según Lasso

Lopez	Importancia	Petro	Importancia	Uribe	Importancia
vial	94,143	paronacionalj	76,582	farc	100
cuidado	87,625	subterraneo	74,784	gbno	92,049
bogota	86,047	hidroituango	74,247	cafe	74,401
bogotanos	79,463	cuerpos	71,331	pte	62,621
concertacion	77,087	centenares	71,302	azucar	62,187
jovenesalau	75,960	humana	69,159	libra	58,415
bici	74,799	tarifas	64,737	informa	57,167
equipobogota	71,259	desastre	60,363	duele	54,848
localidad	68,570	desigualdad	55,732	autoridad	54,201
bogotaregion	65,827	primaria	54,636	bloqueos	50,957
abrazo	62,823	pacto	53,430	borrador	50,094
agradecemos	62,477	era	52,409	ffaa	44,454
recomendamos	61,557	presidencial	48,315	afectar	44,327
aestahora	61,122	coalicion	48,189	fallecimiento	43,305
pico	60,255	duque	47,873	fallecio	42,657
juntos	59,358	importaciones	47,124	metro	40,135
rescate	58,886	progresista	46,526	habana	39,281
pmu	58,213	pequena	46,216	menor	38,543
atendiendo	57,437	neoliberal	45,917	garantias	38,381
humildes	57,299	neoliberalismo	44,415	violenta	37,722
localidades	57,107	pensiones	44,361	agradezco	37,711
empezo	56,602	debe	43,433	luego	37,674
cuidamos	56,598	europea	43,362	bogota	36,795
logramos	56,363	espero	43,150	todas	36,553
empezar	56,058	quitar	43,055	democratico	36,422
tendra	56,057	latinoamericana	42,838	comprar	36,331
ninas	55,945	solicito	42,457	formal	36,191
mejora	55,804	capturados	42,295	insumos	36,099
compromisos	55,307	paramilitares	42,209	convencionescentrodemocrat	36,084
homenaje	54,327	implica	41,549	red	34,818
potelrenacerdebogota	54,244	eeuu	41,351	tema	34,489
kms	53,751	criminales	41,327	cuba	33,282
ciudad	52,775	pacifico	41,280	epm	33,162
conjunto	52,670	petroleo	41,267	seran	32,921
ecologica	51,718	selva	40,805	foro	32,731
movilidadsosten	51,523	paramilitarismo	40,606	disidencias	32,660
feria	51,114	llaman	40,275	dictadura	32,633
istrital	50,207	sustancialmente	39,981	solo	32,177
tapabocas	49,652	gracias	39,803	soldados	32,131
felicidad	49,303	muertos	39,373	ivan	31,848

3.3. Comparación

Antes de hacer la elección del modelo final también se corrieron otras técnicas para realizar la clasificación. El objetivo es clasificar los tweets de acuerdo con su dueño/a, que está dado por la variable *name*, a través de un gran conjunto de variables sobre la frecuencia de las palabras. Teniendo en cuenta lo anterior se estimaron las siguientes técnicas: Ridge, Elastic net, Gradient Boosting (GBM), PCA y Random Forest (RF). Se utiliza la técnica de remuestreo *cross-validation* con k=10, con el objetivo de tener una mejor medida del desempeño de los modelos. En el caso de los modelos Elastic net, GBM y RF, se realizaron transformaciones adicionales a la base de datos para facilitar la corrida de las herramientas, de tal manera que se redujera la dimensionalidad de la base de datos y se garantizara que *train* y *test* tuvieran las mismas variables. Junto a lo anterior, los hiperparámetros de los

mejores modelos y sus resultados son:

```
# Lasso: alpha = 1, lambda = 0.004
# Ridge: alpha = 0 and lambda = 0.03536296
# Elastic Net: alpha = 0.94, lambda = 0.0051
# Lasso + PCA: alpha = 1, lambda = 0.003
# GBM: n.trees = 300, interaction.depth = 3, shrinkage = 1.0
# RF: n.trees = 300, bootstrap = True, verbose = 2, max_features = 'sqrt'
(se considera la raíz cuadrada del total de características)
```

Tabla 3: Comparación del modelo principal con otras técnicas

Estadística	**Lasso**	**Ridge**	**EN**	**Lasso + PCA**	**RF**	**GBM**
Accuracy (Kaggle)	0.77333	0.34333	0.6	0.28666	0.72666	0.74
Accuracy	0.773	0.371	0.6362	0.3752	0.7378	0.7375
Balanced Accuracy	0.833	0.3	0.7233	0.3713	0.7322	0.7309
Kappa	0.658	0	0.4518	0.0605	0.6043	0.6046

Como se mencionó anteriormente, el modelo Lasso exhibe los mejores resultados en las diferentes métricas de evaluación presentadas en el Cuadro 3. De igual forma, los modelos RF y GBM son las dos siguientes mejores herramientas en la clasificación de tweets. Además, tanto el modelo Ridge como el modelo Lasso + PCA presentan malos resultados y cuentan con una amplia diferencia frente a las demás técnicas implementadas. Notamos que la reducción de dimensionalidad proporcionada por el modelo PCA no logra explicar la variabilidad de la base de datos, lo cual podría ser clave para explicar su mal resultado.

4. Conclusiones y recomendaciones

La base de datos proporcionada combina información de los tweets de tres políticos colombianos, que se utilizaron para entrenar y validar un modelo de clasificación de lenguaje natural. El preprocesamiento de los tweets incluyó la eliminación de caracteres, emojis, URL y stop words, la lematización (stemming) y la creación de una matriz TF-IDF para obtener la frecuencia con la que aparecen las palabras en los tweets. La eliminación de palabras poco frecuentes fue crucial para mejorar el rendimiento del modelo. El análisis descriptivo reveló que los políticos hablan frecuentemente sobre la situación social y económica del país, así como de la pandemia y la vacunación.

En conclusión, el modelo Lasso resultante es útil para predecir la cuenta que publicó cada tweet, ya que cada político tiene palabras que utiliza con mayor frecuencia. Sin embargo, es importante considerar que algunas palabras pueden no tener sentido por sí solas o que el

contexto del mensaje en Twitter puede perderse debido a la naturaleza del lenguaje natural. Además, es importante destacar que la base de datos se limitó a los tweets de solo tres políticos colombianos, lo que podría limitar su aplicabilidad a otros contextos.

En cuanto a las variables más útiles para clasificar las publicaciones de cada tweetero, las palabras "farc", "gbno", "cafe" y "pte" resultaron ser las más relevantes para Uribe, mientras que "paronacionalj", "subterraneo", "hidrotuango" y "cuerpos" fueron las más relevantes para Petro, y "vial", "cuidado", "bogota" y "concentracion" para López. Estos resultados sugieren que las palabras más importantes para cada político están relacionadas con sus agendas políticas, lo que refleja sus principales preocupaciones y temas de interés. En resumen, este estudio demuestra la importancia de la selección de modelos adecuados y la optimización de los hiperparámetros para lograr un mejor desempeño en la predicción y clasificación de datos. Además, la comparación de diferentes técnicas de clasificación puede ayudar a seleccionar la mejor opción para un problema particular.

Referencias

Arcila-Calderón, C. et al. (2017). "Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático". En: *Profesional de la Información* 26.5, págs. 973-982. URL: https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/epi.2017.sep.18.

Riofrio, C. E. et al. (2022). "Identificación de ideología política mediante un modelo Transformer para estilometría y Clasificación por votos en Machine Learning". En: *Polo del Conocimiento* 7.9, págs. 1457-1474. URL: https://www.polodelconocimiento.com/ojs/index.php/es/article/view/4642.