Taller 2 - Big Data

Diplomado de Análisis y Modelamiento de datos - Machine Learning José Miguel Tobar - jtobarr@utem.cl

**Generación del Dataset**

* 1. **Detección del Problema**

El día 01 de junio del 2022 el presidente de la república Gabriel Boric presentó su primera cuenta pública, donde realizó un balance de su mandato y presentó las principales propuestas que impulsará durante los próximos cuatro años.

La tasa de desaprobación de Boric subió al 57% en la última encuesta semanal Cadem publicada a última hora del domingo, el nivel más alto desde que asumió el cargo en marzo. Debido a esto se hace necesario ver la percepción del publico en general luego del discurso en la cual realiza la cuenta pública. Esta opinión podría reflejar a corto plazo el aumento o disminución de esta cifra.

* 1. **Generar base de datos**

La base de datos se realizó con un conjunto de datos obtenidos desde TWITTER, en la cual se utilizo el tag **cuenta pública.** La cantidad de datos corresponde a 203 opiniones emitidas en esta red social.

**RESULTADOS**

**3.1.- ¿Qué problema (de negocio) resuelve el modelo de clasificación en sus datos?**

El problema que llega a resolver este modelo es en el área política, específicamente lo que respecta a la opinión pública. Conocer la opinión publica de los oyentes permite tomar mejores decisiones futuras y determinar los puntos clave para la ciudadanía, de esta misma manera esta información permitirá tomar las medidas necesarias para detener la cifra de desaprobación.

**3.2.- ¿Cuáles son las categorías de clasificación? ¿Cuántos ejemplos hay de cada categoría?**

En nuestro modelo existen 2 categorías: Aprobación y Desaprobación esto enfocado en la opinión que tiene cada oyente de la cuenta publica emitida por el presidente.

Aprobación: 104 Registros

Desaprobación: 99 Registros

**3.3.- Muestre los resultados del entrenamiento de su modelo (pestaña Stats), tanto para el overall como para cada una de las categorías. (pantallazo)**

Se adjunta primer pantallazo correspondiente a los datos contenidos en el overall, donde muestra la cantidad total de datos de entrenamiento que son 203, un Accuracy de 81% y un F1 SCORE de 81%

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

En el segundo pantallazo podemos ver la categoría correspondiente a Aprobación con un total de 104 datos, 78% de precisión y 88% de recall.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

El tercer y ultimo pantallazo contiene los datos asociados a la categoría Desaprobación. Con un total de 99 datos con 85% de precisión y 74% de recall.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

3.4.- ¿Qué significa Accuracy? ¿Qué significa F1 Score?

Las dos métricas mencionadas en este apartado corresponden a evaluaciones de desempeño de nuestro modelo. Específicamente el Accuracy nos entrega el porcentaje de casos que el modelo ha acertado, para nuestro caso es del 81%, este valor podría posiblemente aumentar si incrementamos la cantidad de datos ingresados. En lo que respecta al F1 Score es una combinación entre la métrica Accuracy y Recall dando un único valor. Esta métrica es la más apropiada cuando tenemos conjuntos de datos no balanceados. Se calcula como la media armónica de Accuracy y Recall. Su fórmula es:

F1 = (2 \* Accuracy \* Recall) / (Accuracy + Recall)

3.5.- Escriba un ejemplo de texto por cada una de las categorías:

|  |  |
| --- | --- |
| Categoría | Ejemplo |
| Aprobación | Cuando vote a Boric como presidente esperaba escuchar una persona autentica que se preocupara por los que nunca fuimos escuchados, grande presidente! |
| Desaprobación | No sé si Boric sabe que es una cuenta pública, porque lo que hizo fue una payasada, maldito burro que tenemos de presidente |

Se adjunta el código en lenguaje PYTHON para realizar pruebas del modelo.

from monkeylearn import MonkeyLearn

ml = MonkeyLearn('f5630ed1d81817bf3d1f35cb5b2a9404d4feedb2')

data = ["La primera cuenta pública que escucho y disfruto completa, escucho al presidente"]

model\_id = 'cl\_ZqvBUTmF'

result = ml.classifiers.classify(model\_id, data)

print(result.body)