

ISIS3301 - Inteligencia de Negocios

Juan Nicolas García - 201717860

Informe Proyecto Analítica de textos – Etapa 1

Introducción

Partiendo de una necesidad concreta de clasificación supervisada, este documento describe el desarrollo de un sistema de analítica de textos que avanza de forma ordenada desde el perfilamiento y la preparación de los datos hasta el entrenamiento de modelos y la rotulación del archivo de prueba. En primer lugar, aplicamos pasos fundamentales de NLP como la normalización, la lematización con el modelo de español de spaCy, la depuración de ruido y de stopwords y, además, la extracción de características mediante TF-IDF. En este marco, el problema se orienta a comprender y organizar opiniones vinculadas con prioridades públicas alineadas a los Objetivos de Desarrollo Sostenible de la ONU, por ejemplo, pobreza, salud y educación, de modo que el sistema pueda apoyar el seguimiento temático y la toma de decisiones basada en evidencia. Por último, evaluamos alternativas de modelado que incluyen Regresión Logística, LinearSVC y un Random Forest entrenado sobre representaciones reducidas con Truncated SVD.

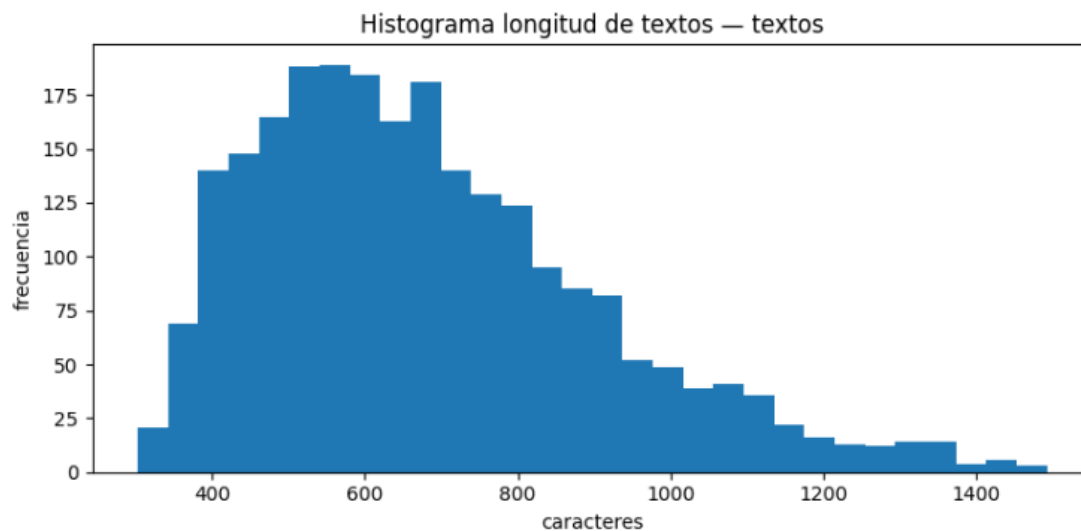
Documentación del proceso de aprendizaje automatico



Entendimiento y preparación de los datos

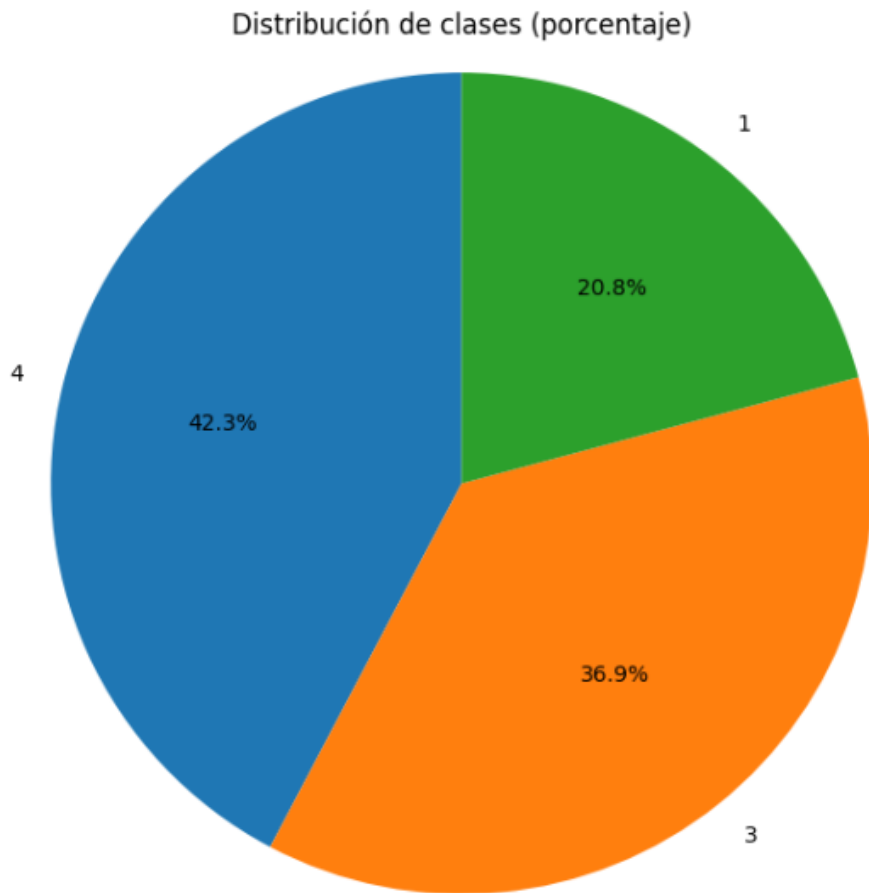
	columna	tipo	nulos	%_nulos	unicos	ejemplo
0	textos	object	0	0.0	2424	"Aprendizaje" y "educación" se consideran sinó...
1	labels	int64	0	0.0	3	4
	filas	columnas	duplicados	%_duplicados		
0	2424	2	0	0.0		
	min	p25	p50	p75	max	media
0	303	513.0	647.0	809.25	1492	683.223185

En primer lugar, se realizó el perfilamiento de los datos con el objetivo de identificar posibles problemas de calidad. La tabla evidencia que no se presentan valores nulos ni duplicados y que el número de observaciones se mantiene constante. Además, se aprecia la distribución de los tipos de datos y ejemplos representativos de cada columna, lo que facilitó la validación inicial de la estructura del dataset.

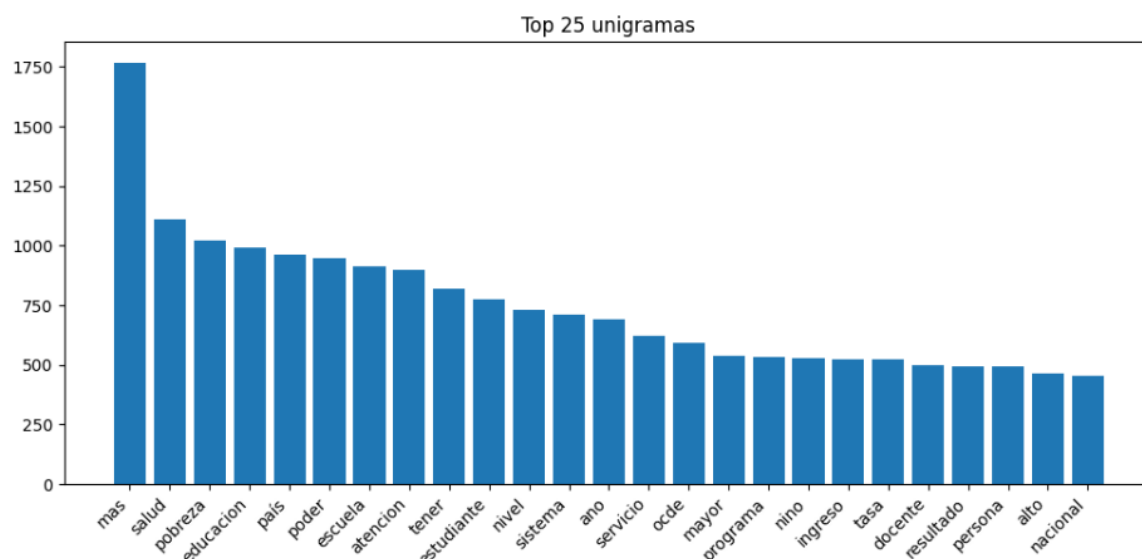


Luego, se evaluó la longitud de los textos, lo que permitió conocer el rango y la distribución de caracteres en el corpus. El histograma muestra que la mayoría de los textos se concentran en una longitud intermedia, aunque también existen registros con mayor extensión. Este análisis es relevante ya que ayuda a anticipar posibles sesgos asociados con textos excesivamente cortos o largos, los cuales pueden afectar la capacidad del modelo para extraer información útil.

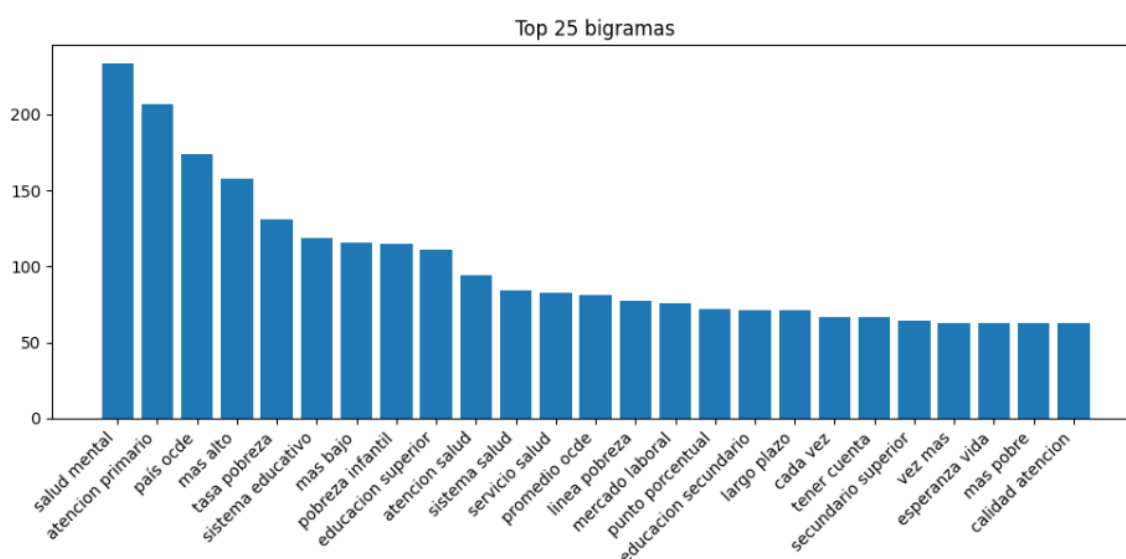
clase	conteo	%
0	4	1025
1	3	894
2	1	505



Asimismo, se examinó la distribución de las clases en el dataset. El gráfico circular permite observar la proporción relativa entre las categorías, lo cual es fundamental para reconocer posibles desbalances. En este caso, si bien las tres clases tienen representación, existe una clase predominante que ocupa más de cuarenta por ciento de los registros. Esta situación requiere atención, ya que un desbalance marcado puede afectar el desempeño del modelo, haciendo necesario recurrir a técnicas de ajuste como el uso de pesos balanceados.



Como parte del proceso de limpieza se aplicaron técnicas de normalización, eliminación de caracteres especiales, stopwords y lematización en español. A partir de estos textos depurados se extrajeron unigramas, los cuales permiten identificar las palabras más frecuentes en el corpus. La figura muestra que términos como 'más', 'salud', 'pobreza' y 'educación' aparecen con gran relevancia, lo que anticipa la centralidad de estas temáticas en el conjunto de datos.



Además, se analizaron los bigramas más comunes, lo que permite observar combinaciones de palabras con mayor valor semántico. Entre ellos destacan 'salud mental', 'atención primaria' y 'país OCDE', indicando relaciones clave dentro de las opiniones ciudadanas. Este análisis facilita la identificación de expresiones recurrentes que aportan contexto adicional a los temas discutidos.

como ocurre en el procesamiento de texto. Su estrategia de maximizar el margen entre clases le otorga una gran capacidad de generalización, incluso cuando el dataset no está perfectamente balanceado. Además, con kernels adecuados puede capturar relaciones no lineales, ofreciendo un desempeño competitivo frente a otros enfoques.