ISIS3301 - Inteligencia de Negocios

Juan Nicolas García - 201717860

Informe Proyecto Analítica de textos – Etapa 1

Sección 0. 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.

Sección 1. Documentación del proceso de aprendizaje automatico

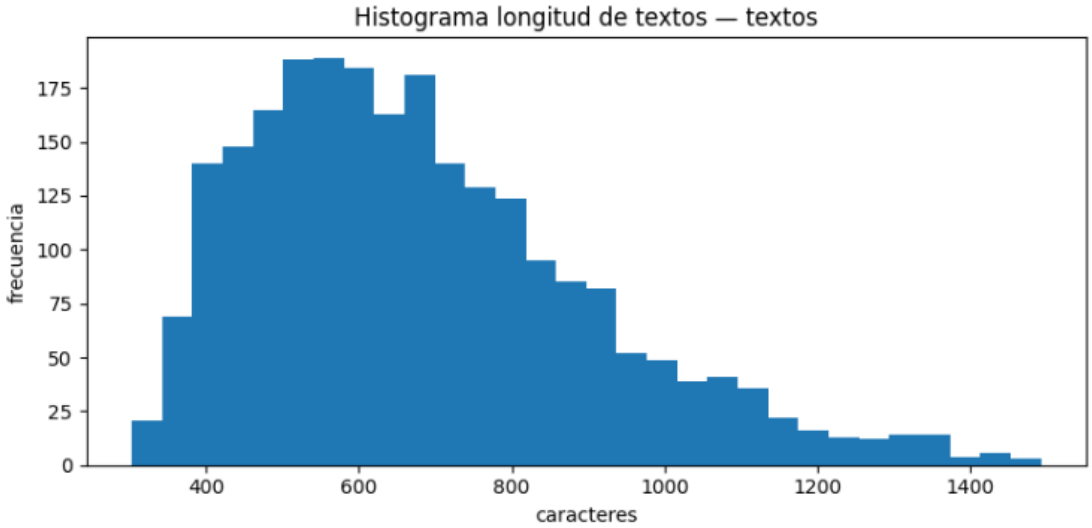
Sección 2. Entendimiento y preparación de los datos

	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

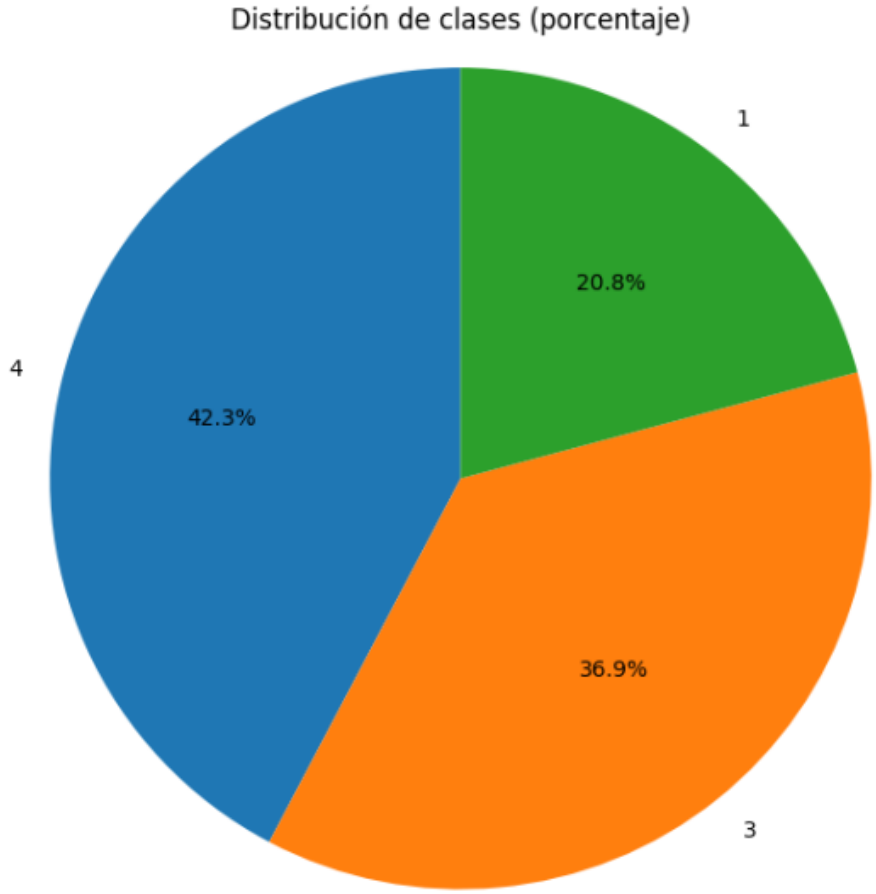
	min	p25	p50	p75	max	media
0	303	513.0	647.0	809.25	1492	683.223185

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.

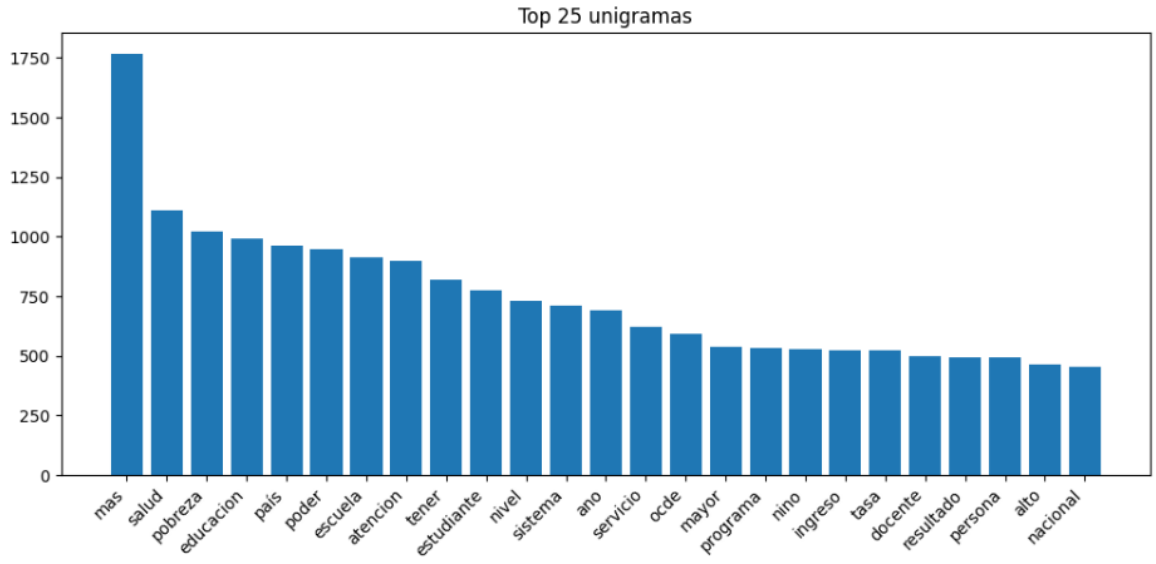


Luego, se evaluó la longitud de los textos, lo que permitió conocer el rango y la distribución de caracteres en el corpus. El histograma muestra que la mayoría de los textos se concentran en una longitud intermedia, aunque también existen registros con mayor extensión. Este análisis es relevante ya que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

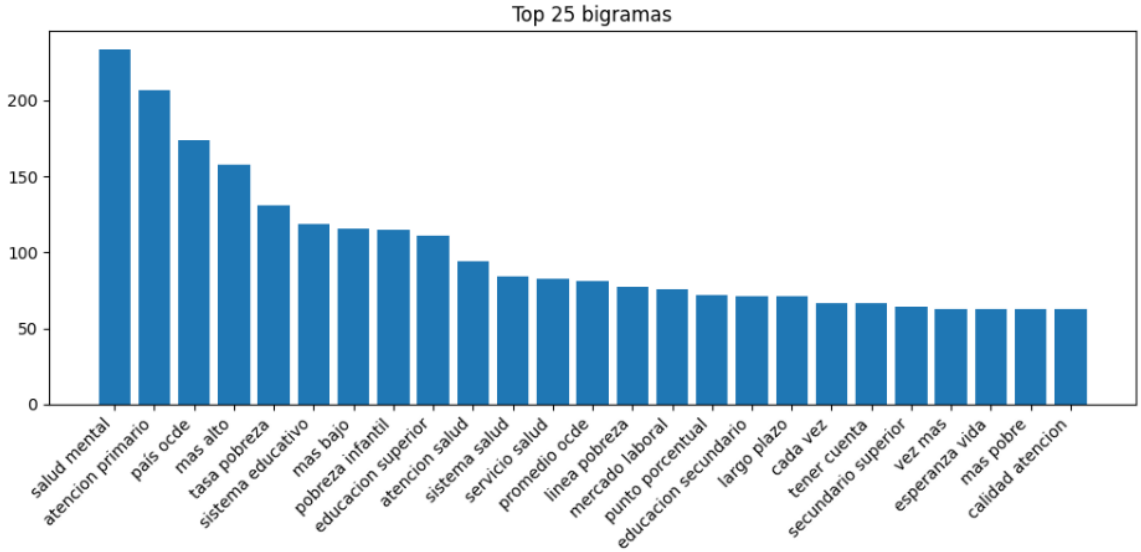
	clase	conteo	%
0	4	1025	42.285479
1	3	894	36.881188
2	1	505	20.833333



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.

Por último, se generó una nube de palabras que representa de manera visual la importancia relativa de los términos dentro del corpus. Palabras como 'escuela', 'poder', 'programa' y 'estudiante' destacan por su tamaño, lo cual refuerza los hallazgos obtenidos en los análisis de unigramas y bigramas. Esta visualización resulta útil para una interpretación rápida de las tendencias del dataset.

En conjunto, la limpieza y el preprocesamiento no solo permitieron mejorar la calidad del texto, sino también extraer patrones semánticos clave. Estos pasos resultan fundamentales para garantizar que las transformaciones posteriores, como la vectorización y el modelado, se basen en datos coherentes y representativos del problema planteado.

Sección 3. Modelado y evaluación

Regresión logística

Se implementó por su rapidez y claridad interpretativa en tareas de clasificación. Este algoritmo resulta eficaz cuando las clases pueden separarse aproximadamente de manera lineal, y además ofrece probabilidades que facilitan comprender la confianza de cada predicción. Gracias a su bajo costo computacional, constituye un punto de partida sólido para el problema.

Random Forest

El Random Forest se utilizó como una alternativa robusta frente a patrones más complejos y relaciones no lineales. En este caso, se aplicó primero una reducción de dimensionalidad con Truncated SVD sobre la matriz TF-IDF, lo que permitió simplificar el espacio de características antes de entrenar el bosque de árboles. Con ello se logró reducir el sobreajuste y mantener un buen balance entre precisión y generalización.

SVM (LinearSVC)

El modelo SVM fue elegido por su capacidad de trabajar en espacios de alta dimensionalidad,

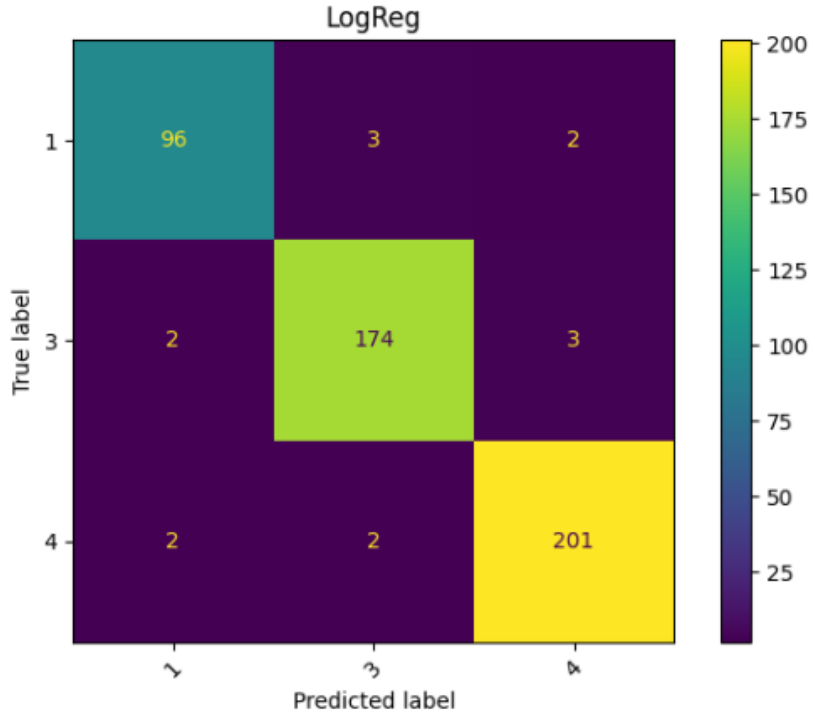
como ocurre en el procesamiento de texto. Su estrategia de maximizar el margen entre clases le otorga una gran capacidad de generalización, incluso cuando el dataset no está perfectamente balanceado. Además, con kernels adecuados puede capturar relaciones no lineales, ofreciendo un desempeño competitivo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

Sección 4. Resultados

En esta sección se presentan los resultados obtenidos tras la aplicación de los tres modelos evaluados (Regresión Logística, LinearSVC y Random Forest con SVD). El análisis se enfoca en las métricas de desempeño y en cómo estas contribuyen a los objetivos del proyecto, enmarcado en la identificación de temáticas relacionadas con los Objetivos de Desarrollo Sostenible (ODS).

[Logistic Regression]

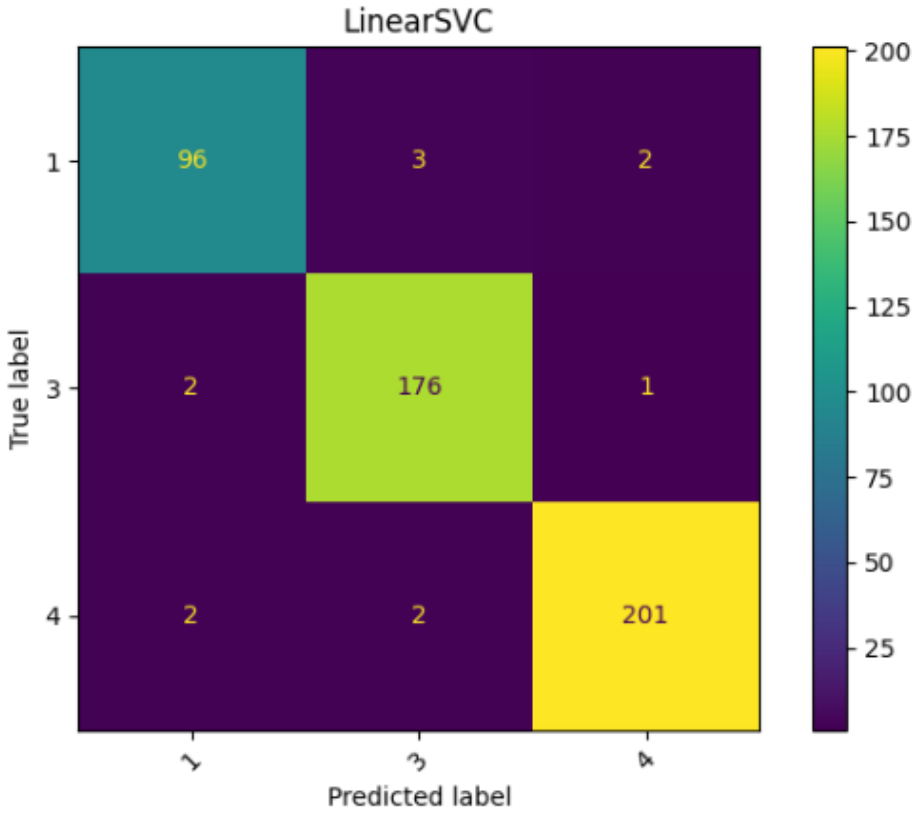
	precision	recall	f1-score	support
1	0.96	0.95	0.96	101
3	0.97	0.97	0.97	179
4	0.98	0.98	0.98	205
accuracy			0.97	485
macro avg	0.97	0.97	0.97	485
weighted avg	0.97	0.97	0.97	485



La regresión logística alcanzó un accuracy del 97%, con valores de precisión, recall y F1-score uniformes en las diferentes clases. Esto demuestra su capacidad de interpretar relaciones lineales en los datos y mantener un equilibrio en la clasificación. El modelo se muestra sólido para identificar patrones generales de las opiniones.

[LinearSVC]

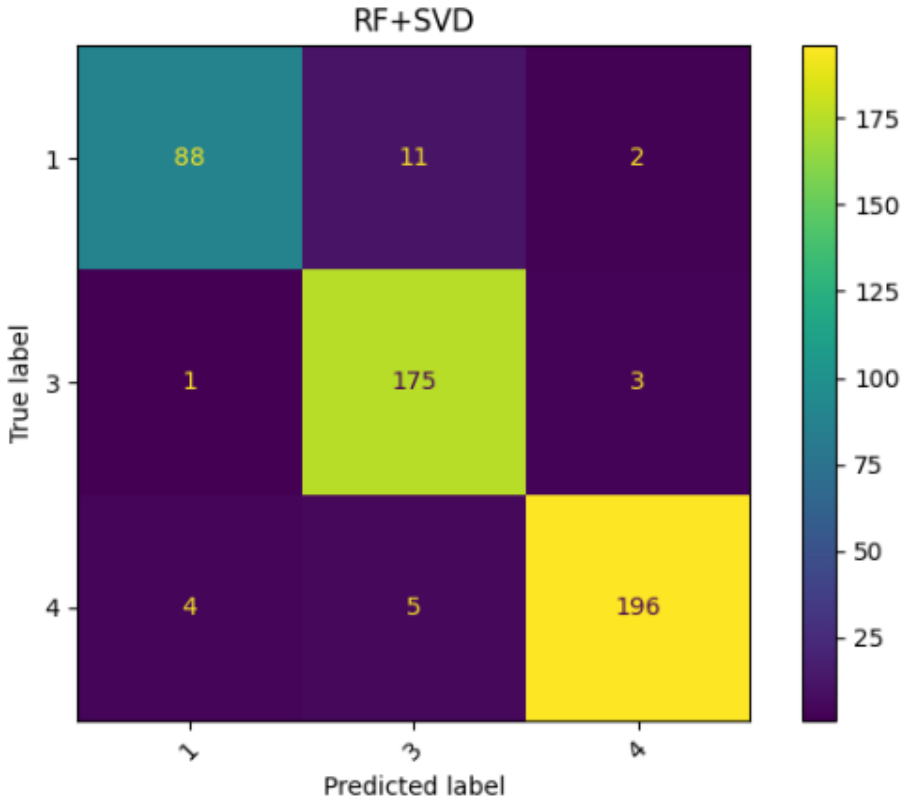
	precision	recall	f1-score	support
1	0.96	0.95	0.96	101
3	0.97	0.98	0.98	179
4	0.99	0.98	0.98	205
accuracy			0.98	485
macro avg	0.97	0.97	0.97	485
weighted avg	0.98	0.98	0.98	485



El modelo LinearSVC obtuvo la mayor precisión con un accuracy cercano al 98%. Sus métricas muestran un excelente equilibrio en la clasificación de todas las clases, lo que refuerza su idoneidad para trabajar con representaciones de texto en alta dimensionalidad. Este resultado sugiere que LinearSVC es la alternativa más confiable para garantizar predicciones consistentes en el contexto de este proyecto.

[RandomForest + SVD]

	precision	recall	f1-score	support
1	0.95	0.87	0.91	101
3	0.92	0.98	0.95	179
4	0.98	0.96	0.97	205
accuracy			0.95	485
macro avg	0.95	0.94	0.94	485
weighted avg	0.95	0.95	0.95	485



El Random Forest con reducción de dimensionalidad mediante SVD alcanzó un accuracy del 95%. Aunque su rendimiento fue ligeramente inferior en comparación con los otros modelos, mostró una alta capacidad para clasificar de forma correcta las clases más representadas. No obstante, se observaron dificultades para capturar con la misma precisión las clases minoritarias, lo que explica su menor desempeño global.

	modelo	accuracy	f1_weighted	f1_macro
0	LinearSVC	0.975258	0.975240	0.971962
1	LogReg	0.971134	0.971110	0.968464
2	RandomForest_SVD	0.946392	0.946153	0.939560

La comparación final de métricas evidencia que LinearSVC es el modelo con mejor rendimiento, seguido de cerca por la regresión logística. El Random Forest, pese a ser robusto, no alcanzó el mismo nivel de desempeño en este contexto. A partir de estos resultados, se seleccionó LinearSVC como modelo final para la asignación de etiquetas en los datos de prueba, dado que garantiza un balance adecuado entre precisión y generalización.

Relación con ODS y estrategias para la organización

Los resultados permiten identificar de manera más precisa los temas predominantes en las opiniones ciudadanas y su conexión con los Objetivos de Desarrollo Sostenible. Palabras clave como 'salud', 'pobreza' o 'educación' reflejan problemáticas prioritarias que pueden orientar la toma de decisiones. Para la organización, contar con este tipo de análisis supone una ventaja al definir estrategias más focalizadas, asignar recursos de manera eficiente y fortalecer el diseño de políticas públicas basadas en evidencia.

Finalmente, los datos de prueba fueron etiquetados con el modelo seleccionado (LinearSVC) y entregados en formato Excel. Este archivo constituye la base para la evaluación comparativa entre grupos y garantiza la transparencia en la asignación de métricas de calidad como el F1-score. En conjunto, los resultados validan la pertinencia del modelo y su utilidad práctica para la organización.