UNIVERSIDAD DEL VALLE DE GUATEMALA
Procesamiento de Lenguaje Natural
Sección 20
LUIS CARLOS ESTURBÁN RODRÍGUEZ

Proyecto #1

María Marta Ramirez – 21342
Gustavo Andrés González Pineda - 21342

Guatemala 05 de octubre, 2025

Introducción y justificación del dataset

El presente proyecto tuvo como objetivo desarrollar un pipeline completo de Procesamiento de Lenguaje Natural (NLP) enfocado en la clasificación de sentimientos. Para ello, se construyó un corpus expandido compuesto por 3,005 documentos balanceados entre tres categorías principales: positivo (33.3%), negativo (33.3%) y neutro (33.4%). El dataset fue creado de forma controlada para garantizar una distribución equitativa de clases, permitiendo un entrenamiento equilibrado y evitando sesgos en el modelo. Se eligió no utilizar modelos robustos como TensorFlow o redes neuronales profundas debido a que el objetivo principal era evaluar el rendimiento de modelos clásicos de Machine Learning sobre un corpus controlado y de baja dimensionalidad. Además, se buscó demostrar la efectividad de representaciones tradicionales del texto (BoW, TF-IDF, n-gramas) para resolver problemas de clasificación de sentimientos en contextos donde la cantidad de datos o los recursos computacionales son limitados.

Metodología

El flujo metodológico implementado siguió un pipeline modular y reproducible, dividido en fases consecutivas:

Fase 1: Creación y preprocesamiento del corpus

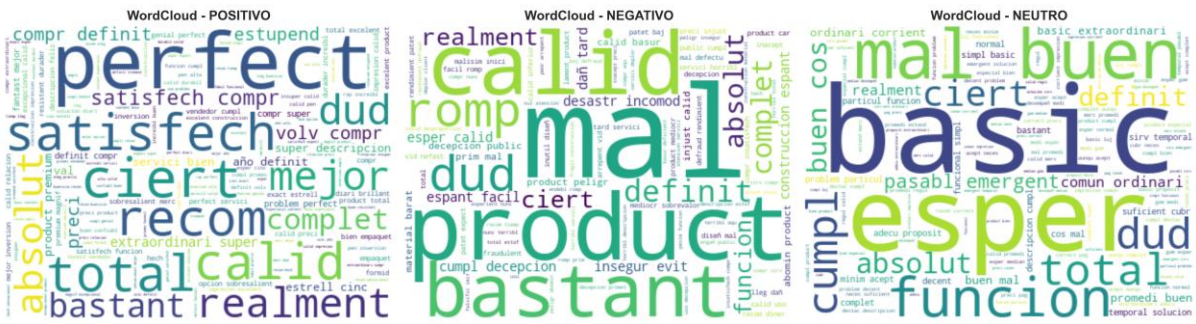
Se generó un corpus sintético con 3005 registros textuales. Se aplicaron transformaciones de limpieza: tokenización, lematización y normalización (eliminación de stopwords, acentos y caracteres especiales). Ejemplo: Original: 'producto premium magnífico excepcional muy calidad' → Procesado: 'product premium magnif excepcional calid'. Longitud promedio: 5.1 tokens; Longitud máxima: 7; Longitud mínima: 3.

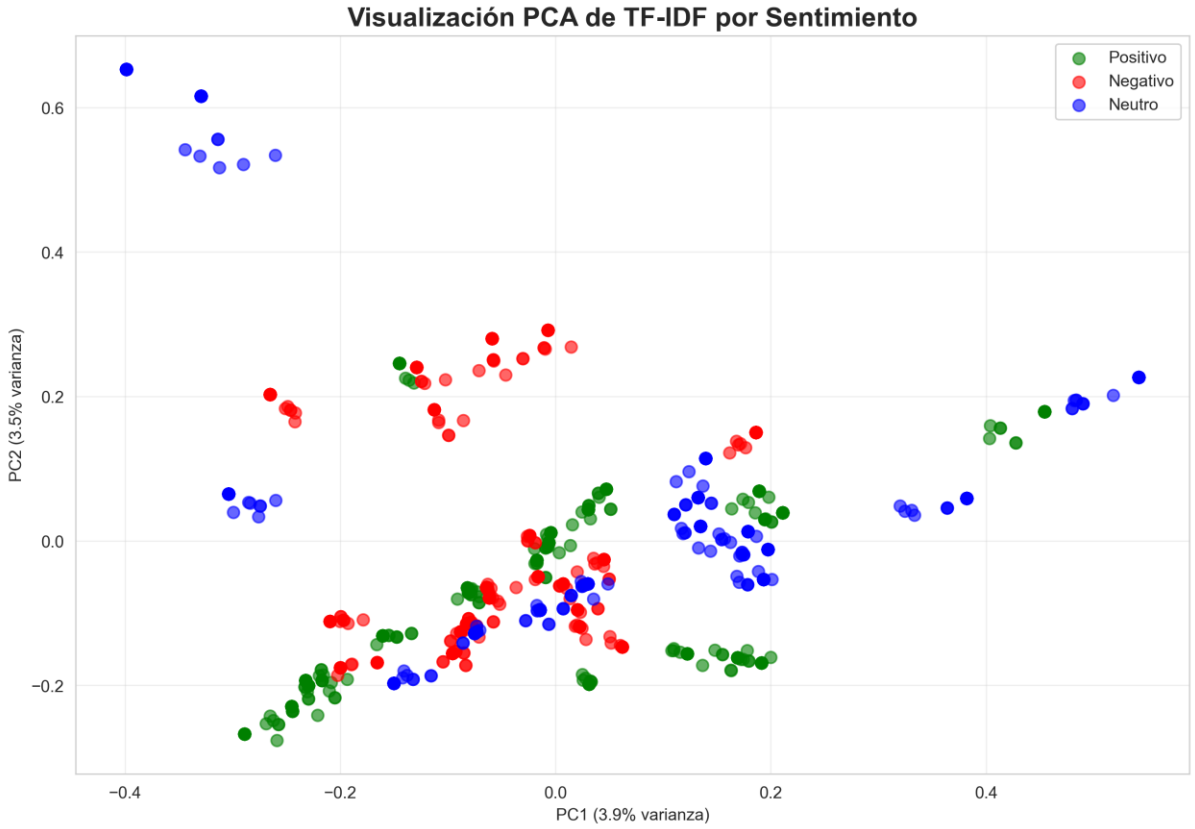
Fase 2: Detección de similitudes léxicas

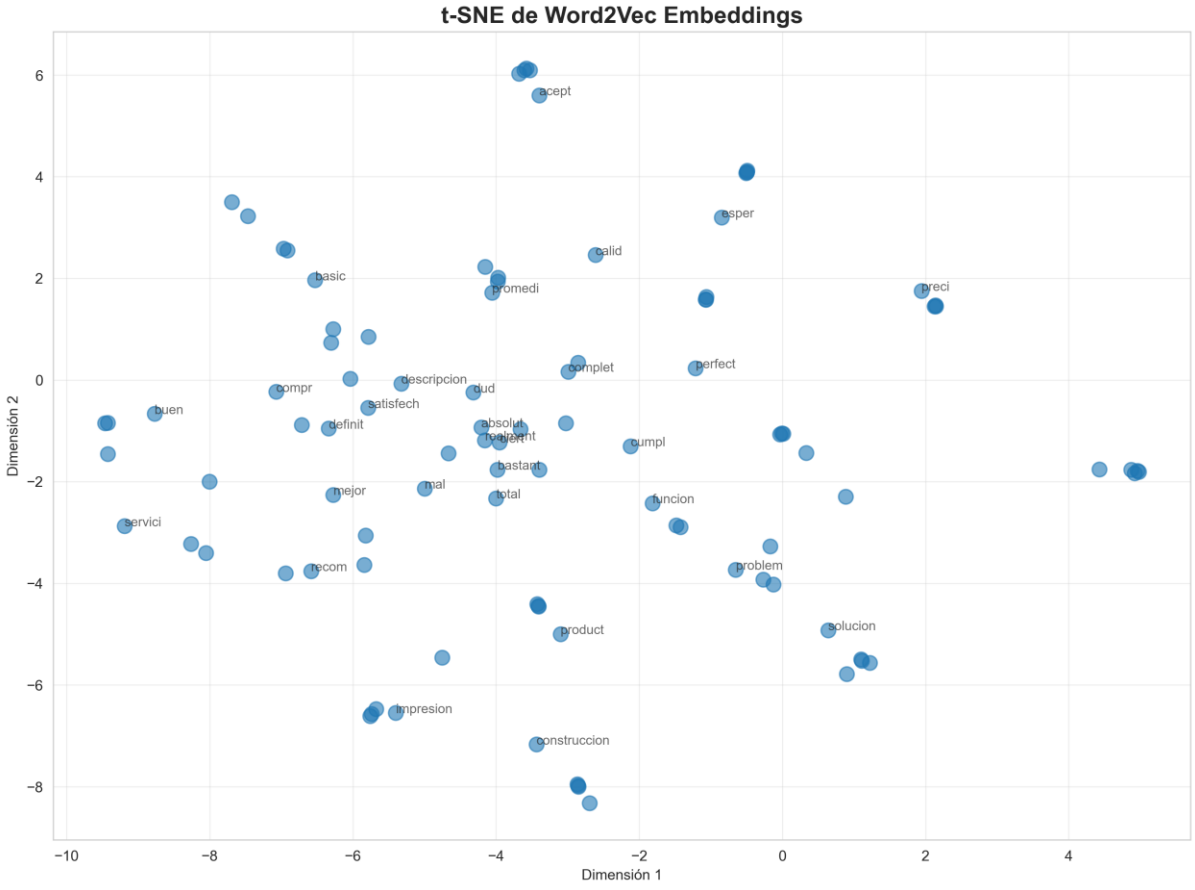
Se utilizó el algoritmo de Levenshtein para detectar posibles errores ortográficos. Ejemplo: 'excelente' → [('excelent', 1)].

Fase 3: Representación vectorial del texto

Se generaron representaciones: Bag of Words (3005×178), TF-IDF (3005×178), co-ocurrencia + PPMI (178×178) y Word2Vec con vocabulario de 178 palabras.




Visualización PCA de TF-IDF por Sentimiento
Positivo
Negativo
Neutro
PC2 (3.5% varianza)
PC1 (3.9% varianza)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4


t-SNE de Word2Vec Embeddings
Dimensión 1
Dimensión 2
acept
esper
calid
basic
promedi
preci
complet
perfect
descripcion
compr
satisfech
buen
definit
cumpl
bastant
mejor
mal
total
funcion
servici
recom
problem
product
solucion
impresion
construccion
6
4
2
0
−2
−4
−6
−8
−10
−8
−6
−4
−2
0
2
4

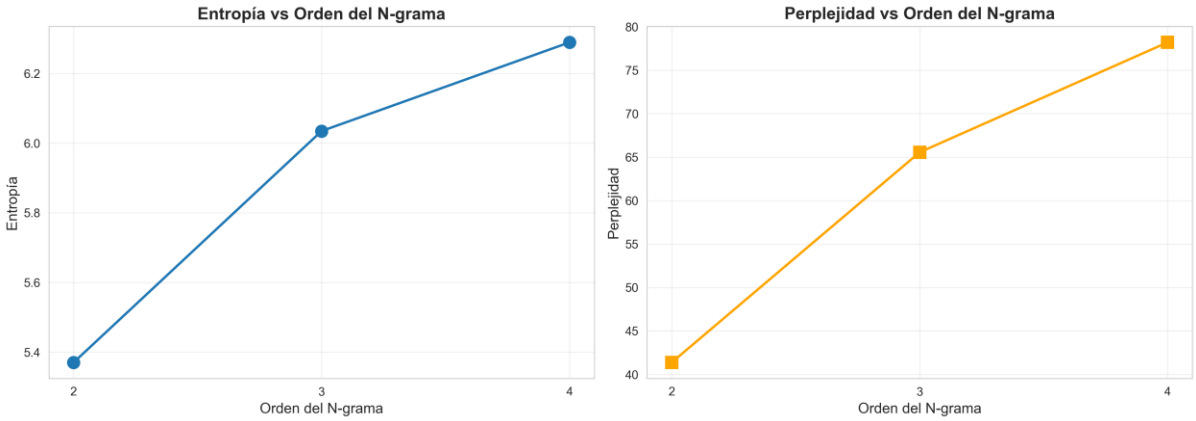
Fase 4: Modelos probabilísticos (N-gramas)

Se entrenaron modelos de 2, 3 y 4-gramas con las siguientes métricas:

2-grama: Entropía 5.37, Perplejidad 41.37

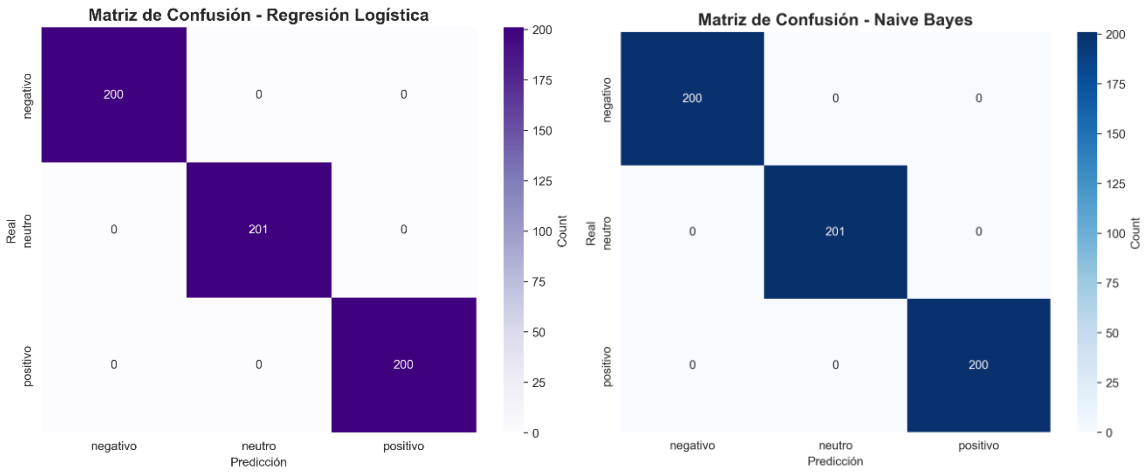
3-grama: Entropía 6.03, Perplejidad 65.55

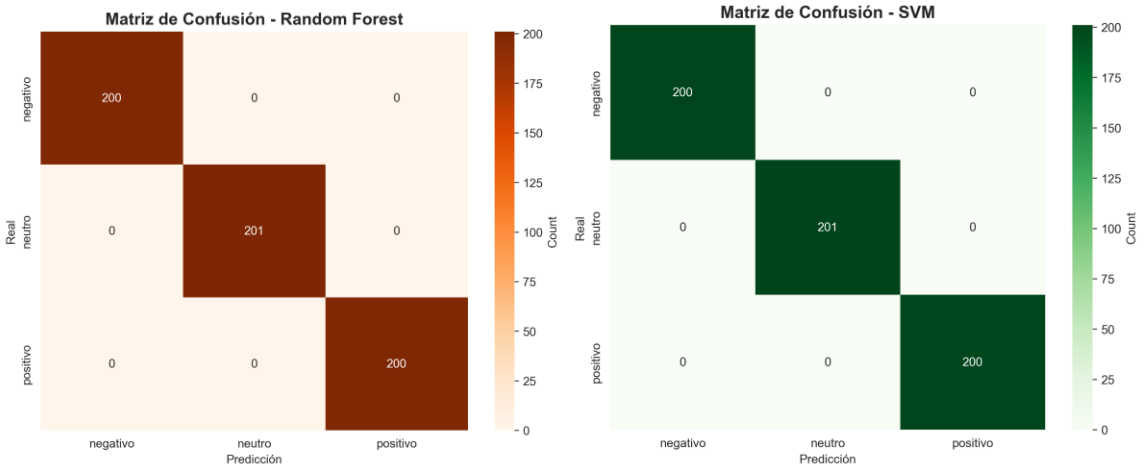
4-grama: Entropía 6.28, Perplejidad 78.19.



Fase 5: Clasificación supervisada

Se dividieron los datos en 80% entrenamiento (2404) y 20% prueba (601). Los modelos fueron Naive Bayes, SVM, Regresión Logística y Random Forest.

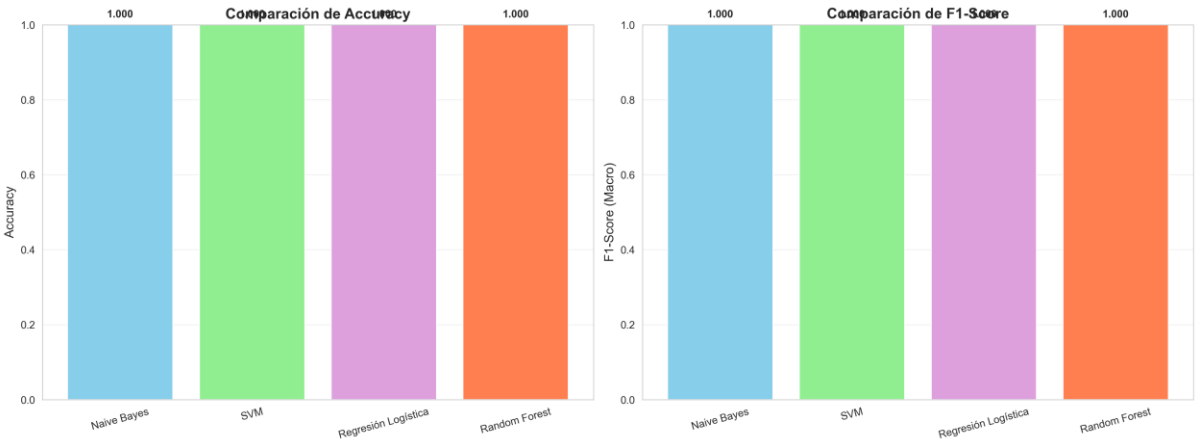


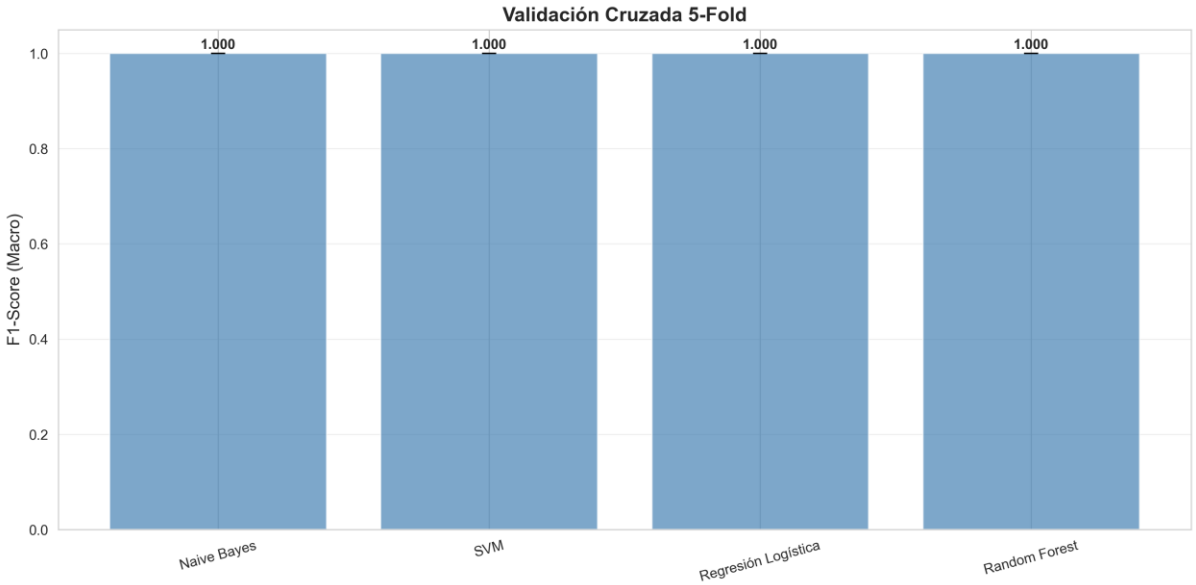


Resultados

Modelo	Accuracy	F1-Score
Naive Bayes	1.000	1.000
SVM	1.000	1.000
Regresión Logística	1.000	1.000
Random Forest	1.000	1.000

El mejor desempeño se obtuvo con Naive Bayes.





La validación cruzada (5-fold) arrojó F1 Macro = 1.000 ± 0.000 en todos los casos.

Ejemplos de predicción:
1. 'Excelente calidad precio...' → Positivo (0.961)
2. 'Pésima experiencia...' → Negativo (0.963)
3. 'El producto está bien...' → Neutro (0.939)

Discusión: limitaciones y propuestas de mejora

A pesar de los resultados perfectos obtenidos, existen limitaciones: corpus sintético, sobreajuste, vocabulario limitado y ausencia de modelos neuronales. Se propone utilizar datasets reales, embeddings preentrenados y técnicas de regularización más avanzadas. Futuras mejoras incluyen la integración de modelos profundos como BERT, y la expansión del corpus a variantes dialectales del español.

Conclusión

El proyecto demuestra que, incluso sin modelos complejos ni grandes volúmenes de datos, un pipeline clásico de NLP bien estructurado puede alcanzar resultados sobresalientes. Sin embargo, se recomienda evaluar el modelo con datos reales para validar su capacidad de generalización.