Universidad Rafael Landívar

Facultad de Ingeniería

Inteligencia artificial

Sección 01

Ingeniero Max Alejandro Cerna Flores

Proyecto 1 IA

José Daniel Man Catellanos 1020820

Luis Pablo Tujab Xuc 1103920

Guatemala, 23 de abril de 2025

Introducción

En la actualidad, las reseñas en línea se han convertido en una herramienta fundamental para tomar decisiones de compra. Cada día, millones de usuarios comparten sus opiniones sobre productos y servicios en plataformas como Amazon, influenciando así la percepción de otros consumidores. Sin embargo, este sistema abierto también ha dado paso a una problemática creciente: el crecimiento de reseñas falsas. Estas pueden ser generadas por competidores, vendedores, o bots, con el fin de manipular la reputación de un producto o empresa.

Ante este desafío, surge la necesidad de implementar soluciones automatizadas que ayuden a identificar y filtrar dichas reseñas falsas. En este contexto, la Inteligencia Artificial se presenta como una herramienta poderosa. Este proyecto se enfoca en el desarrollo de un sistema de clasificación que sea capaz de distinguir entre reseñas genuinas y reseñas falsas utilizando el algoritmo Naïve Bayes, conocido por su eficiencia y simplicidad en tareas de clasificación de texto.

Además, se busca complementar este motor de inferencia con una interfaz web sencilla e interactiva, que permita a cualquier usuario ingresar una reseña y obtener una predicción instantánea sobre su autenticidad. De esta forma, el proyecto no solo explora la aplicación teórica del algoritmo, sino que también ofrece una solución práctica que puede ser integrada en plataformas reales.

Definición del problema

La presencia de reseñas falsas en internet representa un problema significativo tanto para consumidores como para empresas legítimas. Estas opiniones manipuladas pueden inducir al error a miles de usuarios, generando desconfianza, malas decisiones de compra, e incluso pérdidas económicas. Aunque algunos sitios web han intentado implementar mecanismos de control, la sofisticación de las reseñas falsas ha hecho que su detección sea cada vez más compleja.

El problema que abordamos en este proyecto es el de clasificar automáticamente reseñas como genuinas o falsas, basándonos únicamente en el texto escrito por el usuario. Para ello, entrenaremos un modelo de clasificación binaria utilizando el algoritmo Naïve Bayes, el cual deberá identificar patrones lingüísticos, vocabulario sospechoso y otras señales en el lenguaje de una reseña falsa. El principal reto de este problema radica en la ambigüedad y variabilidad del lenguaje humano. Las reseñas falsas pueden estar bien redactadas, imitando el estilo de las reseñas genuinas, lo cual obliga al modelo a encontrar patrones más sutiles para realizar su predicción. Por eso, es muy importante aplicar un proceso riguroso de limpieza y preparación de los datos, así como una evaluación del desempeño del modelo para asegurar su efectividad.

Descripción dataset

El dataset utilizado para este proyecto es el disponible en OSF.io Fake Reviews Dataset, que contiene reseñas reales y falsas etiquetadas. Las principales características del dataset son:

- 20k reviews falsas
- 20k reviews verdaderas
- Diferentes categorías para las reseñas dependiendo de la clase de producto
- En la columna rating es la calificación que el usuario le dio al producto en una escala del 1 al 5 donde 1 es una mala experiencia y 5 es una muy buena experiencia
- La columna label es referente al tipo de reseña si es una verdadera o es una reseña falsa
- Las reseñas de los usuarios varían en longitud y estilo llegando a ser breves o más extensas

Descripción del procedimiento aplicado

El desarrollo del sistema se divido en las siguientes fases

1. Preprocesamiento de texto
    a. Conversión de mayúsculas a minúsculas
    b. Adaptación para caracteres especiales
2. Representación de texto
    a. Convertir el texto en vectores numéricos
    b. Cálculo de aparición de palabras repetidas
3. Entrenamiento del modelo
    a. Implementar Naïve Bayes
    b. Entrenamiento mediante 80% del dataset
4. Evaluación del modelo
    a. Evaluación de métricas a utilizar para medir los resultados como por ejemplo precision, recall, F1-score
5. Interfaz web
    a. Creación del frontend

Explicación del algoritmo Naiv Bayes

El algoritmo Naïve Bayes es un modelo probabilístico basado en el Teorema de Bayes, bajo el supuesto de que las características (en este caso, las palabras del texto) son independientes entre sí, dado el resultado de la clase. Naïve Bayes ha demostrado ser efectivo para clasificación de texto. El modelo estima la probabilidad de que una reseña sea real o falsa utilizando el siguiente modelo.

$$P(Clase|\ Reseña) = \frac{P(Reseña|Clase) \times P(Clase)}{P(Reseña)}$$

Durante el entrenamiento:

- Se calcula la probabilidad de aparición de cada palabra por clase.
- Se aplican técnicas de suavizado como Laplace smoothing para evitar probabilidades nulas.

Para clasificar una nueva reseña:

- Se calcula la probabilidad de que la reseña pertenezca a cada clase.
- Se selecciona la clase con mayor probabilidad.

Este enfoque es eficiente, rápido, y especialmente útil en problemas donde el tamaño del dataset es grande y el texto es corto, como en las reseñas de productos.

Explicación evaluación del modelo (métricas y resultados obtenidos)

El modelo TF-IDF Naïve Bayes entrenado con α=0.05 y n-gramas de longitud 1 a 4 alcanzó una precisión global del 90.98 % sobre un conjunto de prueba de 8 086 reseñas. La precisión para la clase “genuina” fue de 94.20 %, el recall de 87.17 % y el F1-score de 90.55 %. La matriz de confusión revela 215 falsos positivos y 514 falsos negativos, lo que indica un sesgo ligero hacia el conservadurismo al etiquetar reseñas genuinas como falsas. 80% de entrenamiento 32,846 y 20% de prueba el cual es el 8,086.

Resultados obtenidos con la mejor alpha, tiempo aproximado de pruebas para encontrar el mejor Alpha fue de 20 minutos

```
**Mejor alpha=0.05, F1=0.9119
***** Entrenamiento Final Completado *****
Best alpha=0.05
Test size: 8086
Accuracy : 0.9098
Precision: 0.9420
Recall   : 0.8717
F1-score : 0.9055
Matriz de Confusión (falsa/genuina):
[[3864  215]
 [ 514 3493]]
INFO:     127.0.0.1:65052 - "GET /train HTTP/1.1" 200 OK
```

Factores a favor

- Alta precisión (0.942): pocos falsos positivos (215 sobre ~4 k negatives), lo que indica que cuando predices “genuina” casi siempre aciertas.
- Balance de clases razonable: la matriz muestra ~4 079 ejemplos de “falsa” y ~4 007 de “genuina”, así que no parece haber un sesgo grave en el test.
- Grid search de α exhaustivo: recorrer [0.05,6] en pasos de 0.05 te asegura encontrar un suavizado óptimo para tu data.

Diagramas

Arquitectura de la solución

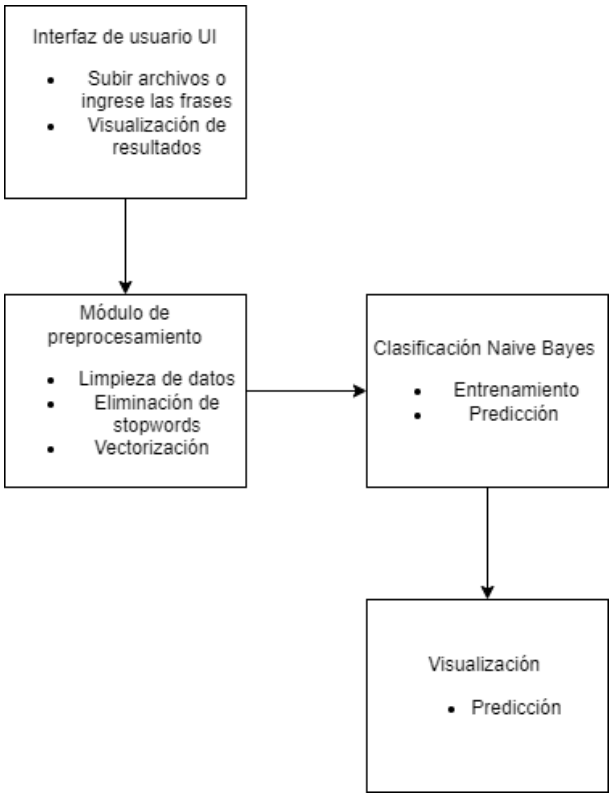


Diagrama de caso de uso

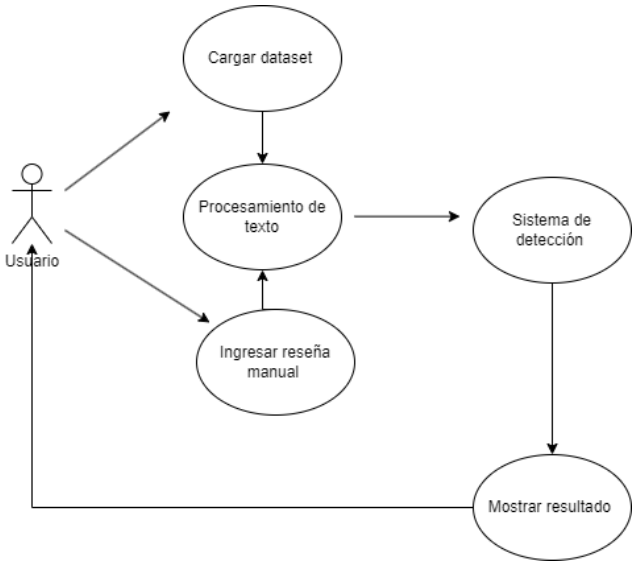


Diagrama de flujo general

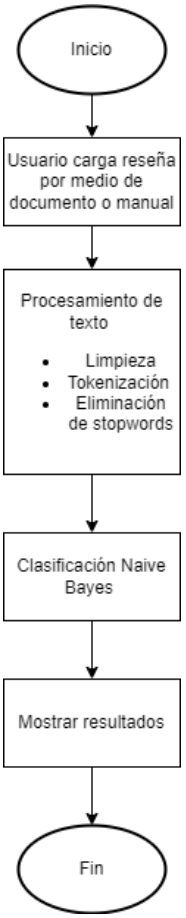


Diagrama de componentes

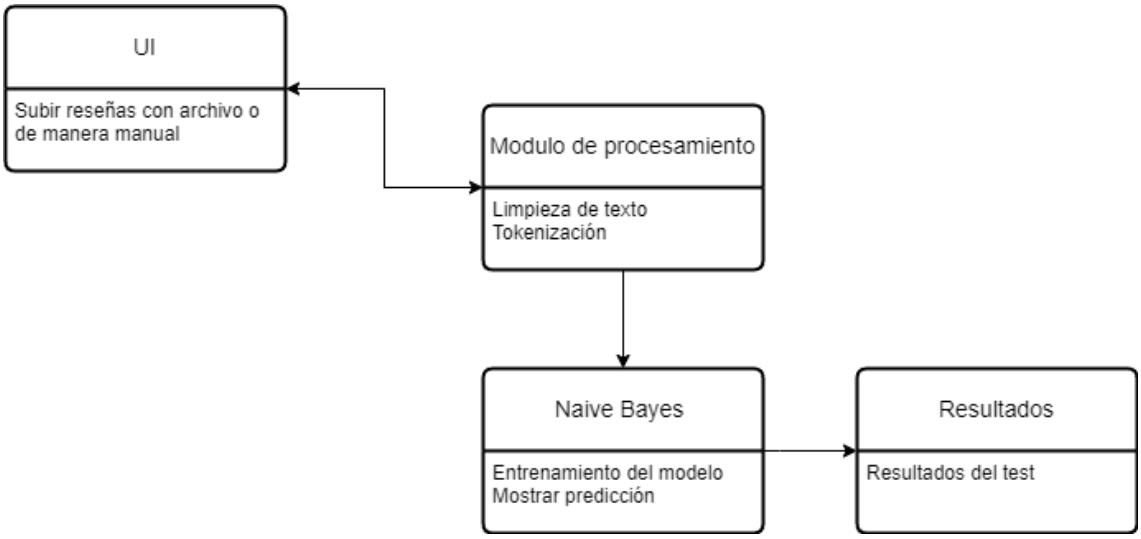
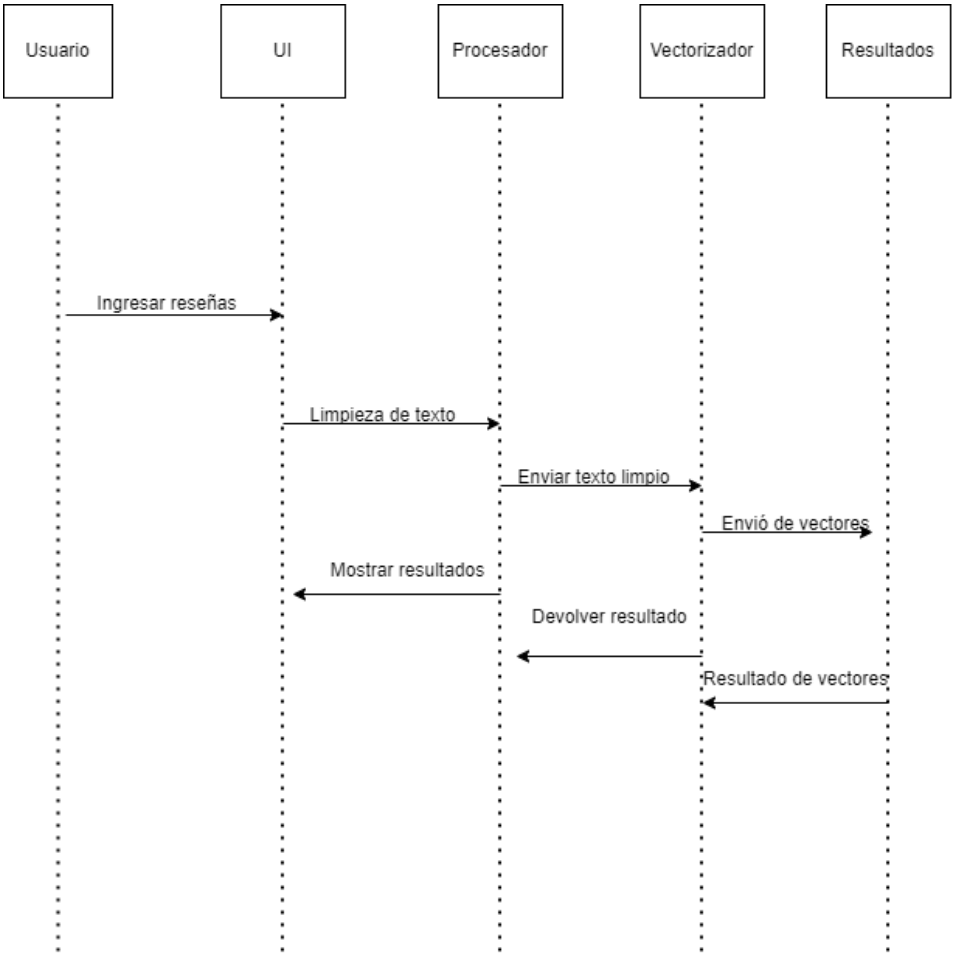


Diagrama de secuencias



Evidencia funcionamiento

Enlace video: https://youtu.be/BrftYf9FjHA

Conclusiones

1. El elevado F1-score (0.9055) y la alta precisión demuestran que la combinación de TF-IDF con n-gramas extensivos captura de manera efectiva las características discriminantes del texto, convirtiendo al modelo en un baseline robusto para detección de reseñas auténticas.
2. El modelo es especialmente fuerte minimizando falsos positivos (solo el 5.28 % de reseñas falsas se clasificaron como genuinas), lo que es crucial cuando el coste de aprobar información engañosa es alto.
3. La exhaustiva búsqueda de α y la validación cruzada garantizan un ajuste cuidadoso del suavizado, pero aún existen oportunidades de mejora como por ejemplo incorporar IDF en la fase de predicción, estratificar la partición train/test y refinar el preprocesamiento lingüístico podría elevar aún más el recall sin comprometer la precisión.