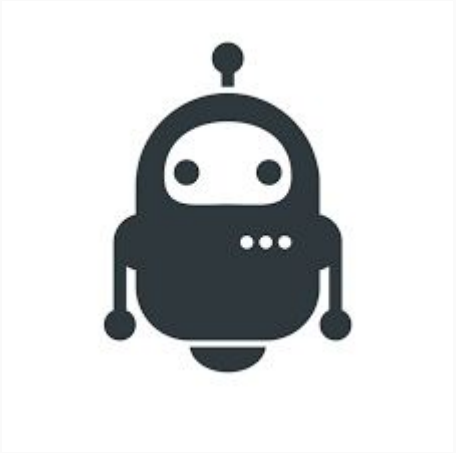
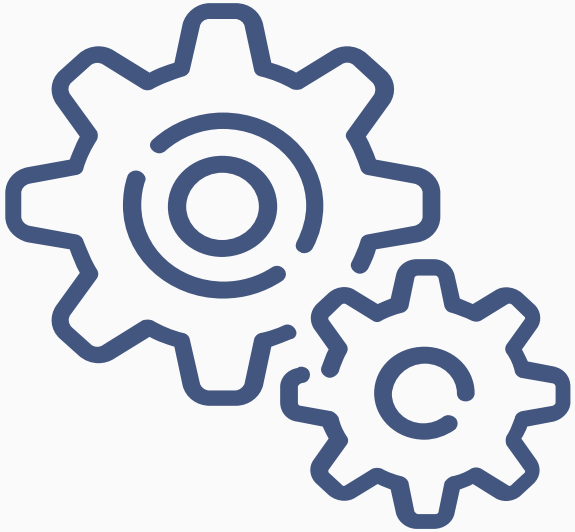
PREDICCIÓN RESEÑAS AMAZON

Índice

PRODUCTOS ELEGIDOS





PREPROCESAMIENTO BÁSICO

PREPROCESAMIENTO BÁSICO

1. LETRAS MINÚSCULAS

2. ELIMINACIÓN SIGNOS PUNTUACIÓN

3. ELIMINACIÓN STOPWORDS

4. TOKENIZACIÓN

VECTORIZACIÓN TF – IDF

BOW

El más simple y rápido

Clasificación textos pequeños

Bueno cuando la semántica no es relevante

VS

TF - IDF

Simple, eficiente, rápido y equilibrado

Reduce palabras irrelevantes

Destaca palabras importantes

No tiene en cuenta la relación semántica

VS

TF / IDF

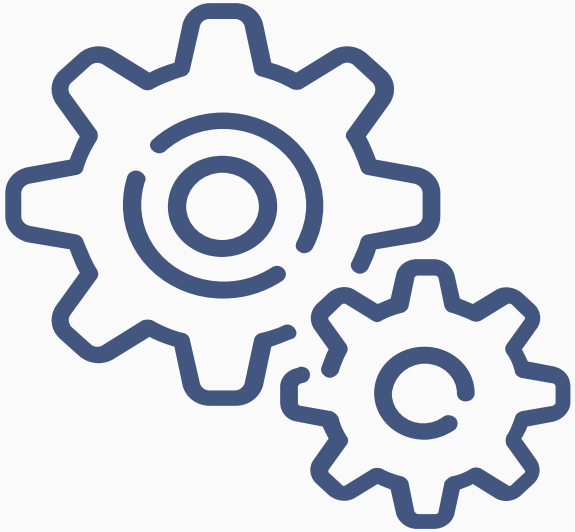
Mide importancia de palabras

Reduce palabras irrelevantes

No distingue palabras comunes y clave

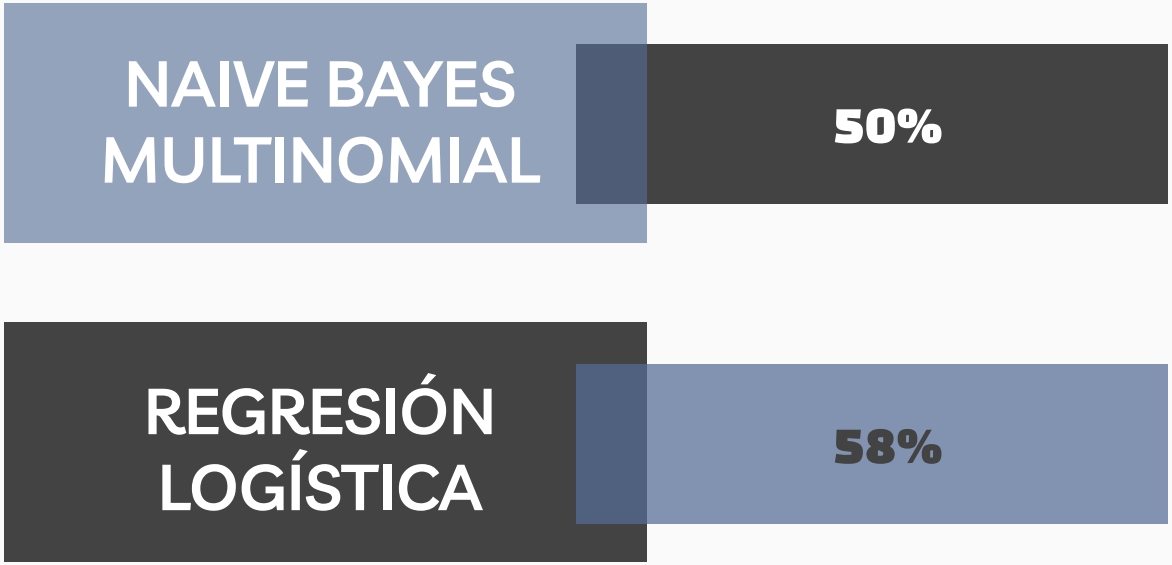
No considera la frecuencia

No suficientes por sí solos



ELECCIÓN MODELO INICIAL

MODELOS SELECCIONADOS



MEJORAS BASADAS EN EL MODELO BASE

Mejoras en cuanto a preprocesamiento

Lemming / Stemming

Recordando...

Análisis lingüístico para sacar el lema o forma canónica de cada palabra

Running -> Run
Better -> Good

Lemming

Se basa en recorte de sufijos. No devuelve palabras con sentido siempre

Running -> Run
Better -> Bet

Stemming

Intuimos

Resultados

Conclusiones

Idea de simpleza durante el resto del proyecto

Modelos sencillos no funcionan con datos complejos

Mejoras en cuanto al tratamiento aplicado

Vectorizador TF-IDF de Sklearn

N-Gramas

Las distintas palabras pueden ser tratadas individualmente, por parejas, de tres en tres...etc

Tratados como un solo token

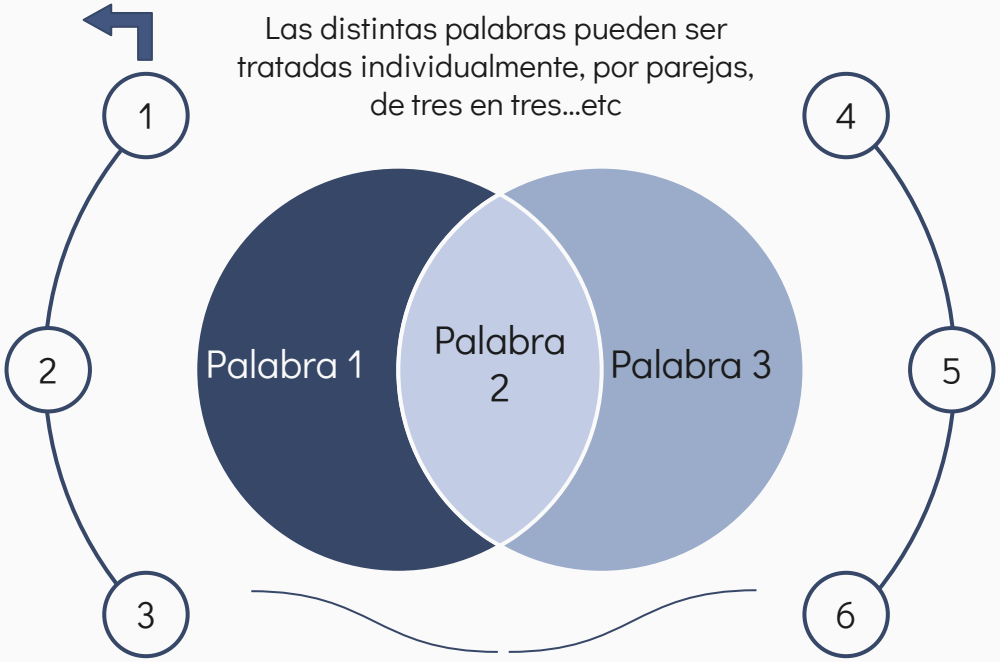
Intuimos que un valor de N mayor conseguirá mayor contexto y por lo tanto mejor resultado

N-Gramas

Una vez más el mejor resultado es el más sencillo

Implementamos incluso rangos en los que generaba combinaciones, ampliando el espacio de vectores con lo que aprendían los modelos

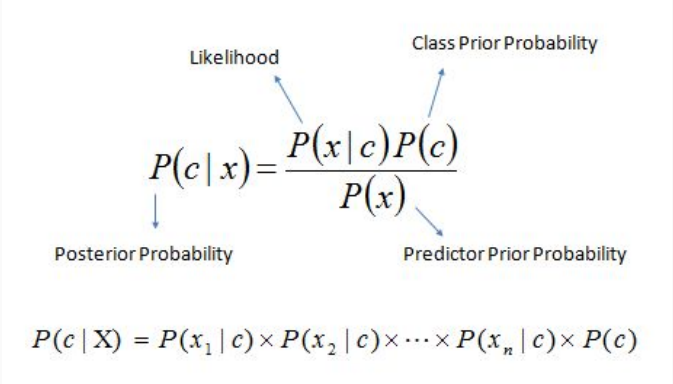
Las distintas palabras pueden ser tratadas individualmente, por parejas, de tres en tres...etc



Tratados como un solo token

N-Gramas

Vimos que el resultado era realmente malo para Naive Bayes



Con diferentes pruebas manuales y discutiendo con otros grupos, dimos con el porqué. Al usar conteo de palabras para calcular las probabilidades, si como palabra teníamos en cuenta agrupaciones de más de 3, las posibilidades de que se repitieran eran muy bajas por lo que no tenía ejemplos suficientes como para generar un buen entrenamiento

Normalización

¿Para qué?

Normalizar los textos de entradas resulta importante para equiparar la relevancia de reviews con distinto conteo de palabras

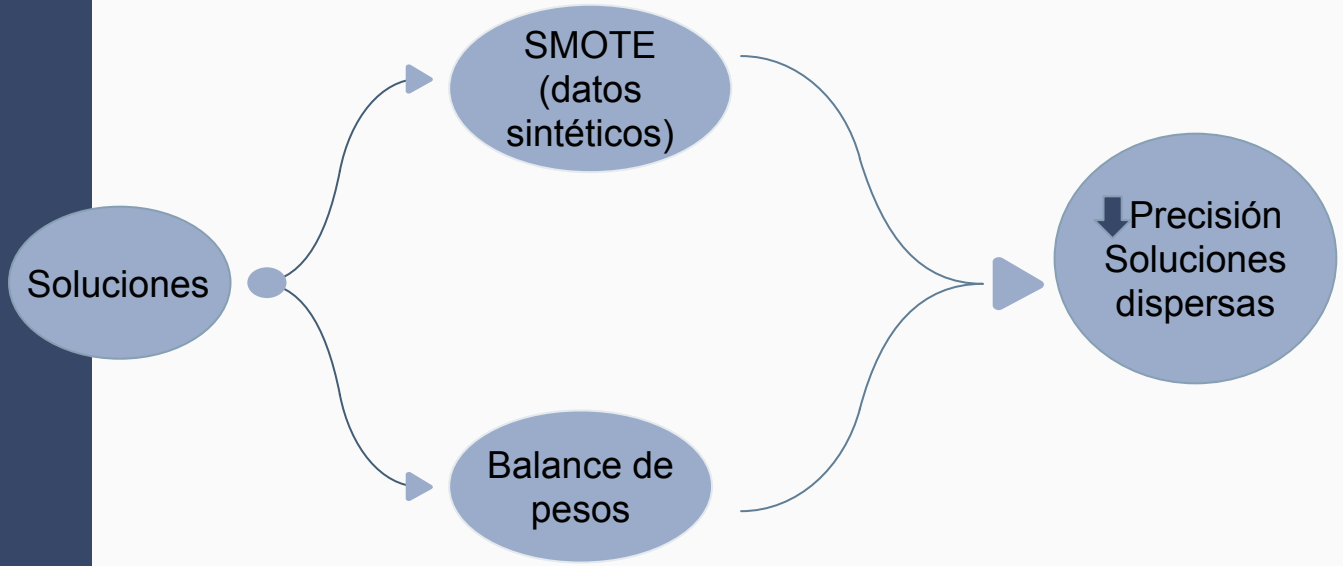
Si no lo normalizamos :
Las reviews con más palabras conseguirán solo por ese efecto valores más altos

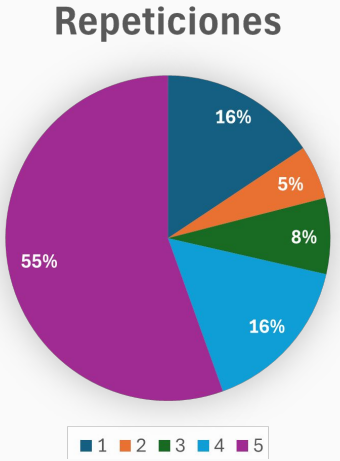
Normalización

¿Resultados?

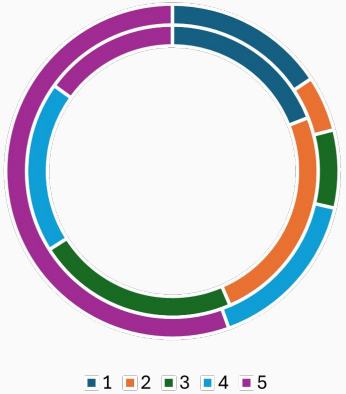
No se apreciaban cambios notables en los modelos con los que evaluamos, pero lo mantuvimos en mente para más adelante

Balanceo de clases

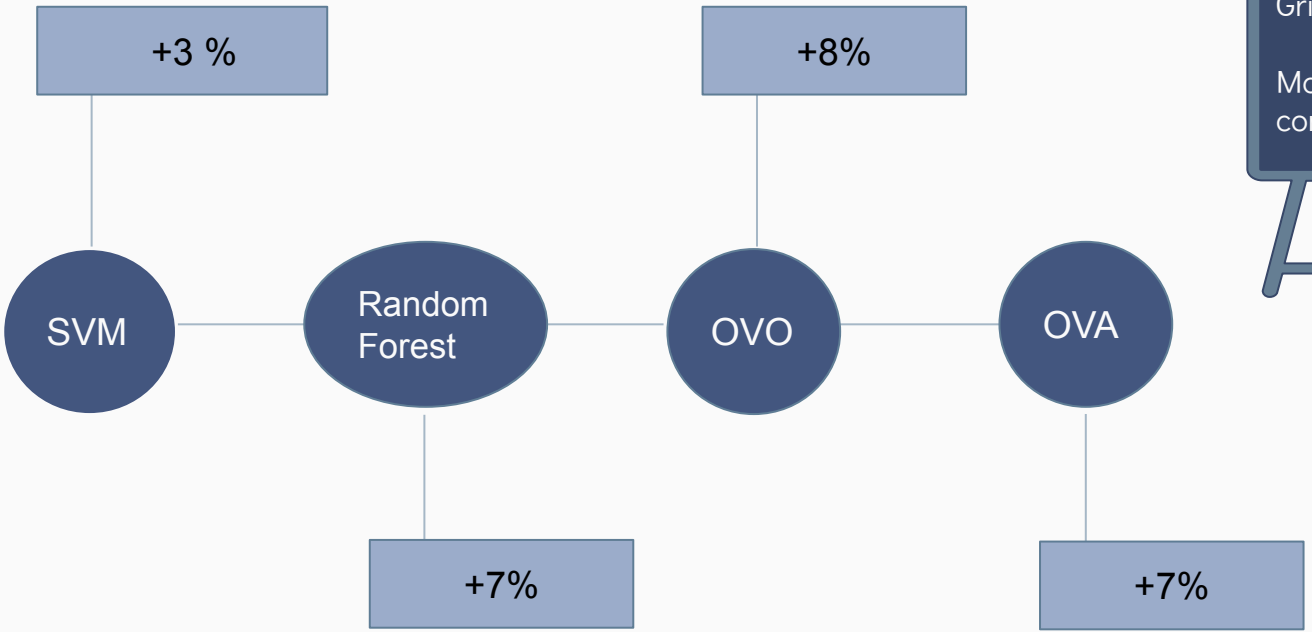




Relación frecuencia - pesos

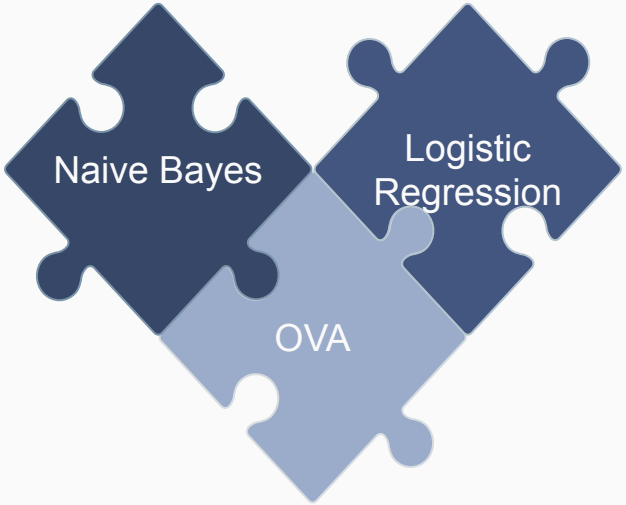


Comprobación de modelos



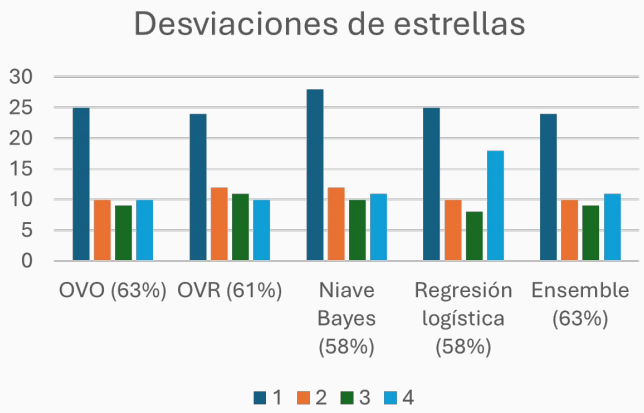
Ensemble

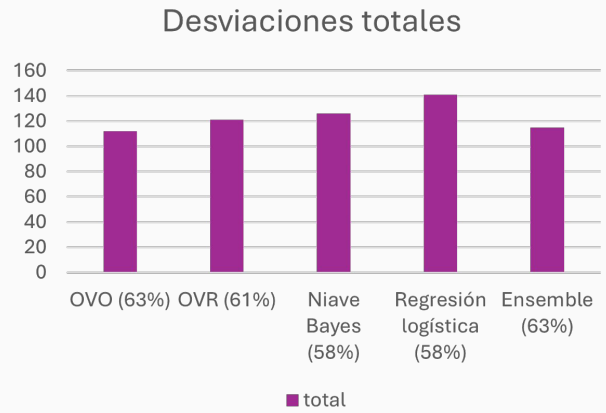
- Equilibrar clasificadores
- Distintos pesos



Nueva métrica: Desviación de estrellas

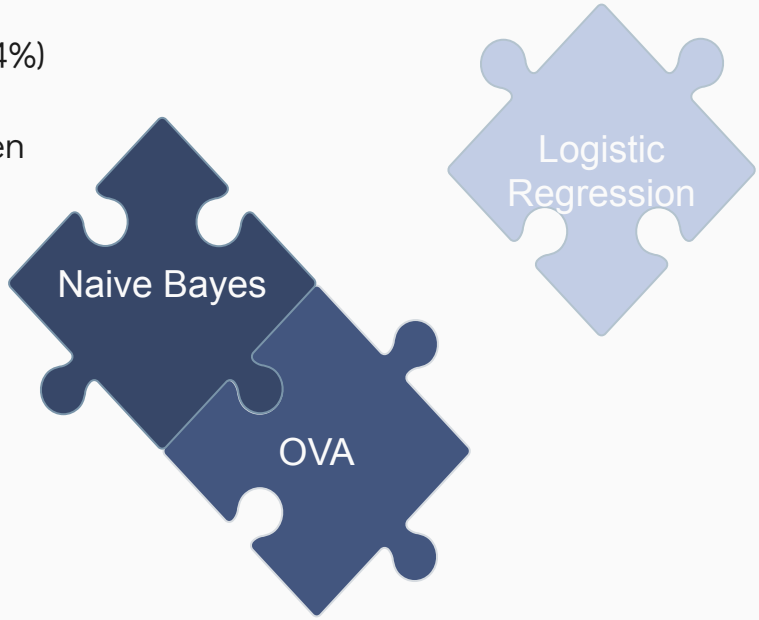
total = ∑ (nº de desviaciones * valor de la desviación)





Nueva métrica: Desviación de estrellas

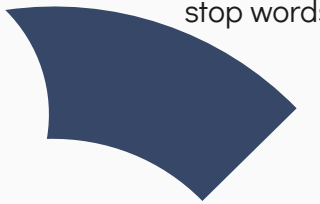
- Mayor precisión (64%)
- Menor desviación en predicciones
- OVA con apoyo de Naive Bayes



Conclusiones

Preprocesamiento

stop words y stemming



Conclusiones

Preprocesamiento

Preprocesamiento

stop words y stemming

vectorización TF-IDF sin n-gramas y con normalización L2

No balancear clases

Afecta negativamente la precisión.

Preprocesamiento

stop words y stemming

vectorización TF-IDF sin n-gramas y con normalización L2

No balancear clases

Afecta negativamente la precisión.

Métricas

Permite un análisis más detallado

Conclusiones

Preprocesamiento

stop words y stemming

vectorización TF-IDF sin n-gramas y con normalización L2

No balancear clases

Afecta negativamente la precisión.

Métricas

Permite un análisis más detallado

Modelo

Ensemble combinando Naive Bayes y OVA, lo que optimiza tanto la precisión como la desviación de las predicciones.

64% de precisión y reducción en la desviación de las predicciones

Preprocesamiento

stop words y stemming

vectorización TF-IDF sin n-gramas y con normalización L2

No balancear clases

Afecta negativamente la precisión.

Métricas

Permite un análisis más detallado

Modelo

Ensemble combinando Naive Bayes y OVA, lo que optimiza tanto la precisión como la desviación de las predicciones.

¡Gracias!

¿Alguna pregunta?