Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos II

Grupo 101

27 de septiembre de 2024

Clasificación de sentimientos base de datos IMBD

Autor:

Catherine Johanna Rojas Mendoza

A01798149

Doctor:

Alfredo Esquivel Jaramillo

Análisis de Sentimientos con Reseñas de IMDB utilizando Técnicas de Aprendizaje Automático

El código aborda el proceso de descarga, preprocesamiento y análisis del conjunto de datos de reseñas de películas de IMDB para la clasificación de sentimientos utilizando varios modelos de aprendizaje automático. El código comienza descargando el conjunto de datos de IMDB y extrayendo su contenido. Posteriormente, se lleva a cabo el preprocesamiento de los datos de texto, que incluye la tokenización, la eliminación de palabras vacías (stop-words) y el uso de expresiones regulares para limpiar el texto.
Se emplea un modelo de *bolsa de palabras* (Bag-of-Words) y *TF-IDF*(Term frequency – Inverse Document Frequency) para representar los datos textuales de manera numérica, lo cual es necesario para que los modelos de aprendizaje automático puedan realizar inferencias. El código implementa varios clasificadores, entre ellos regresión logística, máquina de soporte vectorial (SVM), bosques aleatorios y un clasificador de boosting por gradiente para predecir el sentimiento de las reseñas como positivo o negativo.
Adicionalmente, proporciona funcionalidades para visualizar la importancia de las características y explorar técnicas de vectorización como TF-IDF.
Este proceso nos muestra los componentes esenciales del análisis de sentimientos en el procesamiento del lenguaje natural (NLP), desde la adquisición y preprocesamiento de datos hasta el entrenamiento y evaluación de modelos.

Comparación de resultados con BOW

Entrenar con features BOW con Nsamp = 1000, maxtokens = 50, maxtokenlen = 20

Modelo	Exactitud	AUC (Train)	CV Score (Media)	Tiempo de Entrenamiento (s)
Regresión Logística	0.685	-	-	-
SVC	0.6667	-	-	4
Bosques Aleatorios	0.7083	-	-	4
Gradient Boosting	0.675	0.9509	0.7159	131

Cuadro 1
Comparación de Modelos de Clasificación

Observaciones

El modelo de Gradient Boosting Machines exhibe un rendimiento superior en términos de exactitud (0.8564) durante el entrenamiento y alcanza un AUC considerablemente alto. No obstante, su disminución en el rendimiento en el conjunto de prueba (0.675) sugiere un posible sobreajuste.
Por otro lado, el Bosque Aleatorio logra un buen equilibrio entre tiempo de entrenamiento y exactitud en el conjunto de prueba.
Los tres primeros modelos, Máquina de Soporte Vectorial, Regresión Logística y Bosques Aleatorios, fueron

los más rápidos de entrenar. Sin embargo, la exactitud de la Máquina de Soporte Vectorial y la Regresión Logística es ligeramente inferior en comparación con la obtenida por el Bosque Aleatorio y Gradient Boosting.

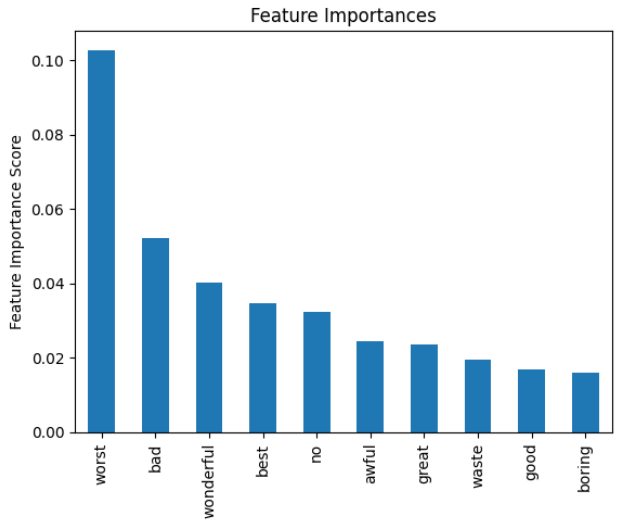

Figura 1
Feature Importances 1

Para este caso, la gráfica muestra que el modelo de clasificación Gradient Boosting Machine considera ciertas palabras como *worst* y *bad* como las más importantes para realizar su predicción. Esto indica que estas palabras tienen una fuerte influencia en la decisión del modelo, posiblemente al identificar opiniones negativas o mensajes con cierto tono emocional. Las palabras con menor importancia, como *good* y *boring*, contribuyen menos a las decisiones del modelo. De manera general podemos decir que, el modelo se basa principalmente en palabras con connotaciones fuertes para clasificar el texto.

Entrenar con features BOW con Nsamp = 1000, maxtokens = 100, maxtokenlen = 100

Modelo	Exactitud	AUC (Train)	CV Score (Media)	Tiempo de Entrenamiento (s)
Regresión Logística	0.7483	-	-	-
SVC	0.7333	-	-	5
Bosques Aleatorios	0.7433	-	-	2
Gradient Boosting	0.7083	0.9677	0.7680	142

Cuadro 2
Comparación de Modelos de Clasificación con Bag-of-Words

Observaciones

La Regresión Logística y el Bosque Aleatorio muestran un buen equilibrio entre tiempo de entrenamiento y precisión en el conjunto de prueba. Gradient Boosting Machines ofrece un alto rendimiento en

entrenamiento (.89), pero la diferencia con la exactitud en prueba sugiere que podría estar ajustándose demasiado a los datos de entrenamiento.

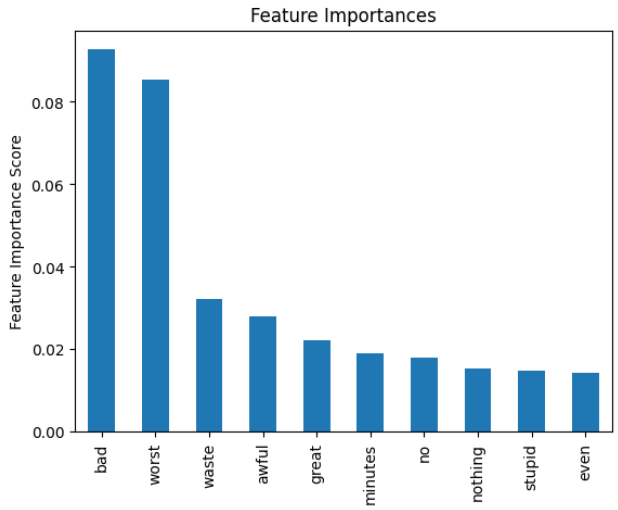

Figura 2
Feature Importances 2

En este caso, las palabras *bad* y *worst* se destacan como las más influyentes para el modelo, indicando su gran peso en la clasificación. Otras palabras como *waste*, *awful* y *great* también resultan relevantes, aunque en menor medida. La presencia de términos con connotaciones negativas (*bad, worst, waste, awful*) señala fuertes indicadores de reseñas negativas, mientras que la inclusión de la palabra *great* sugiere que el modelo también aprende a identificar palabras positivas para clasificar de manera efectiva.

Entrenar con features BOW con Nsamp = 1000, maxtokens = 200, maxtokenlen = 100

Modelo	Exactitud	AUC (Train)	CV Score (Media)	Tiempo de Entrenamiento (s)
Regresión Logística	0.7933	-	-	-
SVC	0.7833	-	-	59
Bosques Aleatorios	0.7683	-	-	2
Gradient Boosting	0.7217	0.9825	0.8281	197

Cuadro 3
Comparación de Modelos de Clasificación con Bag-of-Words

Observaciones

En esta comparación de los diferentes modelos de clasificación se muestra que la Regresión Logística es el modelo más efectivo y equilibrado en términos de exactitud y eficiencia con los hiperparámetros actuales. A pesar de que los modelos más complejos, como el Gradient Boosting, presentan un alto desempeño durante el entrenamiento (0.92), indicando así su capacidad de capturar patrones en los datos, sufre de una

caída de exactitud en el conjunto de prueba, lo cual sugiere un posible sobreajuste. El SVC y los Bosques Aleatorios ofrecen un buen comportamiento entre tiempo de entrenamiento y precisión, aunque no superan a la regresión logística.

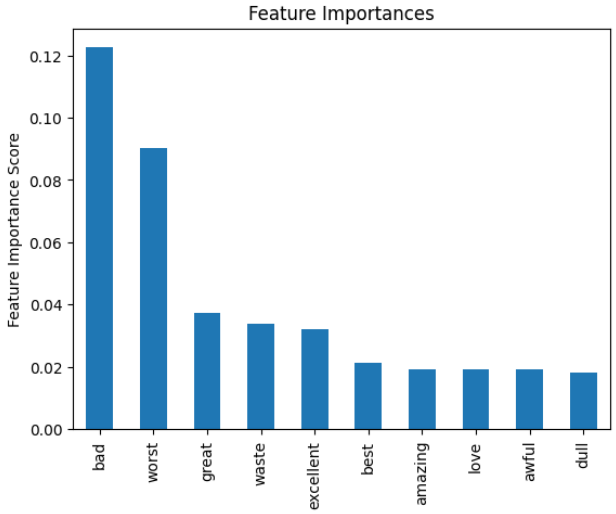

Figura 3
Feature Importances 3

La gráfica muestra que las palabras *bad* y *worst* tienen la mayor puntuación de importancia, lo que significa que el modelo las utiliza como principales indicadores para la clasificación. Otras palabras como *great*, *waste*, y *excellent* también tienen una importancia significativa, sugiriendo que el modelo se basa tanto en términos positivos como negativos para tomar decisiones. La presencia de términos positivos (*excellent, love, amazing*) y negativos (*bad, worst, waste*) indica que el modelo es capaz de capturar la polaridad emocional en el texto, lo cual es crucial para tareas como la clasificación de opiniones o la detección de *spam*.

Comparación de resultados con TF-IDF

Entrenar con features TF-IDF con Nsamp = 1000, maxtokens = 50, maxtokenlen = 20

Modelo	Exactitud	AUC (Train)	CV Score (Media)	Tiempo de Entrenamiento (s)
Regresión Logística	0.7117	-	-	-
SVC	0.7133	-	-	69
Bosques Aleatorios	0.6633	-	-	3
Gradient Boosting	0.6283	0.9638	0.6804	179

Cuadro 4
Comparación de Modelos de Clasificación con TF-IDF

Observaciones

La Regresión Logística y la Máquina de Soporte Vectorial (SVC) mantienen un desempeño decente, con exactitudes de 0.7117 y 0.7133 respectivamente.

Por otro lado, Bosques Aleatorios obtuvo la exactitud más baja (0.6633) pero se entrenan rápidamente en 3 segundos, manteniendo su eficiencia. En comparación, Gradient Boosting presenta una caída considerable en su rendimiento, con una exactitud de 0.6283 en el conjunto de prueba y una notable brecha con su AUC en entrenamiento, lo que indica un posible sobreajuste; además, su tiempo de entrenamiento es considerablemente largo (179 segundos).

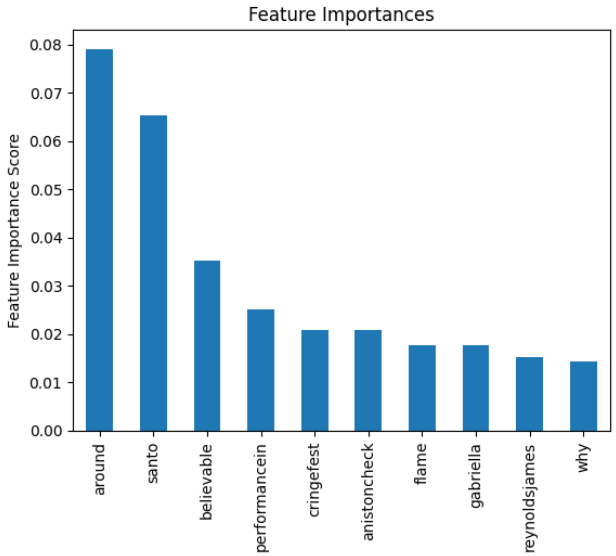

Figura 4
Feature Importances 4

Las palabras *around* y *santo* son las más importantes, seguidas de *believable*. Esto indica que estas palabras tienen un peso significativo en la toma de decisiones del modelo. La presencia de términos más generales *around* y nombres propios *santo, gabriella, reynoldsjames* sugiere que el modelo con TF-IDF está capturando una mezcla de palabras con diferentes tipos de información. Comparado con gráficos previos, la importancia de palabras con carga emocional se reduce, lo cual es típico en representaciones TF-IDF que ponderan términos menos comunes.

Entrenar con features TF-IDF con Nsamp = 1000, maxtokens = 100, maxtokenlen = 100

Modelo	Exactitud	AUC (Train)	CV Score (Media)	Tiempo de Entrenamiento (s)
Regresión Logística	0.7550	-	-	-
SVC	0.7483	-	-	52
Bosques Aleatorios	0.7317	-	-	4
Gradient Boosting	0.6883	0.9777	0.7598	270

Cuadro 5
Comparación de Modelos de Clasificación con TF-IDF

Observaciones

La Regresión Logística obtuvo la mayor exactitud en el conjunto de prueba. El SVC ofrece un desempeño comparable, aunque con un tiempo de entrenamiento significativamente más prolongado.

Los Bosques Aleatorios logran un equilibrio adecuado entre precisión y velocidad de entrenamiento, siendo una opción ideal cuando se busca eficiencia computacional.

Por otro lado, Gradient Boosting muestra signos claros de sobreajuste, ya que su exactitud en entrenamiento fue de 0.9057, pero disminuyó significativamente durante la prueba. Además, su tiempo de entrenamiento es considerablemente largo, lo que lo hace menos práctico para esta configuración.

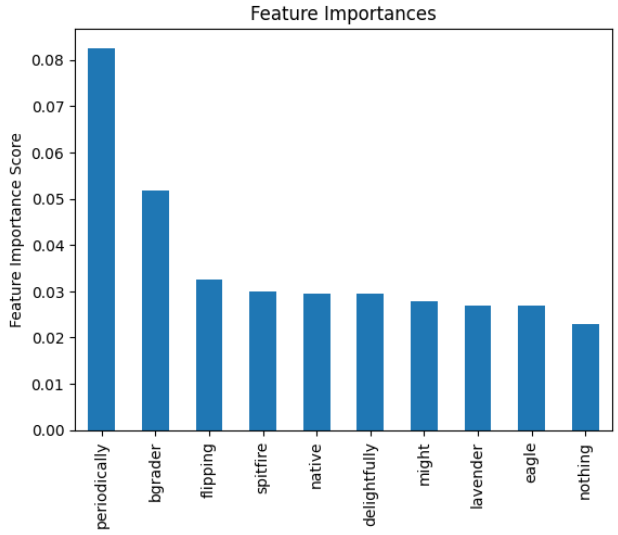

Figura 5
Feature Importances 5

La gráfica destaca las características más importantes identificadas por el modelo de clasificación, donde *periodically* emerge como la más influyente, seguida por *bgrader*, lo que sugiere su relevancia en la toma de decisiones. Las palabras relevantes carecen de carga emocional evidente, indicando que el modelo se enfoca en patrones contextuales específicos. Términos menos comunes, como *bgrader* y *spitfire*, aportan

información distintiva, lo que demuestra que el modelo utiliza detalles particulares para la clasificación. Aunque *periodically* sobresale, las demás palabras tienen importancias similares, sugiriendo que el modelo se basa en una combinación equilibrada de términos para clasificar de manera efectiva.

Entrenar con features BOW con Nsamp = 1000, maxtokens = 200, maxtokenlen = 100

Modelo	Exactitud	AUC (Train)	CV Score (Media)	Tiempo de Entrenamiento (s)
Regresión Logística	0.7867	-	-	-
SVC	0.7933	-	-	3516
Bosques Aleatorios	0.7550	-	-	10
Gradient Boosting	0.7017	0.9833	0.8181	700

Cuadro 6
Comparación de Modelos de Clasificación con TF-IDF

Observaciones

La Máquina de Soporte Vectorial (SVC) obtuvo la mayor exactitud (0.7933), pero a costa de un tiempo de entrenamiento extremadamente largo (3516 segundos), lo que la hace poco práctica en aplicaciones con restricciones de tiempo. La Regresión Logística, con una exactitud de 0.7867, ofrece un rendimiento muy similar al SVC, pero con un menor tiempo de entrenamiento, lo que la convierte en una opción eficiente. Los Bosques Aleatorios presentan un equilibrio adecuado, con una exactitud de 0.7550 y un tiempo de entrenamiento muy bajo, lo que los hace ideales cuando se requiere rapidez y confiabilidad. Por otro lado, el modelo Gradient Boosting muestra signos de sobreajuste, con una alta exactitud en entrenamiento (0.9236) pero una menor capacidad de generalización en el conjunto de prueba y un tiempo de entrenamiento considerable (700 segundos), lo que limita su practicidad en esta configuración.

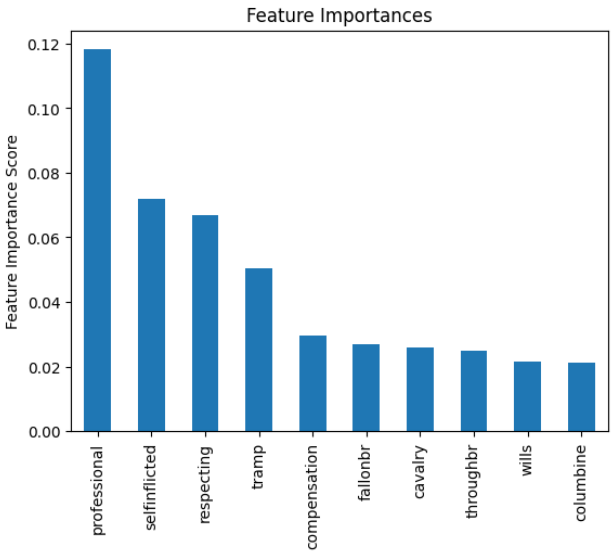

Figura 6
Feature Importances 6

La palabra *professional* tiene la mayor importancia, seguida por *selfinflicted* y *respecting.* Esto sugiere que estas palabras influyen significativamente en la decisión del modelo, lo que implica que su presencia en el texto es un fuerte indicador para clasificarlo en una determinada categoría.

El hecho de que las palabras incluyan términos como *professional, compensation, y cavalry* indica que el modelo está utilizando una combinación de términos formales y específicos para diferenciar las clases. La diversidad de palabras, desde términos profesionales hasta nombres propios (*columbine*), refleja la capacidad del modelo para captar información contextual y de contenido en el texto pero no los sentimientos.

Tabla con todos los resultados

Modelo	Exactitud	AUC (Train)	CV Score (Media)	Tiempo de Entrenamiento (s)
Bag-of-Words				
Nsamp = 1000, maxtokens = 50, maxtokenlen = 20				
Regresión Logística	0.685	-	-	-
SVC	0.6667	-	-	4
Bosques Aleatorios	0.7083	-	-	4
Gradient Boosting	0.675	0.9509	0.7159	131
Nsamp = 1000, maxtokens = 100, maxtokenlen = 100				
Regresión Logística	0.7483	-	-	-
SVC	0.7333	-	-	5
Bosques Aleatorios	0.7433	-	-	2
Gradient Boosting	0.7083	0.9677	0.7680	142
Nsamp = 1000, maxtokens = 200, maxtokenlen = 100				
Regresión Logística	0.7933	-	-	-
SVC	0.7833	-	-	59
Bosques Aleatorios	0.7683	-	-	2
Gradient Boosting	0.7217	0.9825	0.8281	197
TF-IDF				
Nsamp = 1000, maxtokens = 50, maxtokenlen = 20				
Regresión Logística	0.7117	-	-	-
SVC	0.7133	-	-	69
Bosques Aleatorios	0.6633	-	-	3
Gradient Boosting	0.6283	0.9638	0.6804	179
Nsamp = 1000, maxtokens = 100, maxtokenlen = 100				
Regresión Logística	0.7550	-	-	-
SVC	0.7483	-	-	52
Bosques Aleatorios	0.7317	-	-	4
Gradient Boosting	0.6883	0.9777	0.7598	270
Nsamp = 1000, maxtokens = 200, maxtokenlen = 100				
Regresión Logística	0.7867	-	-	-
SVC	0.7933	-	-	3516
Bosques Aleatorios	0.7550	-	-	10
Gradient Boosting	0.7017	0.9833	0.8181	700

Cuadro 7
Comparación de Modelos de Clasificación con Bag-of-Words y TF-IDF

Conclusión

En general, los modelos de Regresión Logística y SVC ofrecen el mejor rendimiento en términos de exactitud. La Regresión Logística destaca por su eficiencia computacional y su desempeño consistente bajo distintas configuraciones y representaciones de datos (Bag-of-Words y TF-IDF). Aunque el SVC logra la mayor exactitud en algunas configuraciones, su elevado costo computacional lo vuelve menos práctico para ciertos contextos, especialmente aquellos con recursos limitados. Por otro lado, los Bosques Aleatorios ofrecen un equilibrio adecuado entre precisión y eficiencia de entrenamiento. El Gradient Boosting, a pesar de su alto desempeño en el conjunto de entrenamiento, sufre de sobreajuste y presenta largos tiempos de

entrenamiento, lo que lo hace menos ideal para este conjunto de datos.

En conclusión, para una combinación óptima de precisión y eficiencia, la Regresión Logística es la opción más recomendable en este escenario. El SVC podría ser una alternativa si se busca la máxima exactitud y se dispone de los recursos computacionales necesarios.

La elección del modelo y la técnica de vectorización impactan directamente en los criterios de clasificación. Los modelos basados en Bag-of-Words sobresalen al capturar la polaridad emocional en el texto, lo que resulta especialmente útil para clasificar opiniones. En cambio, los modelos que utilizan TF-IDF se enfocan en resaltar detalles contextuales y específicos, haciéndolos adecuados para análisis que requieren identificar patrones más complejos en el texto. Esta comparación resalta la importancia de seleccionar la representación de texto más adecuada según el objetivo del análisis de sentimientos, ya sea centrarse en el contenido emocional o en los detalles y contexto específicos del texto.