REVISTA ELECTRÓNICA SOBRE TECNOLOGÍA, EDUCACIÓN Y SOCIEDAD
ISSN: 2448 - 6493
CTES

*Artículos científicos*

# Comparación de algoritmos de machine learning para procesamiento de lenguaje natural

## *Comparison of machine learning algorithms for natural language processing*

## *Comparação de algoritmos de aprendizagem automática para processamento de linguagem natural*

**Yamil Emanuel Castro Solis**
Universidad Politécnica de Victoria, México
yamilyecs@gmail.com
https://orcid.org/0009-0001-1806-5641

**Hiram Herrera Rivas**
Universidad Politécnica de Victoria, México
hiramhr@gmail.com
https://orcid.org/0000-0002-2650-8932

**Resumen**
El presente estudio se centra en la comparación de algoritmos de machine learning para el procesamiento de lenguaje natural, específicamente en tareas de clasificación y análisis de texto. Se analizarán cuatro algoritmos ampliamente utilizados en NLP: regresión logística, árboles de decisión, máquinas de vectores de soporte y redes neuronales. El alcance abarcó la evaluación de estos algoritmos utilizando métricas de evaluación estándar, así como la comparación de sus resultados y el análisis de sus ventajas y desventajas. El estudio pretende contribuir al conocimiento en el campo de la NLP y proporcionar información relevante para futuros trabajos y desarrollos en este ámbito.
**Palabras clave:** Calidad de la información, sistemas de información, sector salud.

**Abstract**
The present study focuses on the comparison of machine learning algorithms for natural language processing, specifically in classification and text analysis tasks. Four algorithms widely used in NLP will be analyzed: logistic regression, decision trees, support vector machines, and neural networks. The scope covered the evaluation of these algorithms using standard evaluation metrics, as well as the comparison of their results and the analysis of their advantages and disadvantages. The study aims to contribute to the knowledge in the field of NLP and provide relevant information for future work and developments in this area.
**Keywords:** Machine learning, natural language processing, algorithms.

REVISTA ELECTRÓNICA SOBRE TECNOLOGÍA, EDUCACIÓN Y SOCIEDAD
ISSN: 2448 - 6493
CTES

**Resumo**
Este estudo centra-se na comparação de algoritmos de aprendizagem automática para o processamento de linguagem natural, especificamente em tarefas de classificação e análise de texto. Serão analisados quatro algoritmos amplamente utilizados em PNL: regressão logística, árvores de decisão, máquinas de vectores de suporte e redes neuronais. O âmbito abrangeu a avaliação destes algoritmos utilizando métricas de avaliação padrão, bem como a comparação dos seus resultados e a análise das suas vantagens e desvantagens. O estudo visa contribuir para o conhecimento no domínio da PNL e fornecer informações relevantes para futuros trabalhos e desenvolvimentos nesta área.
**Palavras-chave:** Aprendizagem automática, processamento de linguagem natural, algoritmos.
**Fecha Recepción:** Junio 2023 **Fecha Aceptación:** Diciembre 2023

## Introducción

El procesamiento de lenguaje natural (PLN) es un campo de estudio que busca desarrollar algoritmos y técnicas que permitan a las computadoras comprender y analizar el lenguaje humano de manera automática. En este trabajo, se realizará una comparación de diferentes algoritmos de machine learning utilizados en el procesamiento de lenguaje natural. Estos algoritmos son ampliamente utilizados en diversas aplicaciones, como chatbots, análisis de sentimientos, traducción automática, entre otros. Es saber cuál tipo de algortitmo utilizar es fundamental para poder atacar de manera correcta el problema, la evaluación y comparación de estos algoritmos se llevó a cabo utilizando métricas específicas, y los resultados obtenidos fueron analizados y discutidos.

### Definición de machine learning

El machine learning, o aprendizaje automático, es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y realizar tareas sin ser programadas explícitamente (Sokolova & Lapalme, 2009). Se basa en la idea de que las máquinas pueden aprender a partir de datos y experiencias anteriores para mejorar su rendimiento en futuras tareas. Utilizando técnicas como el reconocimiento de patrones y la minería de datos, el machine learning ha demostrado ser una herramienta sumamente poderosa en la resolución de problemas complejos y en la toma de decisiones automatizada.

### Tipos de algoritmos de machine learning

Existen diversos tipos de algoritmos de machine learning (Linardatos et al., 2021) que se utilizan para abordar diferentes tipos de problemas. Algunos de los más comunes son: los algoritmos de clasificación, que se utilizan para agrupar datos en categorías o clases; los algoritmos de regresión, que predicen valores continuos basados en datos anteriores; los algoritmos de agrupamiento, que identifican patrones y similitudes en conjuntos de datos sin etiquetar; y los algoritmos de aprendizaje reforzado, que se basan en sistemas de recompensa para aprender cómo tomar decisiones óptimas en situaciones cambiantes (Mahesh & . 2020). La elección del algoritmo adecuado depende del problema en cuestión y de los datos disponibles para el entrenamiento.

**Aplicaciones del machine learning en el procesamiento de lenguaje natural**

El procesamiento de lenguaje natural (NLP) es una disciplina que se centra en la interacción entre las computadoras y el lenguaje humano. El machine learning ha revolucionado el NLP al permitir el análisis, comprensión y generación automática de texto. Algunas de las aplicaciones más destacadas del machine learning en el procesamiento de lenguaje natural incluyen: la traducción automática, el análisis de sentimiento, la extracción de información, la generación de resúmenes, el reconocimiento de voz y las respuestas automáticas en chatbots. Estas aplicaciones han mejorado significativamente la eficiencia y precisión en el tratamiento del lenguaje humano, facilitando la comunicación y la toma de decisiones en diferentes ámbitos.

**Algoritmos de machine learning para procesamiento de lenguaje natural**

Para el procesamiento de lenguaje natural, existen diversos algoritmos de machine learning que se utilizan ampliamente (Sancho Escrivá et al., 2020). Estos algoritmos son herramientas poderosas que permiten analizar y comprender el lenguaje humano de una manera automatizada. Al aplicar algoritmos de machine learning al procesamiento de lenguaje natural, es posible extraer información valiosa de grandes cantidades de texto, como patrones, sentimientos o intenciones. Estos algoritmos son fundamentales en diversas áreas, desde la traducción automática hasta la clasificación de documentos y el análisis de sentimientos.

**Regresión logística**

La regresión logística es un algoritmo de aprendizaje supervisado utilizado en el procesamiento de lenguaje natural. Se emplea para resolver problemas de clasificación binaria, donde se busca predecir una etiqueta discreta (Nick & Campbell, 2007). La regresión logística utiliza la función logística para modelar la relación entre las variables de entrada y la probabilidad de que una instancia pertenezca a una clase en particular. Este algoritmo es especialmente útil cuando se trabaja con datos textuales y se necesita realizar clasificaciones de manera eficiente y precisa.

**Random Forest**

Random Forest o Bosques Aleatorios es un algoritmo de aprendizaje automático que se utiliza ampliamente en el procesamiento de lenguaje natural (Hastie et al., 2009). Este algoritmo se basa en la combinación de múltiples árboles de decisión para realizar la tarea de clasificación o regresión. Cada árbol en el bosque se entrena con un subconjunto aleatorio de datos y características, lo que ayuda a evitar el sobreajuste. Random Forest es conocido por su capacidad para manejar conjuntos de datos complejos y de alta dimensionalidad, y su flexibilidad para lidiar con diferentes tipos de variables y problemas de clasificación (Schonlau & Zou, 2020). Su eficacia en el procesamiento de lenguaje natural se debe a su capacidad para capturar relaciones no lineales y detectar características importantes en los datos de texto (Chen et al., 2020). Random Forest distintos autores lo nombran como mejor que los demás clasificadores en todas las métricas de evaluación aplicadas (Younis et al., 2020).

**Árboles de decisión**

Los árboles de decisión son algoritmos de machine learning que se utilizan para clasificar instancias basándose en una serie de preguntas o condiciones lógicas (de Ville, 2013). Estos árboles se construyen de manera jerárquica, donde cada nodo representa una pregunta y cada rama representa una posible respuesta. Los árboles de decisión son especialmente útiles en el procesamiento de lenguaje natural debido a su interpretabilidad y capacidad para manejar tanto variables categóricas como continuas (Charbuty et al., 2021). Permiten realizar clasificaciones precisas y comprensibles en tareas como la categorización de textos o la detección de spam en correos electrónicos.

**Naive Bayes**

Naive Bayes es un algoritmo de machine learning que se basa en el teorema de Bayes para realizar clasificaciones probabilísticas. Aunque parte de una suposición simplificada y "ingenua" (de ahí su nombre), Naive Bayes es ampliamente utilizado en el procesamiento de lenguaje natural debido a su eficiencia y buen desempeño en tareas como la clasificación de textos (Chen et al., 2020). Este algoritmo se basa en el cálculo de probabilidades condicionales para determinar la clase más probable para una instancia dada. Naive Bayes es especialmente útil en el análisis de sentimientos, la categorización de noticias y la detección de spam (Zhang & Li, 2007).

## Metodología

La metodología de comparación utilizada consistió en evaluar diferentes algoritmos de machine learning para el procesamiento de lenguaje natural, esta metodología permitió determinar cuál algoritmo es más efectivo en términos de precisión y rendimiento para tareas de procesamiento de lenguaje natural.

Se llevaron a cabo experimentos utilizando conjuntos de datos extraídos de “twitter” o ahora llamado “X” y se compararon los resultados obtenidos por cada algoritmo, se utilizó una máquina Huawei MateBook D15 con sistema operativo Windows y el entorno de desarrollo “Jupiter” para realizar las ejecuciones de las pruebas de cada uno de los algoritmos.

Para la recopilación de datos y el poder manipularlos, se utilizó un fichero en formato CSV y algunos de los campos del fichero se encuentran presentados en la tabla número uno.

**Tabla 1.** Tweets etiquetados por neutrales o agresiones

| Id | Label | Tweet |
|---|---|---|
| 100705 | 1 | Hey retard you dont get those two put… |
| 154121 | 1 | The graceful slick is non other than an ungra… |
| 19493 | 1 | You might be a libtard if libtard sjw libera… |
| 135711 | 1 | Controversial remarks alleged called facebook… |
| 108421 | 1 | Go screw contrary to your recent message… |
| 141273 | 0 | Section 1b of the gacr states it complies wi… |
| 9372 | 0 | Say tim could you take a look at the e… |
| 93553 | 0 | Ahcene bendjazia well dont get me started the… |
| 94735 | 0 | No i dont think you mean to be obnoxious I th… |
| 83209 | 0 | Basically what he is trying to do is disamb… |

Fuente: Elaboración propia

Los datos se etiquetaron de manera que los que tienen un número 0 con comentarios neutrales y lo que tienen asignado un numero 1 son agresiones.

Para las pruebas y ejecuciones se utilizaron un total de 159,686 datos ya etiquetados y divididos en dos grupos, de entrenamiento y de pruebas.

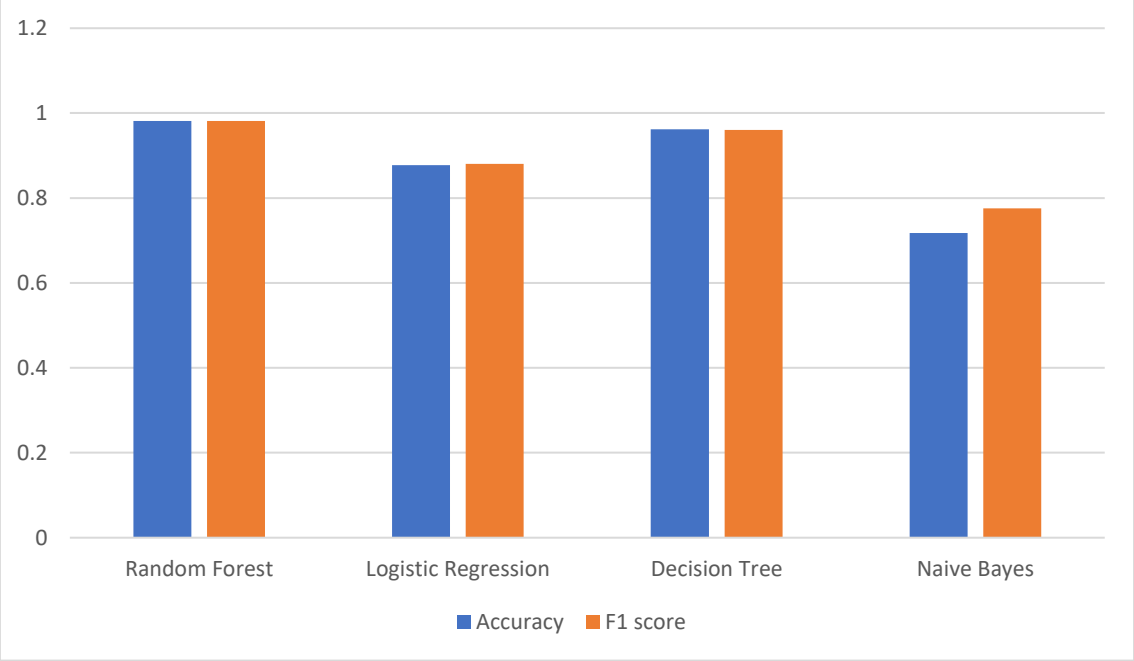
La evaluación de los algoritmos de machine learning para procesamiento de lenguaje natural se llevó a cabo mediante la comparación de diferentes métricas de evaluación.

Estas métricas permitieron medir el rendimiento y la eficacia de los algoritmos en la tarea específica. Los algoritmos fueron evaluados en base a su precisión, exhaustividad, puntuación F1 y exactitud. Estas métricas proporcionaron una visión completa de la capacidad de los algoritmos para procesar y comprender el lenguaje natural.

## Resultados

En la tabla 1 se muestran los resultados de las ejecuciones realizadas para observar la detección de agresiones en los tweets, clasificándolos como agresión o sin agresión, en la tabla se muestra el puntaje que obtuvieron en un rango de 0 a 1.

**Figura 1.** Resultados de ejecuciones al clasificar los tweets




Fuente: Elaboración propia

Los resultados presentados mostraron el rendimiento de diferentes modelos de aprendizaje automático en la tarea de clasificación. Comenzando con el modelo de Random Forest, observamos una accuracy del 98.17%, esta métrica es fundamental ya que representa la capacidad del modelo para realizar predicciones precisas y confiables. Además, la alta puntuación F1 de 0.9813 muestra que el modelo de Random Forest logra un equilibrio óptimo entre precisión y exhaustividad, lo que es esencial para evaluar su rendimiento de manera integral y fiable.

Por otro lado, el modelo de Regresión Logística muestra una precisión sólida del 87.74%, de las muestras clasificadas correctamente por este modelo y la puntuación F1 de 0.8805 indica que el modelo de Regresión Logística logra un equilibrio satisfactorio entre precisión y exhaustividad en la clasificación de las clases.

El modelo de Árbol de Decisión también mostró un rendimiento destacado, con una precisión del 96.22% y una puntuación F1 de 0.9606. Estas métricas indican que el modelo de Árbol de Decisión logró un equilibrio excelente entre precisión y exhaustividad, lo que es fundamental para garantizar predicciones precisas y completas en diferentes escenarios.

En contraste, el modelo de Naive Bayes mostró una precisión significativamente inferior del 71.75% y una puntuación F1 relativamente baja de 0.7755. Estas métricas sugieren que el modelo de Naive Bayes puede estar enfrentando dificultades para encontrar un equilibrio adecuado entre precisión y exhaustividad en la clasificación de las clases.

## Discusión

Al realizar las ejecuciones se puede observar cual algoritmo es mejor y más eficiente, aunque 3 de ellos tuvieron un buen rendimiento al realizar las pruebas, la elección de cual utilizar puede variar dependiendo el caso de cada tipo de problema que se requiera abordar, cabe aclarar que hay distintos autores que se pueden basar en la complejidad del algoritmo ya que unos son más sencillos de manipular que otros. Comparando los resultados con los de distintos autores se puede concluir que el algoritmo de random forest es el que mejor se comporta al clasificar datos de este tipo (Younis et al., 2020).

Como distintos autores mencionan, Random Forest tiene la mayor precisión y puntuación F1, lo que indica que es el modelo más efectivo en este conjunto de datos para la tarea específica que se está evaluando. El alto valor de precisión sugiere que este modelo tiene una baja tasa de falsos positivos y falsos negativos, mientras que el alto valor de F1 indica un buen equilibrio entre precisión y recuperación.

Logistic Regression muestra una precisión y puntuación F1 sólidas, aunque ligeramente más bajas que Random Forest. Aunque tiene un buen rendimiento general, parece haber una diferencia notable en la precisión y el F1 score en comparación con Random Forest.

Decision Tree también muestra un rendimiento sólido, con una precisión y puntuación F1 cercanas a Random Forest. Esto sugiere que Decision Tree también es una buena opción para este conjunto de datos y tarea específica.

Naive Bayes tiene la precisión más baja y una puntuación F1 significativamente más baja en comparación con los otros modelos. Esto podría indicar que Naive Bayes no es tan efectivo en este conjunto de datos en particular o que la naturaleza del problema requiere un enfoque más complejo que el proporcionado por Naive Bayes.

## Conclusiones

En general, la elección del modelo más adecuado dependerá de varios factores, como la naturaleza específica del problema, la disponibilidad de datos, la interpretabilidad del modelo y los requisitos de rendimiento. Los resultados presentados sugieren que Random Forest es el modelo más efectivo en términos de precisión y puntuación F1 para este conjunto de datos en particular, pero es importante considerar otros factores antes de tomar una decisión final sobre qué modelo usar.

## Futuras líneas de investigación

Para la clasificación de datos existen diversas formas de hacerlo, en este manuscrito se presentan 4 de las principales, pero se pueden realizar distintas comparaciones con otros tipos de clasificadores que pueden ser más complejos de manipularlos, pero los resultados pueden ser similares o mejores.

**Referencias**

Charbuty, B., Abdulazeez, A. J. J. o. A. S., & Trends, T. (2021). Classification based on decision tree algorithm for machine learning. *2*(01), 20-28.

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*, 105361. https://doi.org/10.1016/j.knosys.2019.105361

de Ville, B. (2013). Decision trees. *WIREs Computational Statistics*, *5*(6), 448-455. https://doi.org/10.1002/wics.1278

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., Friedman, J. J. T. e. o. s. l. D. m., inference,, & prediction. (2009). Random forests. 587-604.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1).

Mahesh, B. J. I. J. o. S., & ., R. (2020). Machine learning algorithms-a review. *9*(1), 381-386.

Nick, T. G., & Campbell, K. M. (2007). Logistic regression. *Methods Mol Biol*, *404*, 273-301. https://doi.org/10.1007/978-1-59745-530-5_14

Sancho Escrivá, J. V., Fanjul, C., de la Iglesia Vayá, M., Montell, J. A., & Escarti, M. J. (2020). Aplicación de la inteligencia artificial con procesamiento del lenguaje natural para textos de investigación cualitativa en la relación médico-paciente con enfermedad mental mediante el uso de tecnologías móviles.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting communications on statistics and Stata*, *20*(1), 3-29. https://doi.org/10.1177/1536867x20909688

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Younis, U., Asghar, M. Z., Khan, A., Khan, A., Iqbal, J., & Jillani, N. (2020). Applying Machine Learning Techniques for Performing Comparative Opinion Mining. *Open Computer Science*, *10*(1), 461-477. https://doi.org/10.1515/comp-2020-0148 (Open Computer Science)

Zhang, H., & Li, D. (2007, 2-4 Nov. 2007). Naïve Bayes Text Classifier. 2007 IEEE International Conference on Granular Computing (GRC 2007),

| Rol de Contribución | Autor (es) |
|---|---|
| Conceptualización | Yamil Emanuel Castro Solis |
| Metodología | Yamil Emanuel Castro Solis |
| Software | Yamil Emanuel Castro Solis |
| Validación | Yamil Emanuel Castro Solis |
| Análisis Formal | Yamil Emanuel Castro Solis |
| Investigación | Yamil Emanuel Castro Solis |
| Recursos | Yamil Emanuel Castro Solis |
| Curación de datos | Yamil Emanuel Castro Solis |
| Escritura - Preparación del borrador original | Yamil Emanuel Castro Solis |
| Escritura - Revisión y edición | Hiram Herrera Rivas |
| Visualización | Yamil Emanuel Castro Solis |
| Supervisión | Hiram Herrera Rivas |
| Administración de Proyectos | Hiram Herrera Rivas |
| Adquisición de fondos | Yamil Emanuel Castro Solis |