Evaluación de Algoritmos de Aprendizaje Supervisado usando Modelos Binarios para Clasificación de Análisis de Sentimiento
Evaluation of Supervised Learning Algorithms Using Binary Models for Sentiment Analysis Classification

Emmanuel Morales García[1] Cecilia Cruz López[2] Julia Aurora Montano Rivas[3]
Diana Laura Aguirre Capistran[4]

Universidad Veracruzana, Facultad de Estadística e Informática, Av. Xalapa S/N esq. Av. Manuel Ávila Camacho, Col. Del Maestro, Xalapa, Ver., 91020. México
[1] emmorales@uv.mx, [2] ceccruz@uv.mx, [3] julmontano@uv.mx
[4] diana99.capistran@gmail.com

Fecha de recepción: 18 de noviembre de 2023
Fecha de aceptación: 27 de abril de 2024

Resumen. Este estudio tuvo como propósito evaluar algoritmos de aprendizaje supervisado en modelos binarios para mejorar el análisis de sentimiento en la clasificación de datos no estructurados. Se analizaron datos de diversas áreas temáticas, desde ciencias sociales hasta ciencias naturales, con diferentes dimensiones en cada área, reflejando la variabilidad y cantidad de datos recopilados. Los modelos de aprendizaje supervisado lograron altos niveles de precisión, destacándose el modelo análisis discriminante lineal (LDA) como el mejor clasificador en términos de precisión y ROC. Sin embargo, la sensibilidad y especificidad variaron entre modelos. El análisis de sentimientos reveló que predominaba el sentimiento positivo en los datos, respaldado por un conjunto significativo de palabras. Aunque el modelo LDA se mostró idóneo para clasificar los datos, se enfatiza la importancia de considerar el equilibrio entre precisión, sensibilidad y especificidad según los objetivos específicos y la relevancia de falsos positivos y falsos negativos en un contexto particular.

Palabras clave: Máquinas de soporte vectorial, Naive Bayes, Regresión logistica binaria, Árboles de Decisipon, Análisis discriminante lineal.

Summary. This study aimed to evaluate supervised learning algorithms on binary models to improve sentiment analysis in the classification of unstructured data. Data from various subject areas were analyzed, from social sciences to natural sciences, with different dimensions in each area, reflecting the variability and amount of data collected. The supervised learning models achieved high levels of accuracy, with the linear discriminant analysis (LDA) model standing out as the best classifier in terms of accuracy and ROC. However, sensitivity and specificity varied between models. Sentiment analysis revealed that positive sentiment predominated in the data, supported by a significant set of words. Although the LDA model was suitable for classifying the data, the importance of considering the balance between precision, sensitivity and specificity according to the specific objectives and the relevance of false positives and false negatives in a particular context is emphasized.

Keywords: Support vector machines, Naive Bayes, Binary logistic regression, Decisipon trees, Linear discriminant analysis.

1 Introducción

El análisis de sentimiento se ha convertido en un tema relevante dentro del marco de la evaluación de algoritmos de aprendizaje supervisado que se sitúa en el ámbito de la minería de textos y el procesamiento del lenguaje natural, ya que permite comprender la opinión y las emociones expresadas en el texto, lo que tiene aplicaciones en la toma de decisiones empresariales, la retroalimentación del usuario y la detección de tendencias en las redes sociales.

Se basa en la premisa de que la clasificación de sentimiento es una tarea fundamental en el análisis de sentimiento y que los algoritmos de aprendizaje supervisado desempeñan un papel crucial en la construcción de modelos de clasificación precisos. Estos algoritmos, que incluyen técnicas como máquinas de soporte vectorial, Naive Bayes, regresión logística, árboles de decisión y el análisis discriminante lineal, se entrenan utilizando conjuntos de datos etiquetados previamente, lo que les permite aprender patrones y relaciones entre las características del texto y las etiquetas de sentimiento asociadas.

Las máquinas de soporte vectorial (SVM) son efectivos en la clasificación de datos mediante la búsqueda de un hiperplano óptimo que maximiza la separación entre diferentes clases. Son particularmente útiles en problemas

de clasificación lineal y no lineal. En cuanto al algoritmo Naive Bayes basado en el teorema de Bayes, es ampliamente utilizado en problemas de clasificación de texto y análisis de sentimientos. Es especialmente útil cuando se trata de datos con alta dimensionalidad y se utiliza en aplicaciones como la detección de spam y el análisis de texto. También, la regresión logística es un algoritmo que se utiliza principalmente para problemas de clasificación binaria. Modela la probabilidad de que un ejemplo pertenezca a una clase específica y se usa comúnmente en problemas médicos, de marketing y financieros. Asimismo, los árboles de decisión se utilizan para la toma de decisiones basadas en reglas lógicas. Son especialmente útiles en problemas de clasificación y regresión, y su estructura en forma de árbol es fácilmente interpretable. Finalmente, el análisis discriminante lineal (LDA) es eficaz para proporcionar un modelo con el cual se logre una clasificación de individuos [1]. Por lo tanto, estos algoritmos son fundamentales en la clasificación de datos en una variedad de aplicaciones, en esta investigación son usados para clasificación de texto a través de Twitters y poder realizar el análisis de sentimientos.

El contexto también considera el desafío de evaluar la eficacia de estos algoritmos, ya que la precisión en la clasificación de sentimiento es esencial para su aplicación en situaciones del mundo real. La evaluación implica el uso de métricas como exactitud, sensibilidad, especificidad, F-score, matriz de confusión y la curva ROC para medir el rendimiento de los modelos [2]. Además, se deben tener en cuenta aspectos como el desequilibrio de clases, el preprocesamiento de texto y la selección de características para garantizar resultados confiables [3].

El objetivo de este estudio fue evaluar estos algoritmos de aprendizaje supervisado en modelos binarios que coadyuven al análisis de sentimiento para una clasificación correcta de datos no estructurados.

2 Estado del arte

Algunos antecedentes de este enfoque se sitúan en [4] cuyo trabajo se orienta hacia la expansión de la investigación en respuestas a preguntas en una dirección distinta, abordar tareas que involucran múltiples perspectivas y que requieren la habilidad de identificar y estructurar opiniones dentro de un texto específico. Este estudio propone un método para responder preguntas desde una óptica centrada en la extracción de información basada en opiniones. Además, esboza una estrategia para la generación automática de resúmenes basados en opiniones y describe cómo estos resúmenes pueden ser empleados para respaldar diversas tareas relacionadas con la respuesta a preguntas de múltiples perspectivas concluyen con un breve análisis acerca de cómo las representaciones concisas basadas en opiniones podrían ser aplicadas para respaldar diversas tareas relacionadas con la respuesta a preguntas desde diferentes puntos de vista

De igual manera, [5] elaboraron una metodología para extraer el sentimiento de los inversores minoristas a partir de los mensajes publicados en foros sobre acciones. Este algoritmo está compuesto por varios clasificadores que trabajan en conjunto mediante un sistema de votación. Los niveles de precisión alcanzados son similares a los de los clasificadores de Bayes, pero con una tasa menor de resultados falsos positivos y una mayor precisión en la detección del sentimiento. La inclusión de series de tiempo y la consolidación de la información de los mensajes enriquecen la calidad del índice de sentimiento resultante, especialmente cuando se enfrenta a lenguaje coloquial y ambigüedad. Las aplicaciones prácticas de esta metodología revelan una correlación con los valores de los mercados de acciones, específicamente en el caso del sentimiento agregado en el sector tecnológico, que puede predecir los niveles de los índices bursátiles, aunque no a nivel de acciones individuales. Se concluyó que el algoritmo tiene el potencial de evaluar cómo los anuncios de gestión, comunicados de prensa y noticias de terceros impactan en la opinión de los inversores.

En otra investigación realizada por [6] detallan la participación del equipo de investigación ELiRF de la Universidad Politécnica de Valencia en el Taller sobre Análisis de Sentimientos (TASS-2013). El TASS-2013, planteó cuatro desafiantes tareas, la evaluación del sentimiento global en tweets, la identificación de los temas en los tweets (política, economía, deportes, etc.), el análisis del sentimiento a nivel de entidad dentro de los tweets, y la determinación de las inclinaciones políticas (derecha, centro, izquierda, neutral) de los usuarios basándose en sus publicaciones. En este informe presentó estrategias metodológicas empleadas, los logros alcanzados y un análisis minucioso de los mismos. Se usaron técnicas de aprendizaje automático, específicamente, la aproximación de máquinas de soporte vectorial (SVM), apoyadas por la herramienta WEKA y la librería externa LibSVM. Los resultados obtenidos se mantuvieron en su mayoría a la par o incluso superiores en algunos casos en comparación con los equipos líderes en la competición. Un desafío recurrente en todas las tareas del TASS-2013 fue lograr una correcta tokenización de los tweets, por lo que proponen que la organización considere proporcionar tweets ya tokenizados en futuras ediciones para facilitar la comparación de enfoques, sin que esto afecte la evaluación, centrándose exclusivamente en los métodos y características utilizados en la resolución de los problemas.

En [7] los usuarios buscan simplificar el proceso de revisar cada tweet y recopilar datos, así como tienen la habilidad de identificar las opiniones expresadas por otros usuarios en diversas publicaciones. Las publicaciones se sometieron a un proceso de procesamiento para eliminar elementos que no aportaban a la predicción del sentimiento. Se utilizaron algoritmos de aprendizaje automático para categorizar los textos de los tweets en categorías positivas o negativas. Estos algoritmos permitieron un análisis detallado de los datos de salida con el objetivo de comprender mejor las publicaciones de los usuarios y su inclinación hacia una emoción específica.

3 Métodología

Para corroborar y verificar el objetivo de la investigación se desarrollaron algoritmos de clasificación binaria (métodos de aprendizaje supervisado). A continuación, se muestra el diagrama en el que se puede identificar visualmente la metodología propuesta.

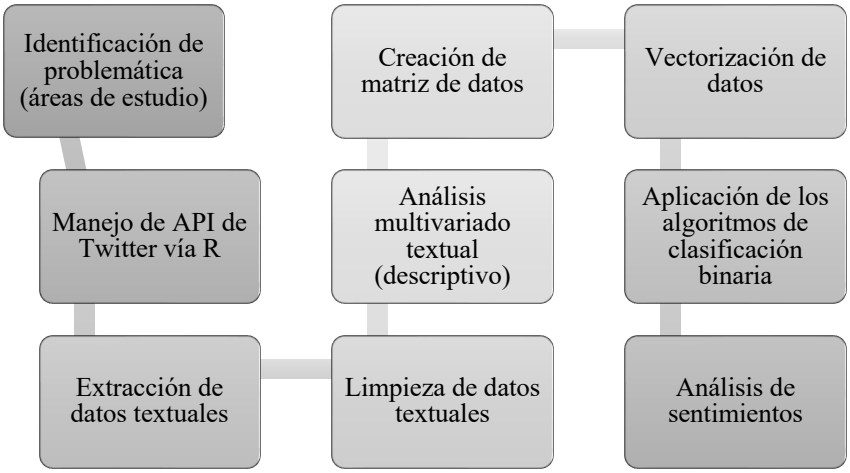

Figura 1. Esquema de propuesta metodológica.

3.1 Descripción de la propuesta metodológica

En este apartado se abordaron los temas de interés en diversas áreas de estudio como, ciencias sociales y humanidades específicamente en la aplicación sobre violencia (derechos de las mujeres y protección), temas como música, arte, lectura, etc. Asimismo, en ciencias naturales se usó la aplicación en el estudio de los animales. También, en ciencias de la salud usando una aplicación sobre la Covid-19, así como en ciencias exactas observando datos sobre evaluación de las publicaciones con impacto estadístico y matemático durante la pandemia, y finalmente cuestiones políticas, la evaluación sobre comentarios con relación al presidente de México Andrés Manuel López Obrador.

El proceso comienza con la extracción de datos textuales de un API de Twitter, para lo cual se requieren credenciales de Twitter y la configuración de una cuenta de desarrollador. Se extrajeron 5000 tweets por cada tema de los mencionados anteriormente. Luego, se llevó a cabo la limpieza de datos, que incluyó la eliminación de enlaces, símbolos, emojis y palabras vacías que no aportaban información relevante. Después de la limpieza, se realizó un análisis multivariado textual descriptivo que implicó la creación de visualizaciones de datos cualitativos, como gráficos de barras y nubes de palabras, para comprender el comportamiento de los datos no estructurados. Posteriormente, se vectorizaron los datos textuales, lo que implicó la creación de una matriz de documentos-términos y la conversión de ésta en frecuencias de ocurrencia de términos en la colección de documentos. Finalmente, se prepararon los datos para su procesamiento con algoritmos de clasificación binaria.

3.2 Algoritmos de clasificación binaria

Se aplicaron diversos algoritmos de clasificación binaria, como máquinas de soporte vectorial, Naive Bayes, regresión logística binaria, árboles de decisión y análisis discriminante lineal (LDA, por sus siglas en inglés). Se utilizaron métricas como exactitud, sensibilidad, especificidad, F-score, matriz de confusión y la curva ROC para evaluar estos métodos. Finalmente, se llevó a cabo un análisis de sentimiento, donde se visualizaron los datos

procesados por el algoritmo que demostró mejor eficiencia en los pasos anteriores. Todo lo anterior se programó a través del software R-Project.

4 Resultados

Como se mencionó anteriormente se tomaron diversas temáticas para crear las bases de datos, una vez construidas se enlistan las variables que se tomaron en cuenta en cada aplicación.

1. Texto de los tweets por cada usuario.
2. Nombre de los usuarios que publicaron los tweets.
3. Dispositivo fue publicado el tweet.
4. Lo que contiene cada tweet.
5. Cuantas compartidas tuvo el tweet.
6. Lugar de donde se compartió el tweet.

Ahora se visualiza la tabla de dimensiones por cada área temática evaluada, se consideran las palabras limpias y con estos datos se pueden lograr crear las matrices binarias.

Tabla 1. Descripción de los datos y sus dimensiones.

Áreas temáticas	Dimensión
Ciencias Social	750
Ciencias de la Salud	411
Ciencias Naturales	349
Ciencias Exactas	233
Ciencias Políticas	290

Con los resultados de la Tabla 1, se procede al entrenamiento de los modelos de aprendizaje supervisado, dividir los datos en entrenamiento y prueba.

4.1 Evaluación de los modelos de clasificación binaria

Tabla 2. Métrica para evaluar los modelos binarios.

Modelos	Exactitud	Especificidad	Sendibilidad	F score	ROC
SVM	88.4 %	0.89	0.69	0.41	66 %
Naive Bayes	88.5 %	0.97	0.24	0.35	60 %
Regresión L.	89.3 %	0.98	0.26	0.39	72 %
Árboles D.	89.0 %	0.98	0.26	0.38	64 %
LDA	90.6 %	0.97	0.32	0.42	70 %

Dados los resultados de los algoritmos binarios se hizo uso de las estadísticas generales, proporcionadas por la matriz de confusión, como resultados se puede observar, que el LDA, obtuvo una precisión de 90.6 %, una especificidad de 0.97 % que representa la proporción de palabras correctamente clasificadas (verdaderos negativos) y una sensibilidad de 0.32 % que es la proporción de palabras correctamente clasificadas (verdaderos positivos), siendo evaluada con un 70 % (ROC) como buen clasificador para esta aplicación. En caso contrario, Naive Bayes es el que presenta estadísticas, menos confiables para una clasificación. Asimismo, la regresión logística, aunque sea el mejor evaluado para la clasificación, de acuerdo con el resultado del área bajo la curva, el porcentaje de precisión fue menor que LDA (Tabla 2).

4.2 Análisis de sentimiento

Para finalizar, se obtuvo el análisis de sentimiento, para los datos binarios, LDA fue evaluado como el mejor algoritmo; es decir, presentó mejor precisión de clasificación. De esta manera se observa el sentimiento que predomino fue el positivo con 1,925 palabras (Figura 2).

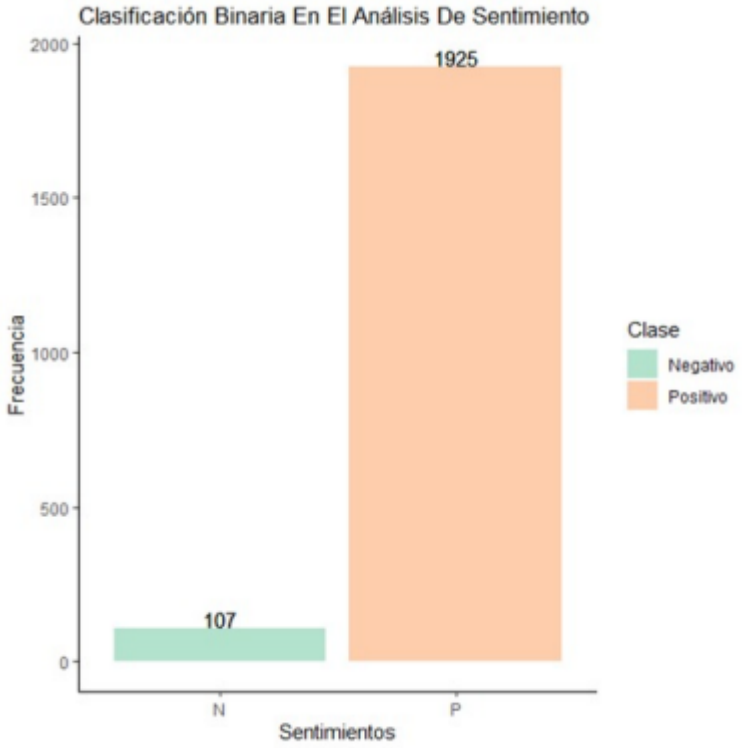

Figura 2. Análisis de sentimientos (resultados LDA).

5 Conclusiones y direcciones para futuras investigaciones

Se analizaron datos de diversas áreas temáticas, abarcando campos que iban desde ciencias sociales y salud hasta políticas y ciencias naturales. Cada área temática se caracterizó por tener un número distinto de dimensiones, lo que reflejó la variabilidad y la cantidad de datos recopilados en el estudio. Se utilizaron modelos de aprendizaje supervisado para llevar a cabo la clasificación binaria de estos datos, y los resultados mostraron un rendimiento generalmente alto. En particular, el modelo LDA se destacó como el mejor clasificador en términos de precisión y área bajo la curva (ROC). Se observó que la especificidad y la sensibilidad variaron entre los modelos, lo que sugirió que el modelo LDA equilibró de manera más efectiva la capacidad de identificar verdaderos positivos y verdaderos negativos. Por otro lado, Naive Bayes mostró una alta precisión en general pero una baja sensibilidad, lo que indicó una tendencia a identificar menos verdaderos positivos, posiblemente indicando limitaciones en su capacidad para identificar casos positivos reales en los datos. Aunque la regresión logística tenía un ROC más alto que LDA, mostró una precisión ligeramente inferior, lo que sugiere que pudo haber sido más conservadora en sus clasificaciones, minimizando los errores, pero a costa de clasificar menos casos como positivos.

En relación con el análisis de sentimientos realizado con el modelo LDA, se encontró que el sentimiento predominante en los datos fue positivo, respaldado por 1,925 palabras. Este hallazgo es de gran relevancia, ya que indica que la mayoría de los comentarios o tweets en los datos se clasifican como positivos según el enfoque de clasificación utilizado. Estos resultados sugieren que el modelo LDA es una elección adecuada para clasificar los datos de la investigación, especialmente en aplicaciones relacionadas con el análisis de sentimientos. Sin embargo, es esencial considerar el equilibrio entre precisión, sensibilidad y especificidad según los objetivos específicos y la importancia relativa de los falsos positivos y falsos negativos en un contexto particular.

Para futuras investigaciones, se planea implementar una variedad de algoritmos adicionales que aborden tanto la clasificación binaria como la clasificación multiclase, incluyendo campos como aleatorios de Márkov, redes bayesianas y clasificación multiclase bayesiana, entre otros. Además, se propone el desarrollo de una interfaz gráfica en Python que facilite el procesamiento de este tipo de datos y métodos, lo que podría contribuir significativamente a la eficacia y accesibilidad de este enfoque en aplicaciones futuras.

Referencias

[1] F. Vanhoenshoven, G. Napoles, R., Falcon, K. Vanhoof, and M. Koppen, Detecting malicious URLs using machine learning techniques. IEEE Symposium Series on *Computational Intelligence* (SSCI). doi:10.1109/ssci.2016.7850079, 2016.

[2] K. A. Carvajal Jaramillo. Aplicación de modelos de aprendizaje supervisado para predicción del tipo de contacto de clientes asignados a un BPO de cobranza. Tesis de Especialidad. Especialidad en Estadística Aplicada. Fundación Universitaria Los Libertadores. Bogotá, Colombia, 2022.

[3] M. M. Loja Paucar Desarrollo de un prototipo para lectura y registro automático de información de visitas en puntos de control de acceso a un establecimiento. Tesis de grado, Ingeniería en Electrónica, Automatización y Control, Universidad de las Fuerzas Armadas, 2023.

[4] C. Cardied, J. Wiebe, T. Wilson and D.J. Litman, Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering. *In New directions in question answering* (pp. 20-27), AAAI Technical Report SS-03-07, 2003.

[5] S. Das and M. Chen, Yahoo! para Amazon: Extracción de sentimientos de Small Talk en la Web, *Ciencias de la gestión* 53 (9), 1375-1388, 2001.

[6] F. Pla and L. F. Hurtado, Análisis de sentimientos en Twitter. In *XXIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* (SEPLN 2013). TASS (pp. 220-227), 2013.

[7] J. Ramon, A. Reyes y P. Palos, Un aanálisis de sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de #BlackFriday, *Espacios,* 39, 16, 2018.