# Diseño de Experimentos para la Comparación de Modelos de Clasificación de Textos en Noticias Reales y Falsas

Álvaro Salgado López

## I. Introducción

La proliferación de noticias falsas en medios digitales ha generado la necesidad de desarrollar métodos automáticos para su detección. Dentro del procesamiento de lenguaje natural (NLP), el aprendizaje automático ofrece herramientas eficientes para la clasificación de textos, permitiendo diferenciar entre información veraz y engañosa con alta precisión.

Este estudio tiene como objetivo comparar distintos modelos de clasificación de texto y evaluar el impacto de sus hiperparámetros en el desempeño de la detección de noticias falsas. Para ello, se utilizó el conjunto de datos "Fake and Real News" de Kaggle, compuesto por noticias etiquetadas como reales o falsas.

En el diseño del experimento, se implementaron varios algoritmos de clasificación, incluyendo Naive Bayes, Regresión Logística y Máquinas de Vectores de Soporte (SVM). Para optimizar los hiperparámetros se empleó GridSearchCV, un método de búsqueda exhaustiva basado en validación cruzada, lo que permitió identificar las configuraciones óptimas para cada clasificador.

El desempeño de los modelos se evaluó utilizando métricas como precisión, recall, F1-score y accuracy, además de analizar la matriz de confusión para interpretar los errores de clasificación. Los resultados de este análisis permitirán determinar qué modelo y combinación de hiperparámetros ofrece la mejor solución para la detección automática de noticias falsas.

## II. Metodología

En esta sección se describe el procedimiento seguido para la clasificación de noticias falsas y reales, desde la preparación de los datos hasta la evaluación de los modelos empleados.

### A. Preprocesamiento de los Datos

El conjunto de datos utilizado proviene de la plataforma Kaggle y contiene noticias etiquetadas como reales o falsas. Antes de aplicar los modelos de clasificación, se realizó un preprocesamiento del texto con el objetivo de mejorar la calidad de los datos y eliminar elementos que pudieran introducir ruido en el análisis.

Las etapas del preprocesamiento fueron las siguientes:

- Normalización del texto: Conversión de todos los caracteres a minúsculas para evitar discrepancias entre palabras con diferentes capitalizaciones.
- Eliminación de caracteres especiales: Se eliminaron signos de puntuación y otros símbolos no alfabéticos.
- Construcción del corpus de texto: Se concatenaron los títulos y el contenido de las noticias en una única columna, con el propósito de preservar la información de ambas fuentes.

Una vez limpiado el texto, se aplicó una técnica de representación numérica de los datos. Para ello, se utilizó la técnica TF-IDF (Term Frequency-Inverse Document Frequency), la cual asigna pesos a cada palabra en función de su frecuencia en los documentos y en el conjunto total de datos. En este caso, se estableció un límite de 5000 características y se eliminaron las palabras vacías en inglés para reducir la dimensionalidad del espacio de características.

### B. División del Conjunto de Datos

Para evaluar el desempeño de los modelos, se realizó una partición del conjunto de datos en 80 % para entrenamiento y 20 % para prueba. Se utilizó una estrategia de división estratificada para garantizar que la distribución de clases en cada subconjunto reflejara la del conjunto original.

### C. Modelos de Clasificación

Se implementaron y compararon diversos modelos de clasificación con el objetivo de determinar cuál proporciona el mejor desempeño en la tarea de detección de noticias falsas. Los modelos evaluados fueron los siguientes:

- Naive Bayes Multinomial
- Regresión Logística
- Máquinas de Soporte Vectorial (SVM)

Cada uno de estos modelos fue entrenado utilizando el conjunto de datos preprocesado y vectorizado con TF-IDF.

### D. Ajuste de Hiperparámetros

En particular, para la Regresión Logística, se realizó una búsqueda aleatoria con los siguientes hiperparámetros:

- C: Parámetro de regularización con valores extraídos de una distribución uniforme en el intervalo [0.1, 10].
- Penalty: Penalización L1 y L2 para regularización.
- Solver: Se utilizó liblinear para permitir el uso de la penalización L1.
- Validación cruzada: Se empleó un esquema de validación de 5 particiones (5-fold cross-validation).

5. Evaluación de los Modelos

Los modelos fueron evaluados con base en las siguientes métricas:

- Precisión (Accuracy): Proporción de predicciones correctas en el conjunto de prueba.
- Matriz de confusión: Representación de los errores de clasificación en términos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.
- Puntajes de Precision, Recall y F1-score: Medidas de desempeño para evaluar la capacidad del modelo de distinguir entre noticias falsas y reales.
- Análisis de palabras más importantes: Se identificaron las palabras más relevantes para la clasificación de noticias falsas y reales a través del análisis de los pesos asignados por el vectorizador TF-IDF.
- Conteo de palabras: Se realizaron análisis de frecuencia para comparar las palabras más comunes en noticias falsas y reales. A partir de estos conteos, se generaron representaciones visuales mediante nubes de palabras y gráficos de barras para facilitar la interpretación.

## III. Resultados

En este estudio se compararon tres modelos de clasificación de texto para detectar noticias falsas, utilizando el conjunto de datos de noticias falsas y reales disponible en Kaggle. Los modelos evaluados fueron Naive Bayes, Regresión Logística y Support Vector Machine (SVM). Además, se utilizó la técnica de RandomizedSearchCV para optimizar los hiperparámetros del modelo de Regresión Logística.

### A. Desempeño de los Modelos

Los tres modelos fueron evaluados utilizando las métricas de precisión (accuracy), recall, f1-score y la matriz de confusión. A continuación, en la Tabla I se presentan los resultados de cada modelo:

| Modelo | Precisión | Recall | F1-score |
|---|---|---|---|
| **Naive Bayes** | 0.93 | 0.93 | 0.93 |
| **Noticias Reales (0)** | 0.93 | 0.93 | 0.93 |
| **Noticias Falsas (1)** | 0.94 | 0.94 | 0.94 |
| **Regresión Logística** | 0.99 | 0.99 | 0.99 |
| **Noticias Reales (0)** | 0.98 | 0.99 | 0.99 |
| **Noticias Falsas (1)** | 0.99 | 0.98 | 0.99 |
| **SVM** | 0.99 | 0.99 | 0.99 |
| **Noticias Reales (0)** | 0.99 | 1.00 | 0.99 |
| **Noticias Falsas (1)** | 1.00 | 0.99 | 0.99 |

TABLE I
RESULTADOS DE PRECISIÓN, RECALL Y F1-SCORE DE LOS MODELOS EVALUADOS.

Naive Bayes alcanzó una precisión de 0.93 en ambas clases (noticias verdaderas y falsas), mostrando una buena capacidad para manejar la dispersión de las clases en el conjunto de datos. Este modelo es conocido por ser eficiente y rápido, especialmente en tareas de clasificación de texto, y su rendimiento en este caso no fue la excepción.

Por otro lado, Regresión Logística alcanzó una precisión significativamente más alta de 0.99, tanto en noticias verdaderas como falsas. Este modelo mostró una alta capacidad de generalización y una excelente distinción entre las clases, lo que lo posicionó como una opción muy sólida para la tarea de clasificación de noticias.

Máquinas de Vectores de Soporte (SVM) también mostró un desempeño sobresaliente con una precisión de 0.99, sin embargo, se observó que el tiempo de entrenamiento fue considerablemente mayor en comparación con los otros modelos. Este comportamiento se debe a la complejidad inherente al entrenamiento de SVM, especialmente cuando se trabaja con grandes volúmenes de datos y una cantidad significativa de características, como es el caso del conjunto de datos utilizado en este estudio.

Por esta razón, aunque SVM mostró un rendimiento muy cercano al de la regresión logística, se optó por utilizar Regresión Logística para la optimización de los hiperparámetros, debido a su menor tiempo de entrenamiento sin sacrificar el desempeño. La optimización de los hiperparámetros de la regresión logística se realizó mediante el uso de RandomizedSearchCV, lo que permitió encontrar los mejores valores de los parámetros clave (C, penalty, solver) y mejorar aún más su precisión.

La validación cruzada utilizada durante la optimización mostró que el modelo de regresión logística con los hiperparámetros optimizados alcanzó una precisión en el conjunto de prueba de 0.9959, lo que confirma que este modelo no solo es eficiente en cuanto a tiempo, sino también altamente preciso.

### B. Relevancia de las palabras

Las palabras más frecuentes en las noticias falsas y reales, según el análisis de TF-IDF, revelan patrones significativos en el uso del lenguaje en ambas categorías.

En las noticias falsas, los términos más frecuentes incluyen nombres y temas políticos asociados con figuras prominentes, como “america”, “trump”, “clinton”, “obama”, “hillary”, y “donald”. Estos términos reflejan una tendencia a centrarse en figuras y eventos políticos específicos, particularmente relacionados con las campañas electorales. La presencia de palabras como “media”, “news”, “video”, “time”, y “state” también sugiere un enfoque hacia los temas de actualidad, los medios de comunicación y la cobertura informativa.

En contraste, las noticias reales muestran una distribución de palabras más asociada con temas generales de política y gobernanza, como “campaign”, “election”, “government”, “party”, y “republican”. Las noticias reales tienden a enfocarse en términos más formales y específicos relacionados con el gobierno y las elecciones, con términos recurrentes como “president”, “states”, “washington”, y “united”.

En la Fig 1 se aprecia el conteo de las palabras para cada categoría noticias

Estos resultados evidencian la diferencia en el enfoque temático entre las noticias reales y las falsas. Las noticias falsas tienden a estar más centradas en nombres específicos y temas relacionados con figuras políticas, mientras que las noticias reales se enfocan más en procesos electorales y políticos, con un vocabulario más institucional y generalizado.

Este análisis permite visualizar las diferencias clave entre las dos categorías de noticias, ofreciendo una base sólida para

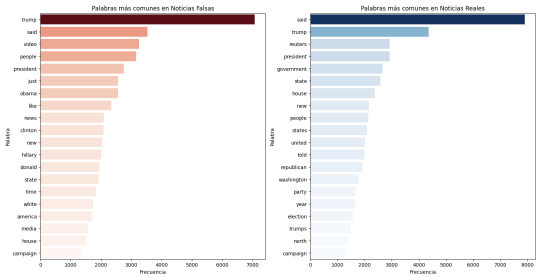


Fig. 1. Conteo de palabras

la clasificación y la detección de noticias falsas a través del modelo entrenado.

## IV. Conclusión

El objetivo principal de este estudio fue realizar un diseño de experimentos para comparar diferentes modelos de clasificación y sus hiperparámetros en relación con la tarea de clasificación de noticias reales y falsas. Se utilizaron tres modelos principales: Naive Bayes, Regresión Logística y SVM, con el propósito de identificar cuál de ellos se adapta mejor a los datos y ofrece un desempeño superior.

Los resultados mostraron que todos los modelos tuvieron un desempeño sobresaliente, con SVM y Regresión Logística alcanzando una precisión cercana al 99%. Sin embargo, el uso de técnicas de optimización, como el RandomizedSearchCV para la Regresión Logística, permitió mejorar la precisión del modelo y reducir los tiempos de entrenamiento, lo que hace que este modelo sea la opción más eficiente en términos de recursos computacionales. La optimización de los hiperparámetros a través de esta técnica contribuyó a una precisión final de 99.6

A través del análisis de las palabras más frecuentes en las noticias reales y falsas utilizando el método TF-IDF, se observó una clara diferencia en los términos utilizados en cada categoría, lo cual validó que las características textuales son clave para la clasificación. Las noticias falsas tendieron a tener un vocabulario más centrado en nombres de figuras políticas y controversias, mientras que las noticias reales mostraron un enfoque más institucional y político.

Finalmente, la comparación de los modelos mediante métricas como precisión, recall y F1-score, junto con la validación cruzada, demostró que la Regresión Logística optimizada fue la que mejor desempeño tuvo, lo que confirma que la selección adecuada de hiperparámetros es crucial para mejorar el rendimiento de los modelos de clasificación de texto.

Este análisis resalta la importancia de aplicar un diseño de experimentos adecuado en tareas de clasificación, permitiendo la evaluación precisa y eficiente de los modelos y la mejora de su rendimiento a través de la optimización de parámetros.