Diseño de Experimentos para Comparar Modelos y sus Hiperparámetros en la Clasificación de Textos

César Alexis Gómez Vieyra
Maestría en Ciencia de Datos
Facultad de Ciencias Físico Matemáticas
San Nicolás de los Garza, México
cesar.gomezvi@uanl.edu.mx

***Resumen*—Este documento presenta un diseño de experimentos para comparar diferentes modelos de clasificación de textos y sus hiperparámetros. Se describen los métodos utilizados, los datos, los resultados obtenidos y las conclusiones derivadas del análisis.**

***Palabras Clave*—Clasificación de textos, diseño de experimentos, modelos de aprendizaje automático, hiperparámetros.**

I. Introducción

La clasificación de textos es una tarea fundamental en el procesamiento del lenguaje natural (PLN). Este trabajo se enfoca en comparar diferentes modelos de clasificación y sus hiperparámetros para determinar cuál ofrece el mejor rendimiento en un conjunto de datos específico.

II. Metodología

II-A. Modelos de Clasificación

Se seleccionaron los siguientes modelos para la comparación:

- Regresión logística
- Redes neuronales
- Agrupamiento de k-medias

II-B. Conjunto de Datos

El contenido del dataset consiste en una colección de 5,157 correos electrónicos, categorizados en dos clases principales: ”Spamƶ ”Ham”(No-Spam). Cada correo electrónico incluye el texto completo del mensaje, que abarca tanto la línea de asunto como el cuerpo del mensaje. La mayoría de los correos electrónicos (87 %) están clasificados como ”Ham”, mientras que el 13 % restante son ”Spam”.

II-C. Transformación de Datos

Los datos anteriormente mencionados se utilizaron para realizar varios conteos, tales como el conteo de palabras, de caracteres, de caracteres especiales y de números,estas cuatro transformaciones ayudarán a realizar las pruebas pertinentes, en el siguiente diagrama de caja (figura 1), podremos ver cómo se distribuyen estos datos.

II-D. Evaluación

Los modelos se evaluaron utilizando la métrica de precisión

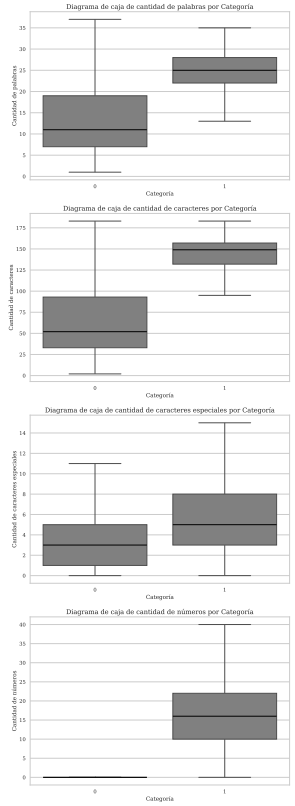

Figura 1. Diagrama de caja para cada variable

III. Resultados

III-A. Comparación de Modelos

Los resultados obtenidos para cada modelo y sus hiperparámetros se presentan en la Tabla I.

Modelo	hiperparametros	presicion
K-means	variables('word', 'char')	78.6 %
K-means	variables('word', 'special_char2')	76.3 %
K-means	variables('word', 'numbers')	78.7 %
K-means	variables('char', 'special_char2')	78.8 %
K-means	variables('char', 'numbers')	78.8 %
K-means	variables('special_char2', 'numbers')	96.3 %
Redes neuronales	Capas: 3, Neuronas: 8, Tasa de Aprendizaje: 0.01	97.5 %
Redes neuronales	Capas: 3, Neuronas: 16, Tasa de Aprendizaje: 0.01	97.7 %
Regresion logistica	çonst=-2.7469, word=-3.4921, char=4.3244, sc=-1.2734, num=4.1955"	96.8 %

Cuadro I
HIPERPARÁMETROS Y PRECISIÓN DE DIFERENTES MODELOS

III-B. Análisis de Resultados

K-means

Las gráficas muestran los resultados del algoritmo de K-Means aplicado a un conjunto de datos con diferentes variables, como *word* (cantidad de palabras), *char* (cantidad de caracteres), *special_char2* (caracteres especiales) y *numbers* (cantidad de números). En general, se observa que los puntos tienden a agruparse en la región inferior izquierda de las gráficas, indicando que la mayoría de los documentos tienen bajos conteos en estas variables. Sin embargo, hay algunas correlaciones notables, como entre *word* y *char*, donde un aumento en el número de palabras se asocia con un aumento en el número de caracteres. Además, hay puntos dispersos que sugieren la presencia de *outliers* o documentos con características significativamente diferentes. En todas las gráficas, los puntos se agrupan en varias regiones, lo que sugiere la existencia de distintos clústeres formados por el algoritmo de K-Means.

Redes neuronales

Para los algoritmos de redes neuronales da como resultado que el pasar de 8 a 16 no hay una diferencia significante por lo cual podemos considerar que son idénticos resultados, asi que lo podemos resumir en que la red neuronal fue entrenada durante 20 épocas, mostrando una mejora significativa en la precisión y una reducción en la pérdida tanto en el conjunto de entrenamiento como en el de validación. Al inicio, en la primera época, la precisión del entrenamiento fue de 72.63 % con una pérdida de 0.4976, mientras que la precisión de validación fue de 95.63 % con una pérdida de 0.1304. A lo largo del entrenamiento, la precisión del modelo en el conjunto de entrenamiento aumentó consistentemente, alcanzando valores superiores al 97 % desde la segunda época en adelante. La precisión de validación se mantuvo estable alrededor del 96 %, con ligeras variaciones. Al final del entrenamiento, en la vigésima época, la precisión del entrenamiento fue de 97.72 % con una pérdida de 0.0688, y la precisión de validación fue de 96.64 % con una pérdida de 0.1129. Estos resultados indican que el modelo tiene un buen rendimiento tanto en el conjunto de entrenamiento como en el de validación, con una alta precisión y una baja pérdida, sugiriendo que el modelo generaliza bien a datos no vistos.

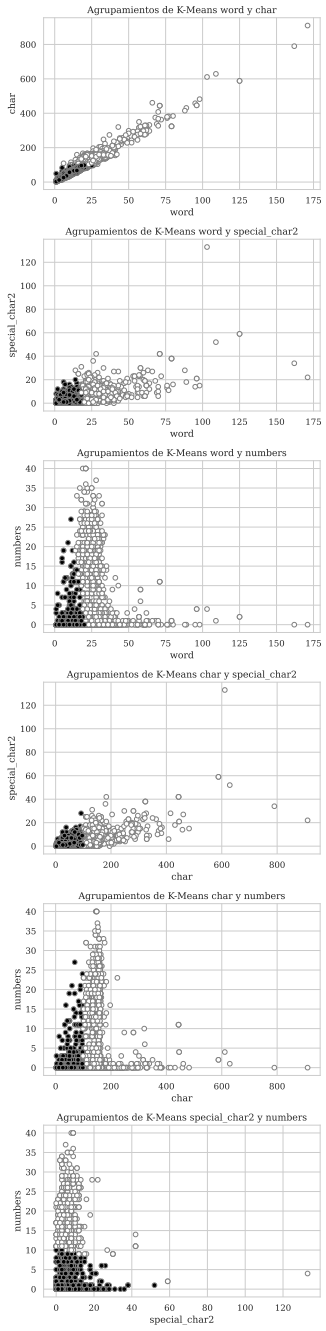

Figura 2. K-means por pares de las variables

Regresión Logística

El modelo de regresión logística tiene como variable dependiente category y se ajustó utilizando el método de Máxima Verosimilitud (MLE) con 4457 observaciones y 4 grados de libertad. El Pseudo R-cuadrado es 0.7731, lo que indica que el modelo explica aproximadamente el 77.31 % de la variabilidad en la variable dependiente. El logaritmo de la verosimilitud es -398.64, y el valor p de la prueba de razón de verosimilitud (LLR) es menor a 0.0001, lo que sugiere que el modelo en

su conjunto es significativo. Los coeficientes de las variables independientes son los siguientes: const (-2.7469), word (-3.4921), char (4.3244), special char2 (-1.2734) y numbers (4.1955). Todos los coeficientes son estadísticamente significativos ($p < 0.05$).

Dep. Variable:	category
No. Observations:	4457
Model:	Logit
Df Residuals:	4452
Method:	MLE
Df Model:	4
Pseudo R-squ.:	0.7731
Time:	03:03:44
Log-Likelihood:	-398.64
converged:	True
LL-Null:	-1757.1
Covariance Type:	nonrobust
LLR p-value:	0.000

	coef	**std err**	**z**	$P > \|z\|$	**[0.025**	**0.975]**
const	-2.7469	0.114	-24.164	0.000	-2.970	-2.524
word	-3.4921	0.458	-7.620	0.000	-4.390	-2.594
char	4.3244	0.467	9.259	0.000	3.408	5.241
special_char2	-1.2734	0.165	-7.720	0.000	-1.597	-0.950
numbers	4.1955	0.253	16.611	0.000	3.701	4.691

Cuadro II
RESULTADOS REGRESIÓN LOGÍSTICA

IV. CONCLUSIONES

En conclusión, los resultados obtenidos de los diferentes modelos de aprendizaje automático y técnicas estadísticas indican un rendimiento robusto y significativo en la clasificación y predicción de datos. El algoritmo de K-Means sugiere la existencia de distintos clústeres, lo que permite identificar patrones y tendencias en el conjunto de datos. Las redes neuronales, entrenadas durante 20 épocas, mostraron una mejora significativa en la precisión y una reducción en la pérdida, alcanzando una precisión de entrenamiento superior al 97 % y una precisión de validación estable alrededor del 96 %, lo que indica un buen rendimiento y generalización a datos no vistos. Por otro lado, el modelo de regresión logística, ajustado mediante el método de Máxima Verosimilitud, explicó aproximadamente el 77.31 % de la variabilidad en la variable dependiente, con coeficientes estadísticamente significativos, sugiriendo un impacto considerable de las variables independientes en la predicción. En resumen, los modelos utilizados demostraron ser efectivos, proporcionando resultados precisos y significativos para una mejor comprensión y predicción de los datos analizados.

REFERENCIAS

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[2] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, Inc., 1988.

[3] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013.