ULPGC
Escuela de Ingeniería Informática

Práctica Clasificación de Textos

José Juan Hernández Gálvez[1] [1]*jose.hernandez219@alu.ulpgc.es*
Jorge Lang-Lenton Ferreiro[2] [2]*jorge.lang101@alu.ulpgc.es*

Resumen

En el presente estudio, se aborda la clasificación de textos del conjunto de datos 20newsgroup, empleando dos técnicas de clasificación: Naive Bayes y Regresión Logística. El propósito principal es evaluar y comparar el rendimiento de estas técnicas en términos de exactitud, precisión y sensibilidad para discernir cuál de ellas se desempeña de manera más óptima en este dataset específico. La metodología y los resultados detallados ofrecen una comprensión clara de las capacidades y limitaciones de cada enfoque en el contexto de 20newsgroup.

1. Introducción

El análisis y clasificación de grandes volúmenes de datos textuales es una tarea que ha ganado relevancia en los últimos años, especialmente con el surgimiento de herramientas más sofisticadas de procesamiento de lenguaje natural y aprendizaje automático. Uno de los conjuntos de datos más emblemáticos en este ámbito es el dataset 20newsgroup, compuesto por mensajes de veinte grupos de noticias, representando así un desafío interesante para la clasificación textual.

Para abordar este análisis y experimentación, emplearemos el lenguaje de programación Python. En particular, haremos uso de la librería scikit-learn, una potente herramienta destinada al aprendizaje automático que brinda diversos algoritmos y herramientas de preprocesamiento, facilitando un enfoque ya más optimizado.

Antes de sumergirnos en la exploración, es vital tener claridad sobre ciertos términos claves presentes en el documento:

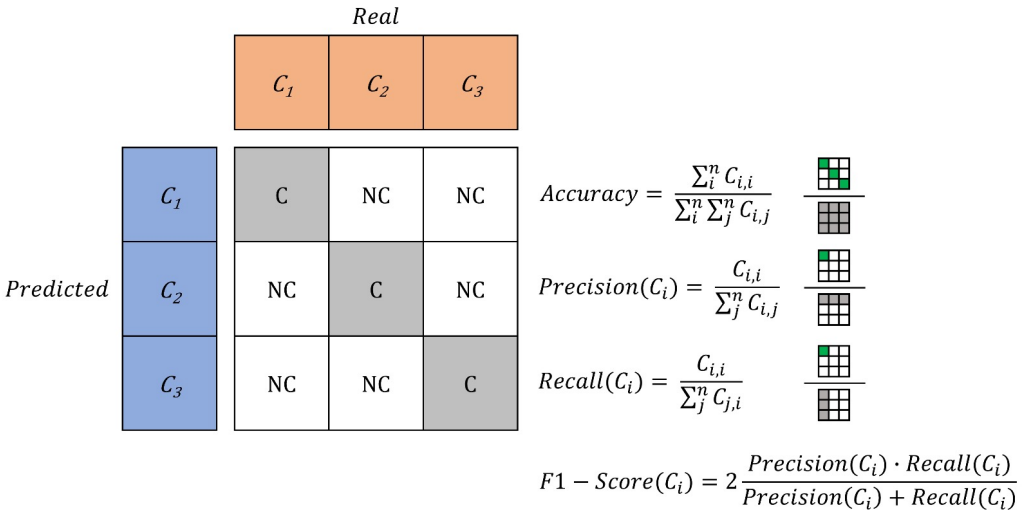

Figura 1: **Matriz de Confusión**: Una tabla visual que muestra el desempeño de un modelo de clasificación, reflejando las predicciones versus los valores reales, en un caso en que hay 3 clases diferentes.

- **Accuracy** (Exactitud): Indica el porcentaje de predicciones correctas del total. Es, en esencia, la suma de las predicciones acertadas dividida entre todas las predicciones.
- **Precision** (Precisión): Dado que el modelo ha predicho que una entrada pertenece a la clase C_n, ¿cuántas veces acertó en esa predicción (sobre todos los positivos predichos)?
- **Recall** (Sensibilidad): Dado que una entrada realmente pertenece a la clase C_n, ¿cuántas veces el modelo lo identificó correctamente como C_n (sobre todos los positivos reales)?
- **F1-Score** (Puntuación F1): Una métrica balanceada que combina la Precision y el Recall, brindando un panorama integral, especialmente útil cuando hay desbalance entre clases.

Entender estas métricas es esencial para una evaluación adecuada de nuestros modelos y para la toma de decisiones fundamentada sobre qué técnica de clasificación se alinea mejor con nuestros objetivos. Con estos cimientos establecidos, nos encontramos preparados para adentrarnos en nuestro análisis.

2. Descripción del dataset 20newsgroup

2.1. Breve historia y origen

El conjunto de datos **20newsgroup** es una colección de aproximadamente 20,000 documentos repartidos en 20 diferentes foros, cada uno abordando un tema distinto. Estos foros de discusión en línea que florecieron durante los primeros días del internet, sirvieron como plataformas donde las personas podían discutir temas específicos. En los años 90, este conjunto de datos empezó a utilizarse ampliamente como recurso para investigaciones en machine learning y procesamiento de texto.

2.2. Características generales del dataset

- **Número Total de Documentos**: El conjunto contiene alrededor de 11,314 documentos.
- **Tamaño Medio de Documentos**: En promedio, un documento contiene alrededor de 317.6 palabras.
- **Variedad de Temas**: Los temas abordados en los newsgroups varían desde deportes y tecnología hasta política y religión, lo que lo hace diverso y representativo de las discusiones en línea de la época.

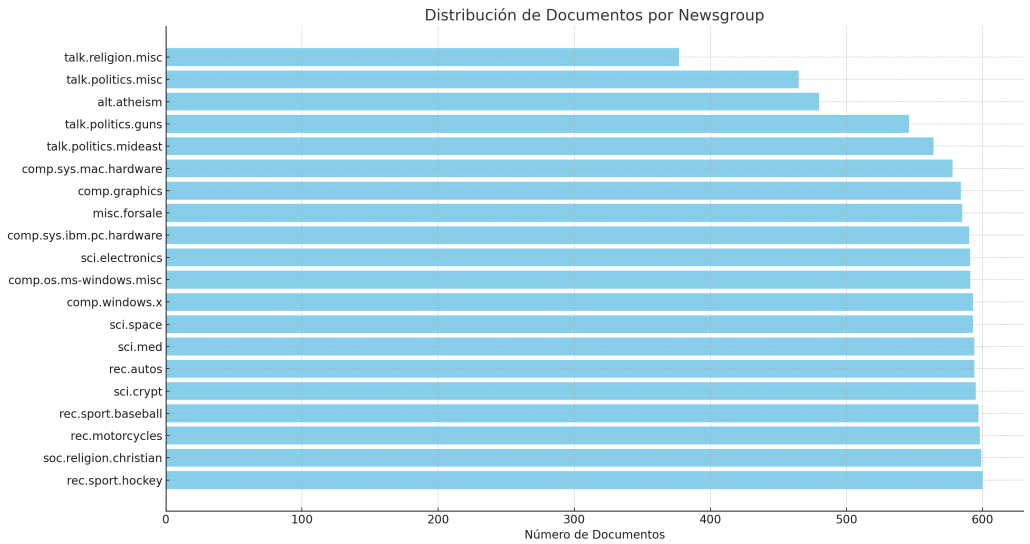

Figura 2: Distribución de documentos por newsgroup

Se puede observar en la Figura 2 justo encima que los newsgroups tienen una distribución relativamente uniforme en términos del número de documentos, con ciertas categorías, como rec.sport.hockeyz "soc.religion.christian", que poseen ligeramente más documentos que otros. En contraste, "talk.religion.miscz "talk.politics.misc"tienen menos documentos en comparación con el resto.

2.3. Relevancia y Aplicaciones Previas

Desde su popularización en los años 90, el dataset "20 Newsgroups"ha sido un recurso clave en la comunidad de Machine Learning y procesamiento de lenguaje natural. Su versatilidad y diversidad lo hacen ideal para tareas como clasificación de texto, clustering y técnicas de reducción de dimensionalidad. Además, ha servido como un estándar para comparar la eficacia de diversos algoritmos y técnicas, permitiendo a los investigadores obtener insights sobre las mejores prácticas y enfoques en el campo del análisis de texto.

3. Preprocesamiento de los datos

El preprocesamiento de datos es una fase crucial en cualquier proyecto de Machine Learning, pues nos aseguramos de que los datos estén en un formato adecuado para ser procesados por los algoritmos. En este proyecto, realizamos varias etapas de preprocesamiento en el conjunto de datos "20 Newsgroups".

3.1. Vectorización de documentos

Dada la naturaleza textual del conjunto de datos, es fundamental convertir los documentos en una representación numérica. Para lograr esto, aplicamos la técnica de Count Vectorization, que convierte los textos en vectores basados en la frecuencia de palabras. Optamos por que este vector tenga 10,000 características distintivas, dejando fuera las "stop-words", ya que, generalmente, carecen de significado semántico relevante.

3.2. División del conjunto de datos

Para entrenar y validar el rendimiento de los modelos, hemos dividido el conjunto de datos en conjuntos de entrenamiento y prueba, utilizando un 80 % de los datos para entrenamiento y el restante 20 % para prueba.

Estas clases y métodos garantizan una estructura organizada y eficiente del preprocesamiento, facilitando su implementación y posterior uso en la fase de modelado.

4. Modelos de ML en Estudio

El mundo del aprendizaje automático y la inteligencia artificial es vasto y diverso, ofreciendo un abanico de técnicas y algoritmos diseñados para abordar una variedad de problemas. Desde la agrupación de datos sin etiquetas hasta la predicción de categorías para datos de entrada, los algoritmos de Machine Learning han revolucionado la forma en que procesamos y entendemos grandes cantidades de información.

4.1. Logistic Regression (Regresión Logística)

La regresión logística es un modelo de clasificación que estima la probabilidad de que una instancia dada pertenezca a una categoría particular. Es especialmente útil cuando la variable objetivo es binaria. A pesar de su nombre, es un algoritmo de clasificación y no de regresión.

4.1.1. Resultados

Class	Precision	Recall	F1-Score	Support
0	0.95	0.92	0.93	97
1	0.67	0.80	0.73	104
2	0.85	0.79	0.82	115
3	0.71	0.76	0.74	123
4	0.84	0.76	0.80	126
5	0.83	0.85	0.84	106
6	0.83	0.83	0.83	109
7	0.90	0.90	0.90	139
8	0.92	0.89	0.90	122
9	0.84	0.96	0.89	102
10	0.94	0.93	0.93	108
11	0.99	0.94	0.97	125
12	0.80	0.83	0.82	114
13	0.93	0.95	0.94	119
14	0.98	0.91	0.95	127
15	0.88	0.90	0.89	122
16	0.92	0.94	0.93	121
17	0.96	0.98	0.97	102
18	0.93	0.88	0.90	107
19	0.86	0.75	0.80	75
Accuracy: 0.8749447635881573				
Macro Avg: 0.88 / 0.87 / 0.87 / 2263				
Weighted Avg: 0.88 / 0.87 / 0.88 / 2263				

Cuadro 1: Resultados de la Regresión Logística en el conjunto de datos 20 Newsgroups

El modelo de Regresión Logística aplicado a este conjunto de datos muestra un rendimiento considerablemente alto. Destaquemos algunas observaciones clave:

- La precisión global (accuracy) del modelo es del 87.49 %, lo que indica que casi 9 de cada 10 predicciones son correctas. Esta es una métrica robusta, especialmente si se compara con los resultados de otros modelos.
- La mayoría de las clases tienen una precisión (precision), recall y F1-score muy altos, con valores por encima del 80 % en casi todas ellas. Esto sugiere que el modelo es capaz de identificar y clasificar correctamente la mayoría de las observaciones en estas categorías.
- Es notable que algunas clases, como la clase 11 y 17, tienen métricas excepcionalmente altas, cercanas al 95 % o incluso superiores en algunas métricas. Estas clases están siendo clasificadas con mucha precisión.
- Aunque algunas clases, como la clase 1, tienen métricas ligeramente más bajas en comparación con otras, todavía muestran un buen rendimiento general, con un F1-score de 0.73.
- Las métricas agregadas Macro Avg y Weighted Avg reflejan este alto rendimiento con valores cercanos o superiores al 87

4.2. Naive Bayes

Este es un conjunto de algoritmos de aprendizaje supervisado basados en la aplicación del teorema de Bayes con el supuesto "naive" (ingenuo) de que cada par de características es independiente entre sí. Es especialmente conocido por su eficiencia y su uso en la clasificación de texto.

4.2.1. Resultados

Class	Precision	Recall	F1-score	Support
0	0.88	0.94	0.91	97
1	0.60	0.87	0.71	104
2	1.00	0.03	0.07	115
3	0.56	0.79	0.66	123
4	0.75	0.89	0.81	126
5	0.82	0.88	0.85	106
6	0.73	0.85	0.79	109
7	0.89	0.91	0.90	139
8	0.91	0.93	0.92	122
9	0.91	0.94	0.92	102
10	0.97	0.93	0.95	108
11	0.99	0.90	0.95	125
12	0.89	0.82	0.85	114
13	0.96	0.92	0.94	119
14	0.97	0.94	0.95	127
15	0.93	0.91	0.92	122
16	0.93	0.93	0.93	121
17	0.98	0.94	0.96	102
18	0.85	0.91	0.88	107
19	0.83	0.65	0.73	75
Accuracy: 0.85				
Macro Avg: 0.87 / 0.84 / 0.83 / 2263				
Weighted Avg: 0.87 / 0.85 / 0.83 / 2263				

Cuadro 2: Resultados de Naive Bayes en el conjunto de datos 20 Newsgroups

El modelo Naive Bayes, cuando se aplica al conjunto de datos "20 Newsgroups", presenta un rendimiento sólido, aunque con algunas particularidades a considerar:

- La precisión global (accuracy) del modelo es del 85 %, lo que indica que 17 de cada 20 predicciones son correctas.
- La mayoría de las clases presentan una alta precisión (precision), recall y F1-score, siendo varias de ellas superiores al 90 %. Esto demuestra que, en general, el modelo tiene un buen desempeño al clasificar las observaciones en estas categorías.
- Sin embargo, hay una peculiaridad destacada: la clase 2 tiene una precisión perfecta de 1.00, pero un recall extremadamente bajo de 0.03, lo que resulta en un F1-score de apenas 0.07. Esto sugiere que, aunque el modelo es muy preciso cuando clasifica una observación como perteneciente a la clase 2, raramente lo hace, lo que lleva a un recall muy bajo.
- Otras clases, como la clase 1 y 3, muestran recall más altos en comparación con su precisión. Esto indica que el modelo tiene una tendencia a sobreestimar la pertenencia a estas clases.
- Las métricas agregadas Macro Avg y Weighted Avg muestran que, en promedio, el modelo tiene un rendimiento equilibrado entre precisión y recall, aunque la métrica F1 está ligeramente más baja debido a clases como la clase 2 que impactan negativamente la media.

4.3. SVM (Máquinas de Vectores de Soporte)

Las SVM son un conjunto de algoritmos de aprendizaje supervisado que se utilizan para clasificación y regresión. Trabajan encontrando el hiperplano que mejor divide un conjunto de datos en clases. Son especialmente efectivas en espacios de alta dimensión y son conocidas por su eficacia en situaciones donde el margen de separación entre clases es pequeño.

4.3.1. Resultados

Class	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	97
1	1.00	0.01	0.02	104
2	0.83	0.04	0.08	115
3	1.00	0.02	0.05	123
4	0.00	0.00	0.00	126
5	0.45	0.09	0.16	106
6	1.00	0.02	0.04	109
7	1.00	0.01	0.01	139
8	0.00	0.00	0.00	122
9	0.05	1.00	0.09	102
10	1.00	0.04	0.07	108
11	1.00	0.09	0.16	125
12	0.00	0.00	0.00	114
13	1.00	0.01	0.02	119
14	1.00	0.02	0.05	127
15	0.64	0.13	0.22	122
16	1.00	0.04	0.08	121
17	0.94	0.15	0.25	102
18	1.00	0.04	0.07	107
19	0.00	0.00	0.00	75
Accuracy: 0.08				
Macro Avg: 0.65 / 0.09 / 0.07 / 2263				
Weighted Avg: 0.66 / 0.08 / 0.07 / 2263				

Cuadro 3: Resultados de SVM en el conjunto de datos 20 Newsgroups

El modelo de Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) muestra un rendimiento bastante atípico y poco óptimo en el conjunto de datos "20 Newsgroups", tal y como se puede ver en métricas presentadas:

- El valor de accuracy es extremadamente bajo, con solo un 8 %. Esto significa que, en promedio, sólo 8 de cada 100 predicciones son correctas. Es una métrica muy preocupante y sugiere que el modelo no está funcionando adecuadamente.
- Varios valores de precisión y recall son extremos (o 0.00 o 1.00), lo que sugiere que el modelo tiene problemas al balancear las clases y/o al hacer predicciones para algunas de ellas. Por ejemplo, las clases 0, 4, 8, 12 y 19 tienen precisiones, recalls y F1-scores de 0.00, lo que indica que el modelo no ha sido capaz de hacer ninguna predicción correcta para estas clases.
- Sorprendentemente, hay clases donde la precisión es perfecta (1.00), pero el recall es muy bajo, lo que lleva a F1-scores muy bajos. Esto sugiere que, aunque el modelo es muy preciso en sus predicciones para estas clases, raramente las predice.
- Un caso particularmente notorio es la clase 9, que tiene un recall de 1.00 pero una precisión de 0.05, lo que lleva a un F1-score de sólo 0.09. Esto sugiere que el modelo está prediciendo casi todas las observaciones como pertenecientes a esta clase, pero en su mayoría de forma incorrecta.
- Las métricas agregadas Macro Avg y Weighted Avg muestran una gran discrepancia entre precisión y recall, con un F1-score medio de sólo 0.07. Esto refuerza la idea de que el modelo no está equilibrando adecuadamente sus predicciones.

4.4. Decision Trees (Árboles de Decisión)

Los árboles de decisión son modelos de aprendizaje supervisado que predicen una variable objetivo basándose en reglas simples inferidas a partir de las características de los datos. Pueden ser utilizados tanto para tareas de clasificación como de regresión. Son especialmente útiles por su facilidad de interpretación visual.

4.4.1. Resultados

Class	Precision	Recall	F1-Score	Support
0	0.62	0.69	0.65	97
1	0.47	0.49	0.48	104
2	0.62	0.69	0.65	115
3	0.52	0.48	0.50	123
4	0.67	0.56	0.61	126
5	0.71	0.66	0.69	106
6	0.56	0.69	0.62	109
7	0.71	0.65	0.68	139
8	0.74	0.75	0.74	122
9	0.57	0.71	0.63	102
10	0.82	0.70	0.76	108
11	0.87	0.81	0.84	125
12	0.47	0.54	0.50	114
13	0.59	0.70	0.64	119
14	0.78	0.73	0.76	127
15	0.82	0.80	0.81	122
16	0.75	0.71	0.73	121
17	0.83	0.88	0.85	102
18	0.67	0.54	0.60	107
19	0.51	0.43	0.46	75
Accuracy: 0.66				
Macro Avg: 0.67 / 0.66 / 0.66 / 2263				
Weighted Avg: 0.67 / 0.66 / 0.67 / 2263				

Cuadro 4: Resultados de los Árboles de Decisión en el conjunto de datos 20 Newsgroups

El modelo de Árboles de Decisión, según las métricas presentadas para el conjunto de datos "20 Newsgroups", muestra un rendimiento moderado. Veamos algunas observaciones basadas en los resultados:

- El valor de accuracy es del 66 %, lo que indica que dos tercios de las predicciones del modelo son correctas. Aunque esta precisión es significativamente mejor que el modelo SVM anteriormente evaluado, todavía hay margen de mejora.
- No hay clases con precisión, recall o F1-scores extremadamente bajos, lo que sugiere que el modelo es capaz de realizar predicciones adecuadas en todas las clases. Sin embargo, hay variaciones en el rendimiento entre clases.
- Las clases 10, 11, 15, y 17 destacan con F1-scores que superan el 0.80, lo que indica que el modelo es particularmente bueno en la predicción de estas clases.
- Por otro lado, las clases 1, 3, 12, y 19 tienen F1-scores por debajo del 0.50. Estas clases podrían ser las áreas donde el modelo enfrenta desafíos y donde se podría beneficiar de un mayor ajuste o características adicionales.
- Las métricas agregadas Macro Avg y Weighted Avg están en línea con el valor de accuracy, todas rondando el 66-67 %. Esto indica una cierta coherencia en el rendimiento del modelo a través de las diferentes clases.
- Hay cierto equilibrio entre precisión y recall en muchas clases, aunque hay algunas, como la clase 4 o la clase 18, donde hay una discrepancia notable entre estas dos métricas.

4.5. KMeans

Es un algoritmo de clustering que tiene como objetivo dividir un conjunto de puntos en K grupos, donde cada punto pertenece al grupo cuyo centroide (promedio de todos los puntos del cluster) es más cercano. El algoritmo funciona iterativamente para asignar cada punto al cluster más cercano.

4.5.1. Resultados

Class	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	97
1	0.00	0.00	0.00	104
2	0.00	0.00	0.00	115
3	0.00	0.00	0.00	123
4	0.00	0.00	0.00	126
5	0.00	0.00	0.00	106
6	0.05	1.00	0.09	109
7	0.00	0.00	0.00	139
8	0.00	0.00	0.00	122
9	0.00	0.00	0.00	102
10	0.00	0.00	0.00	108
11	0.00	0.00	0.00	125
12	0.00	0.00	0.00	114
13	0.00	0.00	0.00	119
14	0.00	0.00	0.00	127
15	0.00	0.00	0.00	122
16	0.00	0.00	0.00	121
17	0.00	0.00	0.00	102
18	0.00	0.00	0.00	107
19	0.00	0.00	0.00	75
Accuracy: 0.05				
Macro Avg: 0.00 / 0.05 / 0.00 / 2263				
Weighted Avg: 0.00 / 0.05 / 0.00 / 2263				

Cuadro 5: Resultados de KMeans en el conjunto de datos 20 Newsgroups

Los resultados del algoritmo KMeans muestran un rendimiento bastante bajo. Las métricas de precisión, recall y F1-score están cercanas a 0 para casi todas las clases, excepto para la clase 6, que parece haber sido etiquetada correctamente en todas las instancias, aunque es probable que esto se deba a un error en la asignación de etiquetas más que a una clasificación precisa. Con un accuracy del 5 %, es evidente que este modelo necesita ser mejorado o reajustado.

4.6. Aglomerative Clustering (Clustering Aglomerativo)

Este es un método de clustering jerárquico que comienza considerando cada punto de datos como un cluster individual y luego, iterativamente, combina los clusters más cercanos en clusters más grandes. El proceso se repite hasta que todos los puntos de datos se agrupan en un único cluster o hasta que se cumplan ciertos criterios.

4.6.1. Resultados

Class	Precision	Recall	F1-Score	Support
0	0.04	1.00	0.08	383
1	0.00	0.00	0.00	480
2	0.00	0.00	0.00	476
3	0.00	0.00	0.00	467
4	0.00	0.00	0.00	452
5	0.00	0.00	0.00	487
6	0.00	0.00	0.00	476
7	0.00	0.00	0.00	455
8	0.00	0.00	0.00	476
9	0.00	0.00	0.00	495
10	0.00	0.00	0.00	492
11	0.00	0.00	0.00	470
12	0.00	0.00	0.00	477
13	0.00	0.00	0.00	475
14	0.00	0.00	0.00	466
15	0.00	0.00	0.00	477
16	0.00	0.00	0.00	425
17	0.00	0.00	0.00	462
18	0.00	0.00	0.00	358
19	0.00	0.00	0.00	302
Accuracy: 0.0423				
Macro Avg: 0.00 / 0.05 / 0.00 / 9051				
Weighted Avg: 0.00 / 0.04 / 0.00 / 9051				

Cuadro 6: Resultados del algoritmo Agglomerative Clustering en el conjunto de datos proporcionado

El algoritmo de Clustering Aglomerativo muestra un rendimiento muy bajo. La única clase que tuvo algún nivel de precisión fue la clase 0, pero esto resultó en un F1-score muy bajo debido a la falta de precisión en las otras clases. El accuracy del modelo es solo del 4.23 %, lo que indica que la asignación de clusters no fue efectiva para este conjunto de datos.