Instituto Politécnico Nacional
Escuela Superior de Cómputo
Unidad de Aprendizaje:
Procesamiento del Lenguaje Natural

Práctica 6:

Sentiment Analysis

Profesor: Joel Omar Juárez Gambino

Equipo 3:

- **Gutiérrez Pérez Gabriela G.**
- **Loeza Saldaña Sebastián**
- **Ramos García Luis Gerardo**
- **Rico Mendoza Josué**

Fecha de entrega: 06 de diciembre 2024

Introducción:

El análisis de sentimientos es una tarea fundamental en el procesamiento del lenguaje natural (NLP) que busca identificar la polaridad de opiniones expresadas en texto. En esta práctica, se trabajará con el corpus **Rest_Mex_2022**, que contiene opiniones sobre restaurantes, hoteles y atracciones turísticas, clasificadas en cinco niveles de polaridad: siendo **1**: Muy negativo, **2**: Negativo, **3**: Neutral, **4**: Positivo y **5**: Muy positivo

Para realizar la práctica se realizó un preprocesamiento de texto, esta es una etapa clave en cualquier proyecto de análisis de texto, ya que permite transformar datos textuales crudos en un formato adecuado para ser utilizado por algoritmos de Machine Learning.

Se aplicaron técnicas como la **tokenización** que es el proceso de dividir un texto en unidades más pequeñas llamadas *tokens*, que pueden ser palabras, frases, oraciones o incluso caracteres individuales. Estas unidades son la base para el análisis de texto, ya que los modelos y algoritmos trabajan con estos tokens en lugar de con el texto completo. **Text cleaning** que es el proceso de eliminar o transformar elementos no deseados en el texto para facilitar su análisis, con el fin de reducir el ruido y mantener solo la información relevante se puede elminar la puntuación o caracteres especiales, quitar espacios extra, eliminar números o convertir el texto a minúsculas. También se eliminaron las palabras clasificadas como **stop words**, las cuales son palabras frecuentes en un idioma que generalmente no aportan significado útil al análisis. Estas palabras, como “el”, “de”, “a”, “y”, “es”, suelen eliminarse para centrarse en términos más significativos. Y finalmente un proceso de **lemmatización** que se caracteriza por reducir las palabras a su forma base o “lema”, teniendo en cuenta su significado y función gramatical.

Estos procesos son esenciales para preparar textos antes de su análisis, mejorando la precisión y eficiencia de los modelos que trabajan con datos de lenguaje natural.

Los modelos de Machine Learning requieren datos numéricos. Por ello, el texto debe representarse de forma matemática, así se realiza una representación binaria para indicar la presencia de palabras como valores binarios (0, 1).

Dentro de los modelos para sentyment análisis aplicados en esta práctica se encuentran:

Regresión logística: es un modelo de clasificación supervisado que se utiliza para predecir probabilidades de pertenencia a una clase. Es un modelo con alta interpretabilidad y es sencillo de implementar.

Naive Bayes: es un modelo probabilístico basado en el Teorema de Bayes, muy útil en tareas de clasificación como análisis de sentimientos, spam detection y categorización de texto. Asume que las características son condicionalmente independientes, lo cual simplifica los cálculos aunque dicha suposición rara vez se cumple en la práctica. Es sencillo de implementar y eficiente en problemas de texto con alta dimensionalidad.

Perceptrón multicapa (PLM) es un tipo de red neuronal artificial que se utiliza para resolver problemas de clasificación y regresión. A diferencia del perceptrón simple, el MLP es capaz de modelar relaciones complejas y no lineales. Cada neurona calcula una combinación lineal de

sus entradas y aplica una función de activación. Es altamente flexible, sin embargo puede requerir mucho tiempo y recursos computacionales para entrenar y es suceptible a sobreajuste si no se regulariza adecuadamente.

Al implementar los modelos de aprendizaje supervisado es importante realizar validaciones como el caso de k-folds cross validation, la cual es una técnica para evaluar modelos dividiendo en "K" particiones o folds. Cada vez, un fold se utiliza como conjunto de prueba y el resto como entrenamiento, asegurando que el modelo sea evaluado en todo el conjunto de datos.

Cabe destacar que realizar una evaluación del rendimiento de los modelos es crucial para garantizar que funcione correctamente, para ello se consideró la métrica F1, la cual es la media armónica entre precisión (proporción de predicciones correctas en las positivas) y recall (proporción de elementos positivos correctamente identificados). La versión empleada en esta práctica es "macro" que calcula el F1 para cada clase y luego toma la media, dando igual importancia a todas las clases, incluso si están desbalanceadas.

Es importante mencionar que el dataset presenta desbalance, esto ocurre cuando las clases no tienen la misma cantidad de ejemplos, lo que puede sesgar los modelos hacia la clase mayoritaria. Para abordar el desbalance de clases se implementó el método de undersampling a la clase mayoritaria manteniendo la clase minoritaria sin cambios para mejorar el equilibrio de los datos de entrenamiento.

Todo esto se realizó con el objetivo principal de mejorar el rendimiento de los modelos, empleando técnicas avanzadas de preprocesamiento, representación de texto y aprendizaje automático. Además, se busca abordar desafíos como el desbalance de clases y explorar métodos para optimizar las predicciones.

Metodología:

Para llevar a cabo esta práctica, experimentamos con distintas técnicas de normalización para evaluar los modelos. Por ejemplo, al probar el modelo de **Regresión Logística**, únicamente aplicamos tokenización. En cambio, para los modelos de **Naive Bayes**, realizamos una serie de preprocesamientos que incluyeron tokenización, limpieza de texto, lematización y eliminación de stop words. Mientras que el **Perceptrón** se le aplicaron las técnicas de tokenización, limpieza de texto y eliminación de stop words. Debido a las limitaciones que se tienen en recursos de cómputo este último modelo se decidió probarse al final con los preprocesamientos que se consideran óptimos.
Para los 3 modelos se usará una representación binaria ya que esta fue la que mejor resultados nos proporcionó en la práctica anterior.

Aunado a esto, se procedió a realizar el primer experimento el cuál consistio en el entrenamiento con la técnica de k-fold cross validation en los modelos de Regresión logística y Naive Bayes. Obteniendo un F1 Score Macro promedio de. 0.4569 para el primero y 0.2743 para el segundo.

Estos resultados indican que el modelo de Regresión Logística mostró un mejor desempeño en términos de equilibrio entre precisión y exhaustividad para las diferentes clases, mientras que el modelo de Naive Bayes presentó un rendimiento significativamente inferior. Es por ello que a partir de este punto se proponen mejoras para aumentar el desempeño del modelo.

Mejoras propuestas:

Primera propuesta (uso de un diccionario con un acumulado por cada emoción)

La primer mejor propuesta es el uso de un diccionario de emojis, con la finalidad de capturar las emociones de las personas al hacer uso de dicho recurso.

Después de la implementación del diccionario de emojis nos percatamos que, del total de opiniones, son muy pocas aquellas que hacen uso de ellos. Por esta razón es que se descartó el uso de esta propuesta.

Segunda propuesta (uso de un diccionario con un acumulado por cada emoción)

Como segundo experimento, además de evaluar los modelos con las diferentes normalizaciones y aplicar validación cruzada con k-fold (como se realizó anteriormente), proponemos incorporar un preprocesamiento adicional. Para ello, utilizamos el documento SEL, el cual clasifica ciertas palabras en categorías emocionales (como alegría, tristeza, etc.) y les asigna un nivel de intensidad (nulo, bajo, medio o alto).

Para representar estas intensidades numéricamente, diseñamos la siguiente fórmula:

$$\text{Puntaje} = (\text{Nulo} \times 0) + (\text{Bajo} \times 1) + (\text{Medio} \times 2) + (\text{Alto} \times 3)$$

El preprocesamiento consistió en buscar, para cada opinión, si alguna de sus palabras estaba presente en el diccionario SEL. En caso afirmativo, se aplicó la fórmula para calcular un puntaje acumulativo por cada emoción. El diseño de la fórmula permite que niveles altos de intensidad emocional generen puntajes más elevados, mientras que niveles nulos mantengan el puntaje bajo. Esto proporciona una representación numérica clara y ponderada de las emociones presentes en el texto. Con esto en cuenta, los puntajes del F1 score macro promedio para los 5 fols es de 0.4668 y y 0.2965 para el modelo de regresión logística y Naïve bayes respectivamente.

Tercera propuesta (uso de un diccionario con acumulado negativo y positivo)

Como tercer experimento, integramos todos los preprocesamientos anteriores, incluyendo el uso del diccionario SEL, con una modificación clave: las emociones se agruparon en dos categorías generales, *positivas* y *negativas*. En este enfoque, cada vez que se identificaba una palabra en el corpus que coincidiera con una entrada del diccionario SEL, se sumaba su Frecuencia Ponderada de Aparición (FPA) a la categoría correspondiente.

	Tokenization	AcumulativoPositivo	AcumulativoNegativo	Polarity
0	Pésimo lugar Piensen dos veces antes de ir a e...	0.792	0.000	1
1	No vayas a lugar de Eddie Cuatro de nosotros f...	0.660	2.489	1
2	Mala relación calidad-precio seguiré corta y s...	0.000	0.000	1
3	Minusválido ? ¡ No te alojes aquí ! Al reserva...	1.529	0.000	1
4	Es una porqueria no pierdan su tiempo No pierd...	0.561	0.629	1
...	...	...	...	...
30207	Verdadera joya arquitectónica Es una construcc...	0.763	0.000	5
30208	Romántico Muy al estilo de Romeo y Julieta es ...	0.630	0.000	5
30209	Parece un castillo Ideal para subir las escali...	0.000	0.000	5
30210	Imperdible Es imperdible , de ahí puedes ver m...	1.263	0.000	5
30211	Muy bonita vista No te puedes ir de Guanajuato...	0.000	1.526	5

Este procedimiento permitió acumular el impacto de las palabras según su connotación emocional, simplificando el análisis a una polaridad general entre emociones positivas y negativas. Los resultados obtenidos mostraron un F1 Score Macro promedio de 0.4631 para el modelo de Regresión Logística y 0.2762 para Naïve Bayes, lo que sugiere un desempeño consistente de Regresión Logística como el modelo más robusto en este experimento.

Cuarta propuesta (submuestreo para balancear clases)

En el cuarto experimento, decidimos excluir el modelo de Naive Bayes de ahora en adelante debido a su desempeño consistentemente bajo en los experimentos previos. En su lugar, nos enfocamos en probar solo regresión logística utilizando el conjunto de datos con las normalizaciones aplicadas previamente (solo tokenización para logística). La principal diferencia en este experimento fue la incorporación de un preprocesamiento de submuestreo, cuyo objetivo era abordar el desbalanceo de clases en el conjunto de datos. Este enfoque buscó equilibrar la representación de las clases, mejorando así el rendimiento de los modelos en escenarios donde las clases minoritarias suelen tener menor impacto en las métricas de desempeño. Para este experimento, se obtuvo un F1 Score macro promedio de 0.4780, lo que sugiere una ligera mejora en comparación con experimentos anteriores. Esto indica que el submuestreo fue efectivo para mitigar el problema del desbalanceo de clases, logrando un desempeño más equitativo entre las diferentes clases y un impacto positivo en el rendimiento general del modelo.

Quinta propuesta (submuestreo + acumulado positivo y negativo)

En el quinto experimento, incorporamos nuevamente el conjunto de datos SEL y aplicamos la técnica de submuestreo. Este experimento se dividió en dos subexperimentos.El primer subexperimento consistió en agregar dos columnas que clasificaran las emociones en categorías generales: *positivas* y *negativas*. Con este enfoque, el modelo de Regresión Logística obtuvo un F1 Score Macro promedio de 0.4801. El segundo subexperimento siguió un enfoque similar, pero en lugar de agrupar las emociones, se agregó una columna por cada emoción específica. Con esta configuración, el modelo alcanzó un F1 Score Macro promedio de 0.4798. Estos resultados pueden sugerir que clasificar las emociones en categorías generales o específicas tiene un impacto mínimo en el desempeño del modelo. Esto podría deberse a que la información esencial capturada por ambas configuraciones es equivalente

en términos de su utilidad para la clasificación. No obstante, agrupar las emociones en categorías generales podría ser más práctico en términos de simplicidad y eficiencia computacional.

Sexta propuesta (uso de perceptrón con submuestreo + acumulado positivo y negativo)
Finalmente, se evaluó el modelo Perceptrón utilizando procesos de normalización y una serie de preprocesamientos que incluyeron tokenización, limpieza de texto y eliminación de stop words. Además, se volvió a aplicar la técnica de submuestreo para abordar el desbalance de clases en el conjunto de datos. Asimismo, se incorporó el conjunto de datos SEL, añadiendo una columna por cada emoción específica. Este enfoque permitió enriquecer la representación de las emociones en el modelo. El Perceptrón alcanzó un F1 Score Macro de 0.4842, lo que representa el mejor desempeño logrado hasta ahora. Este resultado sugiere que la combinación de submuestreo con una representación detallada de las emociones contribuyó positivamente al equilibrio y a la precisión general del modelo.

Resultados con el conjunto de entrenamiento (Cross validation)

	Modelos (en orden de creación)	F1 Score (MACRO)
Primera prueba (versión base)	Regresión logística	0.4569098253655052
	Naive Bayes	0.27432807543570803
En este punto proponemos la primera mejora para el modelo mencionado anteriormente		
Segunda prueba (acumulado por cada emoción)	Regresión logística	0.466764055437772
	Naive Bayes	0.29650836637340444
Siguiendo la idea de usar un diccionario se propone hacerlo de una manera más general (positivo y negativo) en vez de por cada emoción.		
Tercera prueba (acumulado positivo y negativo)	Regresión logística	0.46310505586103484
	Naive Bayes	0.2761822241250599
Notamos que Naive Bayes no es un modelo prometedor para este conjunto de datos, así que se decide abandonarlo y además se propone un balanceo de clases como siguiente mejora.		
Cuarta prueba (balanceo de clases)	Regresión logística	0.4780388073863982
Notamos que las 2 mejoras anteriores aumentaron nuestro F1 MACRO, por lo que propusimos combinar ambas, además de probar las 2 maneras de acumulación (positiva con negativa y por emociones).		
Quinta prueba (balanceo y diccionario)	Regresión logística (acumulado positivo y negativo)	0.48011660803934475

	Regresión logística (acumulado por cada emoción)	0.47984899305710166
Notamos que la mejor manera de representar las emociones es la más general (positivas y negativas), hasta el momento tenemos nuestra mejor combinación (SEL con emociones generalizadas + balanceo de clases con undersampling) por lo que procedemos a probarla con el perceptrón (ya que este modelo es el que toma más tiempo y desde un punto de vista inicial "el que iba a tener mejor desempeño")		
Sexta prueba (Perceptron)	Perceptron con undersambling y acumulado positivo y negativo	0.48428524377372917
Notamos que efectivamente esta última combinación fue la que tuvo el F1 MACRO mayor, pero aun así para el siguiente paso decidimos hacer la prueba del test con el ultimo perceptrón y la Regresión logística (acumulado positivo y negativo) ya que ambas tenían una diferencia mínima en su desempeño.		

Resultados con el conjunto de prueba

Para el perceptron con balanceo (undersampling) en el conjunto de pruebas SEL

```
Classification Report:
              precision    recall  f1-score   support

           1       0.25      0.55      0.34       104
           2       0.16      0.41      0.23       145
           3       0.21      0.32      0.25       422
           4       0.33      0.44      0.38      1163
           5       0.89      0.69      0.78      4209

    accuracy                           0.60      6043
   macro avg       0.37      0.48      0.40      6043
weighted avg       0.70      0.60      0.64      6043

Confusion Matrix:
[[  57   25   19    2    1]
 [  33   59   35   13    5]
 [  51  103  136   97   35]
 [  37   91  203  506  326]
 [  49   94  254  915 2897]]
```

El modelo tiene buen desempeño en la clase mayoritaria (clase 5) con precisión de 0.89 y F1-score de 0.78, pero su rendimiento es bajo en las clases minoritarias, por ejemplo, F1-score de 0.23 y 0.34 para las clases 2 y 1, respectivamente. Aunque el accuracy global es del 60%, en el macro avg presenta un F1-score de 0.40, lo cual indica un bajo rendimiento general al promediar el balance entre precisión y recall en todas las clases

Para regresion logística con balanceo (undersampling) en el conjunto SEL

```
Reporte de clasificación:
              precision    recall  f1-score   support

           1       0.28      0.64      0.39       104
           2       0.17      0.39      0.23       145
           3       0.23      0.37      0.28       422
           4       0.35      0.46      0.40      1163
           5       0.90      0.69      0.78      4209

    accuracy                           0.61      6043
   macro avg       0.38      0.51      0.42      6043
weighted avg       0.72      0.61      0.65      6043
```

```
Matriz de confusión:
[[  67   21   11    3    2]
 [  34   56   37   10    8]
 [  46   98  156   93   29]
 [  41   79  210  539  294]
 [  55   80  265  913 2896]]
```

El modelo mantiene buen desempeño en la clase mayoritaria (clase 5) con un F1-score de 0.78, pero continúa mostrando bajos valores para las clases minoritarias, como el F1-score de 0.23 en la clase 2. Aunque el accuracy es del 61%, el macro avg presenta un F1-score de 0.42, reflejando por mínimo mejores resultados que el percepton multicapa.

Conclusiones

Esta práctica se centró en la aplicación de modelos de aprendizaje automático para el análisis de opiniones en el corpus **Rest_Mex_2022**. El objetivo principal fue identificar la polaridad de las opiniones expresadas en el corpus, que incluye reseñas de restaurantes, hoteles y atracciones turísticas. Para ello, se implementaron y evaluaron tres modelos: regresión logística, Naive Bayes y perceptrón multicapa.

Según los resultados obtenidos con el conjunto de entrenamiento, el modelo de Naive Bayes presentó un desempeño limitado, alcanzando un puntaje F1 macro no superior a 0.29, es por ello que se descartó para las pruebas posteriores, para el modelo de regresión logística mostró resultados consistentes en todas las pruebas y combinaciones del corpus. En la primera prueba con la versión base del conjunto de datos, obtuvo un puntaje F1 macro de 0.45, mientras que, en la quinta prueba, utilizando datos balanceados y un diccionario, logró mejorar su desempeño alcanzando un puntaje de 0.48.

Como última prueba, y tras analizar el comportamiento de las diferentes combinaciones del corpus, se decidió utilizar el conjunto que combinaba emociones generalizadas con balanceo de clases mediante undersampling para el perceptrón multicapa. Este modelo logró su mejor

desempeño, alcanzando un F1 Score Macro de 0.4842, superando ligeramente a la regresión logística.

Dado el buen desempeño tanto del modelo de regresión logística como del perceptrón multicapa con sus respectivos conjuntos de entrenamiento, se procedió a realizar la prueba final utilizando el conjunto de prueba. En este caso, la regresión logística superó apenas al perceptrón multicapa en términos de F1 Score Macro, destacando por una diferencia mínima en su rendimiento. Para este caso la regresión logística no solo tiene una ventaja en su score, sino que también su ejecución es considerablemente más rápida, que lo convierte en nuestro mejor modelo para esta práctica.

El análisis de los resultados obtenidos sugiere que es necesario continuar investigando para mejorar la precisión de los modelos, como por ejemplo explorar otras técnicas para el desbalance de clases; experimentar con algún otro diccionario para la división de opiniones, etc.