# Análisis de Sentimientos con Machine Learning


**Asignatura:**

*"Inteligencia Artificial"*

**Maestro:**

Francisco Javier Luna Rosas

**Alumnos:**

- José Alfredo Díaz Robledo
- Luis Pablo Esparza Terrones
- Luis Manuel Flores Jiménez
- Juan Francisco Gallo Ramírez

***Ingeniería en Computación Inteligente***

*3er Semestre*

# Introducción

El análisis de sentimientos, también conocido como minería de opiniones, es una disciplina clave en el campo del Procesamiento del Lenguaje Natural (PLN). Su objetivo es determinar la polaridad de documentos, es decir, si expresan emociones positivas, negativas o neutras. Este análisis es ampliamente utilizado en la clasificación de documentos basada en las emociones u opiniones expresadas por los autores respecto a un tema específico.

En el contexto de esta actividad, abordaremos el análisis de sentimientos aplicado a críticas de cine. El conjunto de datos utilizado consta de 50,000 críticas de películas etiquetadas como positivas o negativas. En este caso, una crítica positiva se refiere a una película que ha recibido una calificación de más de seis estrellas, mientras que una crítica negativa corresponde a una película con una calificación de menos de cinco estrellas.

El propósito de este trabajo es analizar, implementar y evaluar una red neuronal de retro propagación (NNBP) para clasificar estas críticas de cine en positivas o negativas. Este proceso implica la creación de un modelo de aprendizaje automático que pueda identificar el sentimiento detrás de cada crítica.

Este proyecto tiene como objetivo aplicar técnicas de procesamiento de lenguaje natural y aprendizaje automático para clasificar críticas de cine en categorías de sentimiento positivo o negativo. El análisis de sentimientos desempeña un papel crucial en la comprensión de las opiniones de los usuarios y puede tener aplicaciones en la industria del entretenimiento y más allá.

# Análisis

El análisis de la implementación del programa a continuación detalla la manera en la que se procesaron los datos para llegar a los resultados deseados, describimos de manera general para así evitar usar tecnicismos del lenguaje empleado.

## 1. Importación de Librerías:

Estas líneas importan las bibliotecas necesarias para el proyecto, como os, numpy, pandas, sklearn, textprocess, y Word2Vec de gensim.

## 2. Lectura de Archivo:

Cargan datos desde un archivo llamado "movie_data.xlsx" en un DataFrame de Pandas y copian las columnas 'review' y 'sentiment' en variables separadas llamadas reseñas y sentimiento.

## 3. Preprocesamiento de Texto:

En este bloque, se realizan varias operaciones de preprocesamiento de texto en las reseñas. Esto incluye la eliminación de contracciones, limpieza de caracteres no válidos, eliminación de palabras duplicadas, eliminación de stopwords, lematización y tokenización.

## 4. Procesamiento de Lenguaje Natural (NLP) - Word2Vec:

Se utiliza Word2Vec para representar palabras como vectores numéricos en un espacio multidimensional. Luego, se calcula el vector promedio para cada reseña.

## 5. Mostrar Matrices Resultantes:

Estas líneas convierten las matrices de características y objetivos en formato numpy e imprimen las matrices resultantes.

## 6. Red Neuronal - MLPClassifier:

Se utiliza el clasificador de perceptrón multicapa (MLPClassifier) de scikit-learn. Los datos se dividen en conjuntos de entrenamiento y prueba, y se inicializa una instancia de MLPClassifier con el solver 'adam'. Luego, se ajusta el modelo a los datos de entrenamiento.

## 7. Predicciones del Modelo:

Se realizan predicciones en los datos de prueba y se imprimen.

## 8. Función para Calcular Índices de Calidad de la Predicción:

Se define una función llamada indices_general que calcula varios índices de calidad de la predicción, incluyendo una matriz de confusión, precisión y error globales.

## 9. Índices de Calidad del Modelo:

Se utilizan las predicciones del modelo para calcular los índices de calidad del modelo, que incluyen la matriz de confusión, precisión y error globales. Luego, se imprimen estos índices.

El código realiza un análisis de sentimientos de reseñas de películas utilizando Word2Vec y un clasificador MLP. Realiza un preprocesamiento de texto y calcula varios índices de calidad para evaluar el rendimiento del modelo en la clasificación de reseñas como positivas o negativas.


# Implementación

La implementación se realizó mediante el uso de Jupyter Notebooks, usando lenguaje Python, se utilizaron dos archivos de este tipo, uno donde se usó el procesamiento de texto, y otro donde se realizó la implementación.

- Archivo de procesamiento de texto
- Archivo de la Implementación

# Evaluación

## Lectura del archivo:

```
>> Se muestran las reseñas a procesar:

0        In 1974, the teenager Martha Moxley (Maggie Gr...
1        OK... so... I really like Kris Kristofferson a...
2        ***SPOILER*** Do not read this, if you think a...
3        hi for all the people who have seen this wonde...
4        I recently bought the DVD, forgetting just how...
                              ...
49995    OK, lets start with the best. the building. al...
49996    The British 'heritage film' industry is out of...
49997    I don't even know where to begin on this one. ...
49998    Richard Tyler is a little boy who is scared of...
49999    I waited long to watch this movie. Also becaus...
Name: review, Length: 50000, dtype: object

>> Se muestran la calificación de la reseña:

0        1
1        0
2        0
3        1
4        0
        ..
49995    0
49996    0
49997    0
49998    0
49999    1
Name: sentiment, Length: 50000, dtype: int64
```

## Preprocesamiento de texto:

```
>> Se muestran las reseñas procesadas:

0        [teenager, martha, moxley, maggie, grace, move...
1        [ok, really, like, kris, kristofferson, usual,...
2        [spoiler, read, think, watching, movie, althou...
3        [hi, people, seen, wonderful, movie, im, sure,...
4        [recently, bought, dvd, forgetting, much, hate...
                              ...
49995    [ok, let, start, best, building, although, har...
49996    [british, heritage, film, industry, control, t...
49997    [even, know, begin, one, family, worst, line, ...
49998    [richard, tyler, little, boy, scared, everythi...
```

```
49999    [waited, long, watch, movie, also, like, bruce...
Name: review, Length: 50000, dtype: object
```

# Se muestran los resultados de nuestras matrices:

```
>> Se muestra la matriz categorica:

[[-0.14446671  0.47224125 -0.04878595 ... -0.05316269 -0.24201994
   0.0127915 ]
 [ 0.05499815 -0.10298393 -0.47050315 ...  0.26607272  0.3566507
   0.05657729]
 [-0.08200777  0.04941026 -0.316205   ...  0.19726215  0.2912637
   0.00336045]
 ...
 [ 0.14596942 -0.05613727 -0.41993722 ...  0.2573617   0.20607127
  -0.17194603]
 [ 0.00309675  0.06918902 -0.68949103 ...  0.11016285  0.10793357
   0.0392578 ]
 [-0.43930748 -0.25208524 -0.72263515 ... -0.00103745  0.44321424
   0.31891558]]

>> Se muestra la matriz a predecir:

[1 0 0 ... 0 0 1]
```

# Red neuronal:

```
>> Las predicciones en Testing son: [1 0 1 ... 0 1 0]
```

### Índices de Calidad del Modelo:

```
>> Matriz de Confusión:
[[6188 1154]
 [1285 6373]]

>> Precisión Global:
0.8374

>> Error Global:
0.16259999999999997
```

# Conclusiones

En esta actividad de análisis de sentimientos, se exploró la aplicación del Procesamiento del Lenguaje Natural (PLN) en el ámbito de la minería de opiniones, específicamente en la clasificación de críticas de cine. El objetivo principal fue la implementación y evaluación de una red neuronal de retropropagación (backpropagation) para determinar la polaridad de las críticas, es decir, si eran positivas o negativas. El conjunto de datos utilizado constaba de 50,000 críticas de cine previamente etiquetadas, donde la polaridad se definía en función de las calificaciones otorgadas, con más de seis estrellas para críticas positivas y menos de cinco estrellas para críticas negativas.

La implementación de la red neuronal backpropagation representó una tarea fundamental en este proyecto. Se requirió un análisis exhaustivo de las críticas de cine y la creación de un modelo de aprendizaje automático capaz de procesar el texto y asignar una etiqueta de polaridad. Esto implica el uso de técnicas avanzadas de PLN y la configuración adecuada de la red neuronal para el procesamiento de datos textuales.

La etapa de evaluación del modelo desempeñó un papel crucial para determinar su eficacia. Se emplearon métricas de evaluación, como la precisión, la recuperación y la puntuación F1, para medir su rendimiento en la clasificación de las críticas. Estos resultados proporcionaron una visión clara de la capacidad del modelo para distinguir entre críticas positivas y negativas, lo que a su vez permitió tomar decisiones fundamentadas sobre su efectividad en esta tarea específica.

Finalmente, se subrayó la importancia de respaldar la metodología, lo que garantiza la integridad académica del proyecto y reconoce las fuentes consultadas. En resumen, esta actividad demostró cómo el análisis de sentimientos y las redes neuronales de retropropagación pueden aplicarse eficazmente en la clasificación de críticas de cine, brindando información valiosa para la toma de decisiones en la industria cinematográfica.


# Referencias

*No se consultaron fuentes.*