Universidad Autónoma de Aguascalientes

Centro de Ciencias Básicas

Departamento de Ciencias de la Computación

Optativa Profesionalizante II: Machine Learning y Deep Learning

10° "A"

Tercera Evaluación Parcial

Docente: Dr. Francisco Javier Luna Rosas

Alumno: Joel Alejandro Espinoza Sánchez (211800)

Fecha de Entrega: Aguascalientes, Ags., 2 de junio del 2023.

El análisis de sentimientos, a veces también denominado minería de opiniones, es una conocida sub-disciplina del amplio campo del PLN (Procesamiento del Lenguaje Natural); está relacionado con el análisis de la polaridad de documentos. Una tarea popular en el análisis de sentimiento es la clasificación de documentos basados en las emociones u opiniones expresadas de los autores respecto a un tema en particular. El conjunto de datos de críticas de cine consiste en 50,000 críticas de cine polarizadas etiquetadas como negativas y como positivas. Aquí, positiva significa que una película ha sido clasificada con más de seis estrellas, mientras que negativa significa que una película ha sido clasificada con menos de cinco estrellas.

El alumno deberá elaborar un documento (`*.pdf`) y un archivo auto-reproducible (`*.html`) que analice, implemente y evalúe algoritmos de Deep Learning y Machine Learning para clasificar las críticas de cine. El documento deberá contener:

- Portada
- Evidencias del examen
- Conclusiones
- Referencias (formato APA)

---

a) Una explicación del preprocesamiento de datos para generar un formato adecuado de los datos

Primeramente se cargan los datos que se usarán:

In [1]:
```
# Importa las bibliotecas necesarias, como TensorFlow y Keras

import pandas as pd
import numpy as np
from tensorflow import keras
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, LSTM, Dense

# Carga el archivo CSV que contiene las críticas y sus etiquetas utilizando pandas

data = pd.read_csv("movie_data.csv")
```

En este paso se suelen aplicar las siguientes estrategias para darle una mejor forma al conjunto de datos:

1. **Eliminar los datos irrelevantes:** Eliminar los datos que no son útiles para el análisis de opiniones, como las fechas, las direcciones, etc.
2. **Limpiar el texto:** Eliminar caracteres especiales, signos de puntuación, números y otros caracteres no alfabéticos que no aportan información útil para el análisis.
3. **Convertir el texto a minúsculas:** Convertir todo el texto a minúsculas para facilitar el procesamiento y para que no se distingan las palabras en mayúsculas y minúsculas.
4. **Eliminar las palabras comunes:** Eliminar las palabras comunes que no aportan información útil para el análisis, como artículos, preposiciones, conjunciones, etc.
5. **Lematización:** Lematizar el texto para reducir las palabras a su forma base, lo que puede ayudar a reducir la complejidad del análisis y aumentar la precisión.
6. **Eliminar palabras clave irrelevantes:** Eliminar palabras clave irrelevantes que no aportan información útil para el análisis.
7. **Tokenización:** Separar el texto en palabras individuales para su posterior análisis.
8. **Normalización:** Normalizar el texto para que todas las palabras estén en la misma forma, como eliminar los acentos o caracteres especiales.
9. **Análisis de sentimientos:** Realizar un análisis de sentimientos para asignar a cada opinión un valor de positividad, negatividad o neutralidad.

Se realiza un preprocesamiento con la clase Tokenizer de la librería Keras convirtiendo el texto en tokens y en secuencias:

In [3]:
```
# Preprocesa los datos

reviews = data['review'].values
labels = data['sentiment'].values
```

b) Una explicación del modelo bolsa de palabras o cualquier otro analizador ligústico aplicado al Dataset de críticas de cine

La librería Keras de TensorFlow proporciona la clase Tokenizer y la función pad_sequences para procesar texto y convertirlo en una representación numérica que puede ser utilizada como entrada para modelos de aprendizaje automático.

El modelo de bolsa de palabras que se usó fue dado por TensorFlow de su apartado de Keras siendo el método `Tokenizer` que se utiliza para convertir el texto en una secuencia de enteros (tokens) asignando un número único a cada palabra en el texto. Esto es útil porque los modelos de aprendizaje automático sólo pueden trabajar con datos numéricos, por lo que se debe convertir el texto en una forma numérica para poder ser procesado.

Su implementación se presenta a continuación:

In [4]:
```
# Tokeniza los textos utilizando la clase Tokenizer de Keras

tokenizer = Tokenizer()
tokenizer.fit_on_texts(reviews)
sequences = tokenizer.texts_to_sequences(reviews)
```

c) Una explicación de la transformación de las palabras en vectores de características (utilice la frecuencia de termino - frecuencia inversa de documento "tf-idf" o cualquier otra técnica que permita transformar palabras a vectores de características)

Por otro lado, la transformación de palabras en vectores de características se usa nuevamente TensorFlow con el uso de Keras usando `pad_sequences` que se utiliza para igualar la longitud de las secuencias numéricas resultantes de Tokenizer. Dado que las secuencias de palabras en el texto pueden tener diferentes longitudes, es necesario igualarlas para poder procesarlas en el modelo de aprendizaje automático. pad_sequences se encarga de añadir ceros (o cualquier otro valor definido) al principio o al final de las secuencias para que todas tengan la misma longitud.

La implementación y su uso se presenta en el siguiente segmento:

In [5]:
```
# Establece la longitud máxima de las secuencias

max_sequence_length = 100  # Longitud máxima de las secuencias
padded_sequences = pad_sequences(sequences, maxlen=max_sequence_length)
```

d) Una explicación del modelo CNN (Convoluciones 1D) para clasificar las críticas de cine. La precisión del modelo debe ser del 95% o mayor

A continuación se muestra la implementación del modelo CNN preparando los conjuntos de datos de aprendizaje supervisado y declarando la arquitectura de la red como se muestra en el siguiente segmento de código:

In [6]:
```
# Dividir los datos en conjuntos de entrenamiento y prueba

split_ratio = 0.8  # Proporción de entrenamiento-prueba
split_index = int(split_ratio * len(padded_sequences))

x_train = padded_sequences[:split_index]
y_train = labels[:split_index]
x_test = padded_sequences[split_index:]
y_test = labels[split_index:]

# Construye el modelo de la red neuronal LSTM

vocab_size = len(tokenizer.word_index) + 1  # Tamaño del vocabulario
embedding_dim = 100  # Dimensión de la capa de embedding

model = Sequential()
model.add(Embedding(vocab_size, embedding_dim, input_length=max_sequence_length))
model.add(LSTM(128))
model.add(Dense(1, activation='sigmoid'))

model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Entrena el modelo con los datos de entrenamiento

model.fit(x_train, y_train, epochs=10, batch_size=16)

# Evalúa el modelo con los datos de prueba

loss, accuracy = model.evaluate(x_test, y_test)
print('Accuracy:', accuracy)
```

```
Epoch 1/10
2500/2500 [==============================] - 787s 314ms/step - loss: 0.3866 - accurac
y: 0.8285
Epoch 2/10
2500/2500 [==============================] - 747s 299ms/step - loss: 0.2011 - accurac
y: 0.9237
Epoch 3/10
2500/2500 [==============================] - 690s 276ms/step - loss: 0.1075 - accurac
y: 0.9626
Epoch 4/10
2500/2500 [==============================] - 724s 289ms/step - loss: 0.0592 - accurac
y: 0.9804
Epoch 5/10
2500/2500 [==============================] - 596s 238ms/step - loss: 0.0352 - accurac
y: 0.9888
Epoch 6/10
2500/2500 [==============================] - 543s 217ms/step - loss: 0.0214 - accurac
y: 0.9934
Epoch 7/10
2500/2500 [==============================] - 545s 218ms/step - loss: 0.0156 - accurac
y: 0.9950
Epoch 8/10
2500/2500 [==============================] - 544s 218ms/step - loss: 0.0091 - accurac
y: 0.9969
Epoch 9/10
2500/2500 [==============================] - 544s 218ms/step - loss: 0.0060 - accurac
y: 0.9984
Epoch 10/10
2500/2500 [==============================] - 555s 222ms/step - loss: 0.0072 - accurac
y: 0.9978
313/313 [==============================] - 7s 21ms/step - loss: 0.8612 - accuracy: 0.
8410
Accuracy: 0.8410000205039978
```

e) Una explicación del análisis comparativo de modelos de Deep Learning (Redes Neuronales, Redes RNN) Vs Machine Learning (KNN, Bayes, Arboles de Decisión, SVM, Random Forest, Potenciación, etc.), compare al menos dos clasificadores de cada uno con una precisión del 95% o mayor, el análisis deberá comparar: la precisión del modelo, el error del modelo, precisión negativa (especificidad), precisión positiva (sensibilidad), falsos positivos, falsos negativos, asertividad positiva, asertividad negativa

A continuación se realiza la comparación del modelo implementando en código un modelo Naive Bayes como se aprecia a continuación con las implementaciones de la librería Naive Bayes:

In [9]:
```
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import classification_report
from sklearn.metrics import precision_score
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(padded_sequences, labels, test_siz
```

```
classifier = GaussianNB()
classifier.fit(X_train, y_train)

y_pred = classifier.predict(X_test)

matriz = confusion_matrix(y_test, y_pred)
print('Matriz de Confusión')
print(matriz)

precision = precision_score(y_test, y_pred, pos_label="a", average=None)
print('Precisión del modelo')
print(precision)
```

```
Matriz de Confusión
[[3542 1492]
 [3390 1576]]
Precisión del modelo
[0.51096365 0.5136897 ]
```

```
C:\Users\alexe\Anaconda3\envs\ici-thesis\lib\site-packages\sklearn\metrics\_classific
ation.py:1375: UserWarning: Note that pos_label (set to 'a') is ignored when average
!= 'binary' (got None). You may use labels=[pos_label] to specify a single positive c
lass.
  UserWarning,
```

Podemos observar que la precisión positiva y negativa del modelo utilizado el cual es Naive Bayes, es inferior a la precisión el modelo de la red neuronal convolucional presentado en el presente documento con precisiones cercanas al 50% lo cual no demuestra una gran fiabilidad para usar este modelo en contraste con la red neuronal recurrente que se acerca a un 85% sin embargo este modelo se tarda mucho en procesarse por lo que es necesario considerar qué se desea sacrificar al menos entre estos dos modelos.

Conclusiones

Es interesante e importante poder implementar las bases de estos temas para entenderlos en un futuro, pues, posteriormente no basta con sólo importar librerías que realicen el trabajo pesado, ya que, implementar manualmente estos algoritmos nos enseña a qué hay detrás del algoritmo, cómo funciona y poder comprender realmente qué está ocurriendo como la base de una red neuronal convolucional y la forma en la que ésta aprende. Es muy útil la implementación de estos algoritmos en estas tareas para la vida profesional que nos perpara esta materia de Machine Learning ahora en el final de la carrera.

Referencias

- Anónimo (s.f.) “Red neuronal artificial”. Obtenido de Wikipedia: https://es.wikipedia.org/wiki/Red_neuronal_artificial.
- Data Scientest (2021) “Perceptrón. ¿Qué es y para qué sirve?”. Obtenido de Data Scientest: https://datascientest.com/es/perceptron-que-es-y-para-que-sirve.
- Luna, F. (2023) “El Modelo de McCulloch – Pitts”. Apuntes de ICI 10°.