**Un conjunto de letras negras en un fondo negro

Descripción generada automáticamente con confianza bajaIcono

Descripción generada automáticamente con confianza media**

**Instituto Politécnico Nacional**

**Escuela Superior de Cómputo**

**Ingeniería en Sistemas Computacionales**

**Prof. Joel Omar Juárez Gambino**

**Practice VI.**

**(Sentiment Analysis)**

**Equipo 2:**

**Rodríguez Olmos Noé**

**Rosales Sedano Joaquín**

**López Cedano Axel Adrián**

**Silva Martínez Miguel Ángelo**

**Grupo: 7CM3**

**Procesamiento de Lenguaje Natural**

**09 de mayo de 2024**

**Introducción**

En esta práctica, se busca determinar la polaridad de las opiniones de los usuarios sobre sitios turísticos. El corpus de datos incluye reseñas de restaurantes, hoteles y atracciones turísticas. Cada opinión está valorada en una escala del 1 al 5, donde:

* 1 representa una opinión muy negativa,
* 2 representa una opinión negativa,
* 3 representa una opinión neutral,
* 4 representa una opinión positiva,
* 5 representa una opinión muy positiva.

El objetivo de esta práctica es mejorar el desempeño obtenido por un modelo básico compartido, mediante técnicas avanzadas de preprocesamiento y análisis de texto. Esto incluye normalizar el texto, manejar negaciones y emojis, balancear las clases desbalanceadas y optimizar los hiperparámetros del modelo de aprendizaje automático. A través de estos métodos, se busca proporcionar un análisis más preciso y robusto de la polaridad del sentimiento en las opiniones de los usuarios.

**Metodología**

#### Preprocesamiento del Corpus

El preprocesamiento del texto es un paso fundamental para preparar los datos antes de entrenar los modelos de aprendizaje automático. Los pasos de preprocesamiento aplicados incluyen:

1. **Normalización del Texto**:
   * Convertir todo el texto a minúsculas para garantizar la uniformidad y reducir la dimensionalidad.
   * Separar los signos de puntuación del texto, lo que permite una tokenización más precisa.
   * Eliminar las stop words utilizando la librería nltk, eliminando palabras comunes que no aportan valor significativo al análisis del sentimiento.
2. **Manejo de Negaciones**:
   * Identificar palabras de negación y concatenarlas con las palabras siguientes para captar el efecto de la negación en el sentimiento del texto.
3. **Integración de Emojis**:
   * Utilizar un léxico de emojis para identificar y representar los sentimientos asociados con los emojis presentes en el texto.
4. **Limpieza Adicional**:
   * Eliminar caracteres no alfabéticos para centrarse únicamente en el texto relevante.

#### Representación del texto

Para la representación del texto, se evaluaron varias técnicas. La representación que mejor desempeño mostró fue el **TF-IDF (Term Frequency-Inverse Document Frequency)**, que mide la importancia de una palabra en un documento en relación con un corpus, proporcionando una representación numérica útil para los modelos de aprendizaje automático.

#### Entrenamiento del Modelo de Aprendizaje Automático

Se probaron varios modelos de aprendizaje automático para determinar cuál proporcionaba los mejores resultados en la tarea de clasificación de sentimientos. Después de comparar varios modelos, se determinó que el **modelo de Regresión Logística** era el más eficaz. Se aplicaron las siguientes técnicas:

1. **Vectorización**:
   * El texto preprocesado se transformó utilizando el vectorizador TF-IDF.
2. **Balanceo de Clases**:
   * Para abordar el problema de desbalanceo de clases en el corpus, se utilizó SMOTE (Synthetic Minority Over-sampling Technique), una técnica que genera muestras sintéticas de la clase minoritaria para equilibrar el conjunto de entrenamiento.
3. **Entrenamiento del Modelo**:
   * Se entrenó el modelo de Regresión Logística con los datos vectorizados y balanceados, utilizando hiperparámetros optimizados (*C=5.1, max\_iter=100, penalty='l2', solver='saga').*

#### Predicción de la Polaridad del Sentimiento

Después de entrenar el modelo, se realizaron predicciones sobre el conjunto de prueba y se evaluó su desempeño utilizando varias métricas, incluyendo el F1 macro score, el reporte de clasificación y la matriz de confusión.

**Mejoras propuestas**

#### Aumento de Características

Para mejorar la precisión del modelo, se incorporaron características adicionales relacionadas con el manejo de negaciones y la integración de sentimientos asociados con emojis.

#### Pasos Adicionales de Preprocesamiento

1. **Manejo de Negaciones**:
   * Se introdujo una técnica para identificar y manipular negaciones en el texto, mejorando la precisión del análisis de sentimientos.
2. **Interpolación de Datos con SMOTE**:
   * Se utilizó SMOTE para balancear las clases del conjunto de entrenamiento, lo que ayudó a mejorar el desempeño del modelo en conjuntos de datos desbalanceados.

#### Otros

* **Manipulación de Hiperparámetros**:
  + Se optimizaron los hiperparámetros del modelo de Regresión Logística para mejorar su desempeño, resultando en una mejor clasificación de la polaridad del sentimiento.

**Resultados del entrenamiento de cada experimento**

#### Binarized

• Puntuación F promedio

#### Frequency

• Puntuación F promedio

#### TF-IDF

• Puntuación F promedio

#### Embeddings

• Puntuación F promedio

**Resultados de la prueba**

#### Binarized

• Puntuación F promedio  
• Informe de clasificación   
• Matriz de confusión

#### Frequency

• Puntuación F promedio  
• Informe de clasificación   
• Matriz de confusión

#### TF-IDF

• Puntuación F promedio  
• Informe de clasificación   
• Matriz de confusión

#### Embeddings

• Puntuación F promedio  
• Informe de clasificación   
• Matriz de confusión

**Conclusiones:**En este estudio, se evaluaron varios enfoques para la representación del texto y modelos de aprendizaje automático para la clasificación de la polaridad de opiniones de usuarios sobre sitios turísticos. Las representaciones de texto evaluadas incluyeron binarización, frecuencia, TF-IDF y embeddings, y los resultados se compararon para determinar la eficacia de cada enfoque. A continuación se detallan los resultados obtenidos para cada una de las representaciones de texto:

* **Resultados**
  + **Binarized**

La representación binarizada, que simplemente indica la presencia o ausencia de palabras, fue efectiva en algunos casos, pero en general mostró limitaciones al no capturar la frecuencia o importancia relativa de las palabras en el corpus. Aunque es una técnica simple y rápida, los modelos basados en esta representación no lograron un rendimiento óptimo en términos de precisión y F1 macro score.

* + **Frecuency**

La representación basada en la frecuencia de palabras proporcionó una mejora con respecto a la binarización, ya que considera cuántas veces aparece cada palabra en el texto. Esto permitió al modelo captar mejor la importancia de palabras más comunes en determinadas opiniones. Sin embargo, seguía sin abordar completamente la relevancia relativa de las palabras en el contexto del corpus más amplio.

* + **TF - IDF**

La representación TF-IDF (Term Frequency-Inverse Document Frequency) demostró ser la más efectiva en este estudio. Al medir la importancia de una palabra en un documento en relación con el corpus completo, TF-IDF logró capturar mejor las diferencias sutiles en el lenguaje de las opiniones. Los modelos entrenados con esta representación mostraron el mejor desempeño, alcanzando altos valores en el F1 macro score y ofreciendo una clasificación precisa y robusta de la polaridad del sentimiento.

* + **Embeddings**

Aunque no se implementó en este código, el uso de embeddings (vectores densos que representan el significado semántico de las palabras) es una técnica avanzada que puede capturar relaciones semánticas complejas entre palabras y frases. Los embeddings, como los generados por modelos preentrenados tipo Word2Vec, GloVe o BERT, podrían ofrecer mejoras significativas en el análisis de sentimientos al considerar el contexto y las similitudes semánticas de las palabras en las opiniones.

* **Mejoras Futuras:**
  + **Implementación de Embeddings:** Integrar modelos de embeddings como Word2Vec, GloVe o BERT para capturar relaciones semánticas y contextuales más complejas en las opiniones.
  + **Modelos de Aprendizaje Profundo:** Explorar el uso de redes neuronales profundas, como RNNs, LSTMs y transformers, que han mostrado un gran potencial en tareas de procesamiento de lenguaje natural.
  + **Mejoras en Preprocesamiento:** Incorporar técnicas más avanzadas de manejo de negaciones y emociones, así como el análisis de sarcasmo e ironía, que son comunes en opiniones subjetivas.
  + **Aumento de Datos:** Ampliar el corpus con más datos de opiniones de diferentes fuentes y dominios, lo que podría mejorar la generalización del modelo.
  + **Ajuste de Hiperparámetros:** Continuar afinando los hiperparámetros de los modelos de aprendizaje automático para optimizar su desempeño en diferentes conjuntos de datos y representaciones de texto.
  + **Validación Cruzada y Evaluación Continua:** Utilizar técnicas de validación cruzada para una evaluación más robusta y continua del desempeño del modelo, asegurando que las mejoras se mantengan consistentes en diferentes particiones del conjunto de datos.