**Informe Práctica 01 - Análisis de sentimientos**

**Equipo 5 - Integrantes:**

* Stefany Cantero Cardenas
* Juan Manuel Vera Echeverri
* Sebastian Arango Urrea

**Descripción de la creación del corpus**

Para crear el corpus del proyecto se hizo uso de comentarios en foros de discusión online. Estos fueron ElOtroLado.net para las categorías de textos relacionados con cultura y política, y Rankia.com en la sección de Bolsa para aquellos relacionados con economía. Para extraer los datos se implementaron las librerías BeautifulSoup, Playwright y Scrapy. Adicionalmente, se descargaron los tweets contenidos en los archivos .xml de TASS y se seleccionó el 20% correspondiente. Una vez realizado esto, se obtuvo la polaridad usando la librería Text Blob y se distribuyeron los datos en las carpetas positivo y negativo de acuerdo a su resultado, para obtener al final un total de 7368 datos en el corpus definitivo. Finalmente, se creó un dataframe con las columnas ‘text’ y ‘sentiment’ con el cual se trabajó en las siguientes etapas de la solución.

**Descripción del preprocesamiento, tokenización, reducción , normalización y extracción**

En el preprocesamiento se creó una función para limpiar el texto en la que se llevaron a cabo los procesos de normalización del texto a minúsculas y eliminación de números, tildes, caracteres especiales, signos de puntuación y enlaces. Esta limpieza fue aplicada al dataframe resultante de la creación del corpus.

La tokenización y reducción se hizo con ayuda de la librería NLTK. En esta parte, se implementó una función para obtener los n-gramas y conformar la bolsa de palabras. Hecho esto, se crearon e implementaron las funciones para remover stopwords, n-gramas con frecuencia menor a 3 y realizar stemming. Como resultado, se obtuvieron las cuatro variaciones del corpus necesarias para realizar las aplicaciones de las técnicas de aprendizaje.

**Descripción de la ponderación de características**

Las dos ponderaciones necesarias para la comparación se hicieron usando CountVectorizer() para la Frecuencia Absoluta (TO) y TfidVectorizer para el esquema TF-IDF en las variaciones del corpus.

**Descripción de la configuración de las técnicas de aprendizaje utilizadas para el análisis de sentimientos y cómo se obtuvieron los resultados de las tablas**

Se implementaron tres algoritmos: árboles de decisión, KNN y regresión logística. El particionamiento de los datos fue de 80% para entrenamiento y 20% para validación. Se hizo uso de los resultados de la ponderación de características como entrenamiento y prueba, además del sentimiento como variable objetivo tomando valores en 1 para positivo y 0 para negativo.

Se utilizaron las librerías de scikit-learn para realizar la validación cruzada con k=10 y para obtener las métricas de evaluación de cada algoritmo. Comparando los valores obtenidos con los reales, en una función se ejecutaron todos algoritmos de manera dinámica, con lo cual se llenó un dataframe con todos los valores que se necesitaban.

De esta manera obtuvimos los datos para hacer las comparativas y observar las diferencias entre la efectividad de usar estas técnicas para el preprocesamiento, con múltiples algoritmos de predicción para el análisis de sentimientos.

**Conclusiones de los resultados**

De las tres técnicas de aprendizaje implementadas, la que tuvo mejor rendimiento general fue la Regresión Logística con valores superiores a 80% para cada una de las 4 variaciones del corpus.

En cuanto al F1-Score el máximo valor lo tuvo la regresión logística con un 84.17%. Seguido de este estuvieron los árboles de decisión con un 72.12% y, finalmente, con un resultado similar KNN con 71.65%. De esto, se puede decir que la regresión logística es la

Por técnica de reducción, la regresión es la técnica que mejor se ajusta a los datos recolectados y que tanto precision como recall tienen valores similares en cada técnica de reducción en cuanto al valor de F1-Score

La ponderación TO (frecuencia absoluta) fue aquella que presentó mejores resultados para las tres técnicas de aprendizaje. En el caso de la regresión logística, esta obtuvo valores muy similares de F1-Score con ambas ponderaciones y el mayor valor de las tres técnicas. En cuanto a los árboles de decisión y KNN, estos también obtuvieron valores similares en TO. Mientras que en TF-IDF los árboles de decisión presentaron un valor de F1-Score ligeramente mayor a KNN.