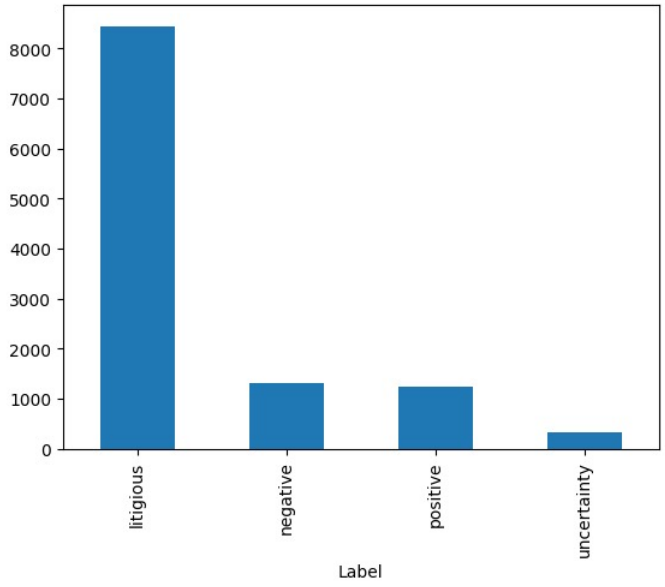
CI-0134: Investigación en Ciencias de la Computación
Emmanuel Zúñiga - B98729

Semana 5: (17 al 21 de marzo)

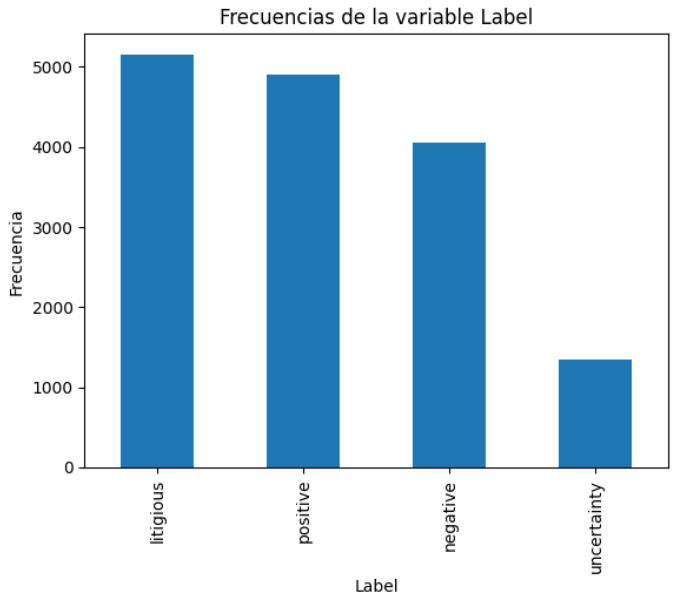
Tareas realizadas:

- ☑ Debido al poco balanceo en el dataset, se decide realizar una integración en conjunto a otros dos datasets con el objetivo de tener un conjunto de datos más balanceado en cuanto a las categorías; a continuación, se presentan los gráficos que muestran el antes y el después de las frecuencias de las clases, con y sin la integración.

  Antes:



  Después:



  Los datasets utilizados fueron obtenidos en la página del Taller de Análisis Semántico en la SEPLN: http://tass.sepln.org/tass_data/download.php?auth=stmaHsRs54EeZv6ebrc

Dichos datasets fueron adaptados de formato XML a CSV y posteriormente, se realizó el mapeo de las clases de dicho dataset a las clases del dataset original del trabajo de investigación.

Queda pendiente conversar con los profesores para incluir datos de tweets de costarricenses para tropicalizar los modelos, dado que la página ofrece un dataset con dichas características. Por otro lado, se plantea eliminar la categoría “uncertainty” para mantener un mejor balance en los datos.

El notebook con la integración de los datos puede consultarse en la carpeta notebooks en el repositorio del proyecto.

Para la otra semana:

- Realizar la experimentación con el dataset generado y evaluación de otras estadísticas como F1.
- Redacción de la metodología.
- Revisión del trabajo relacionado para guiarse en la redacción de la metodología.