RealCheck
Informe Semanal - 3
Objetivos del sprint

- Integrar la búsqueda y selección de fuentes para su posterior análisis.

Resumen

Durante esta semana, se trabajó en la implementación de un algoritmo de similitud de textos para el proyecto de análisis de noticias. También se realizaron mejoras en el sistema de recolección de datos, incluyendo el manejo de errores en las solicitudes de las fuentes y la extracción de los títulos completos de las noticias. Además, se realizaron pruebas con diferentes métodos de procesamiento de lenguaje natural para seleccionar las fuentes más relevantes, y se optó por el uso del modelo pre-entrenado de FastText para obtener mejores resultados. En general, el proyecto sigue avanzando un poco lento según lo planeado pero se espera continuar mejorando la calidad de los resultados obtenidos.

Tareas en progreso	Tareas completadas
● RC-9 Elaborar el documento de arquitectura de software (SAD) ● RC-11 Diseño e implementación de la base de datos	● RC-10 Incorporar la búsqueda de fuentes ● RC-12 Reporte semanal

Problemas y Riesgos:

- Se presentaron problemas con el motor de búsqueda utilizado, ya que no entregaba las peticiones en formato JSON, lo que provocaba que los nombres de los artículos aparecieran incompletos. Por esta razón, se decidió hacer peticiones GET a cada fuente para obtener el título completo de la noticia. Sin embargo, el algoritmo utilizado no realiza scrapping al 100% de las posibles fuentes encontradas, y algunos sitios devuelven un mensaje de "Access Denied". Como resultado, se estima que se está perdiendo aproximadamente el 6% de las fuentes totales encontradas.

Comentarios y observaciones:

- Se llevó a cabo un proceso exhaustivo de exploración y evaluación de diversas técnicas de procesamiento de lenguaje natural para seleccionar las 10 fuentes más adecuadas para el análisis. Se empleó un modelo pre-entrenado de BERT, sin embargo, los resultados obtenidos fueron insatisfactorios. Se procedió a ajustar el modelo mediante una modificación que mejoró ligeramente las probabilidades. También se evaluó el modelo por secuencia difusa, el cual ofreció mejores resultados desde el inicio. Al final, se optó por utilizar el modelo pre-entrenado de FastText denominado "cc.es.300.bin" que arrojó resultados superiores a los modelos anteriores.