

Ingeniería de Requerimientos acelerada con Cómputo Paralelo aplicada a Idiomas Indígenas

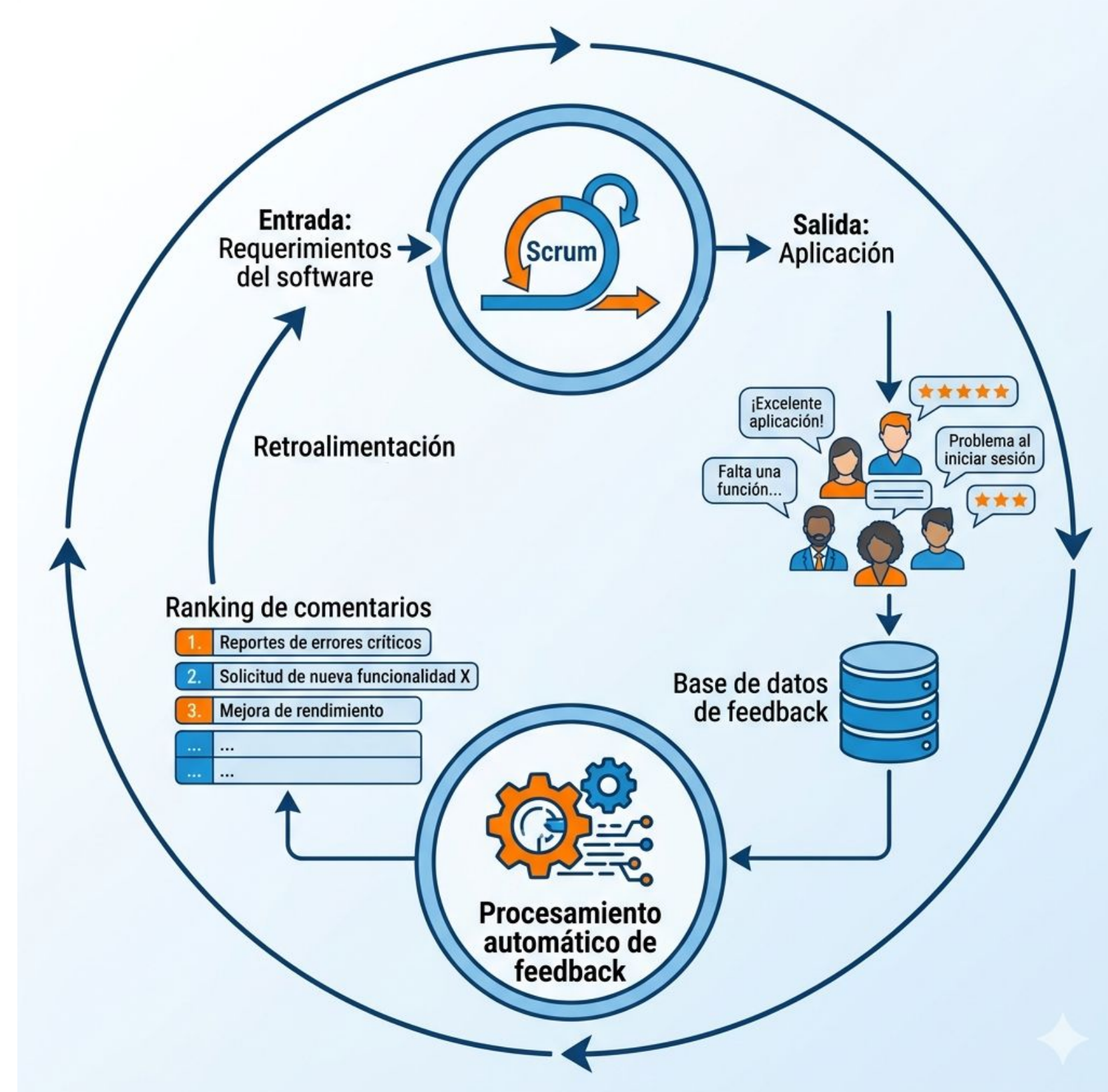
Andrés Rojas Paredes, Brenda Mareco, Christian Chamizo
Instituto de Ciencias, Universidad Nacional de General Sarmiento, Malvinas Argentinas
arojas@campus.ungs.edu.ar



Universidad
Nacional de
General
Sarmiento

Introducción

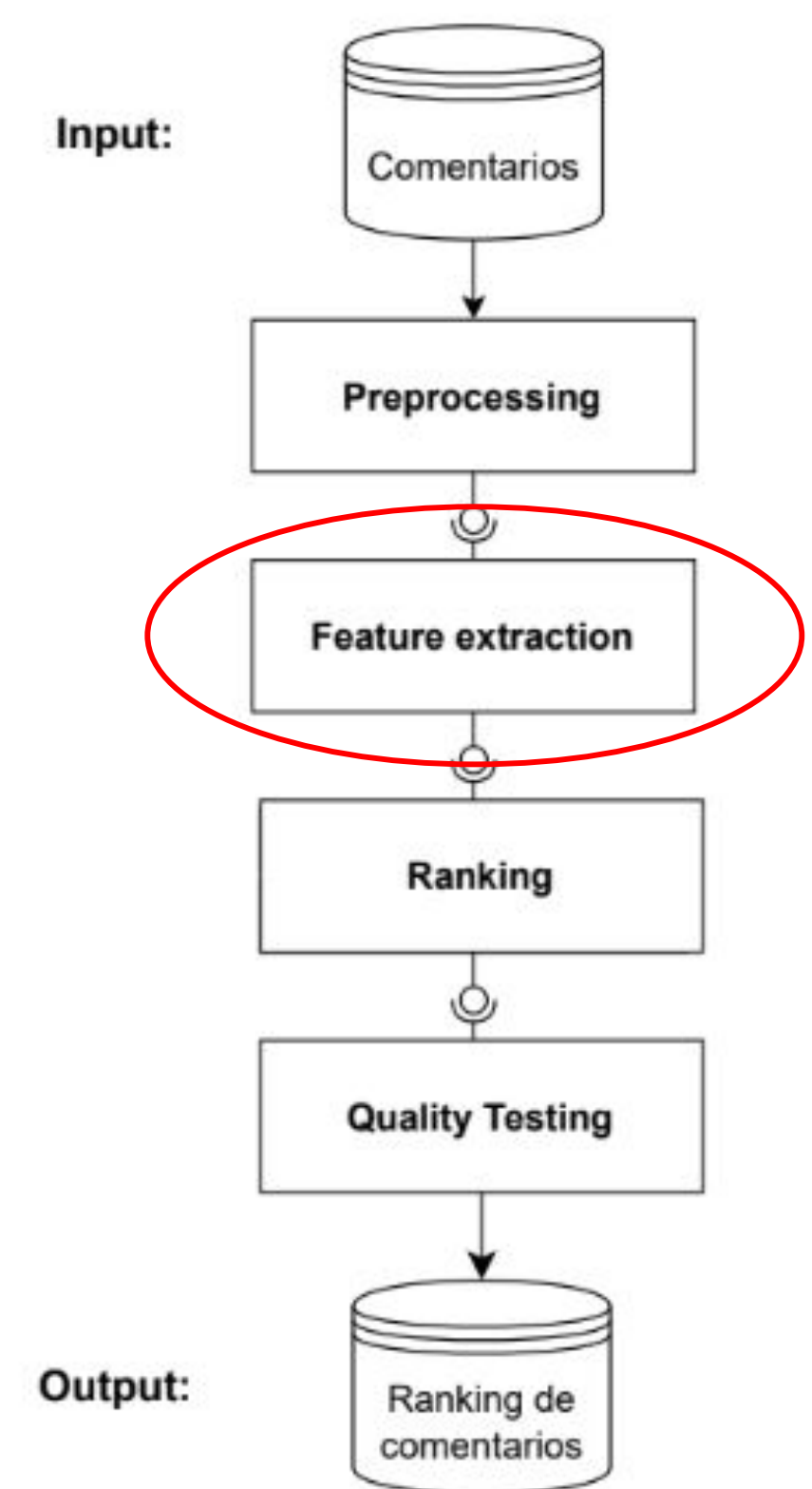
El procesamiento automático de reviews se encarga de procesar los comentarios que los usuarios hacen de una determinada aplicación. La idea es rankear los reviews de tal forma que queden al top del ranking los reviews más importantes o más relevantes para el ciclo de vida de la aplicación.



Se implementa un pipeline cuyo output es el ranking de comentarios o reviews.

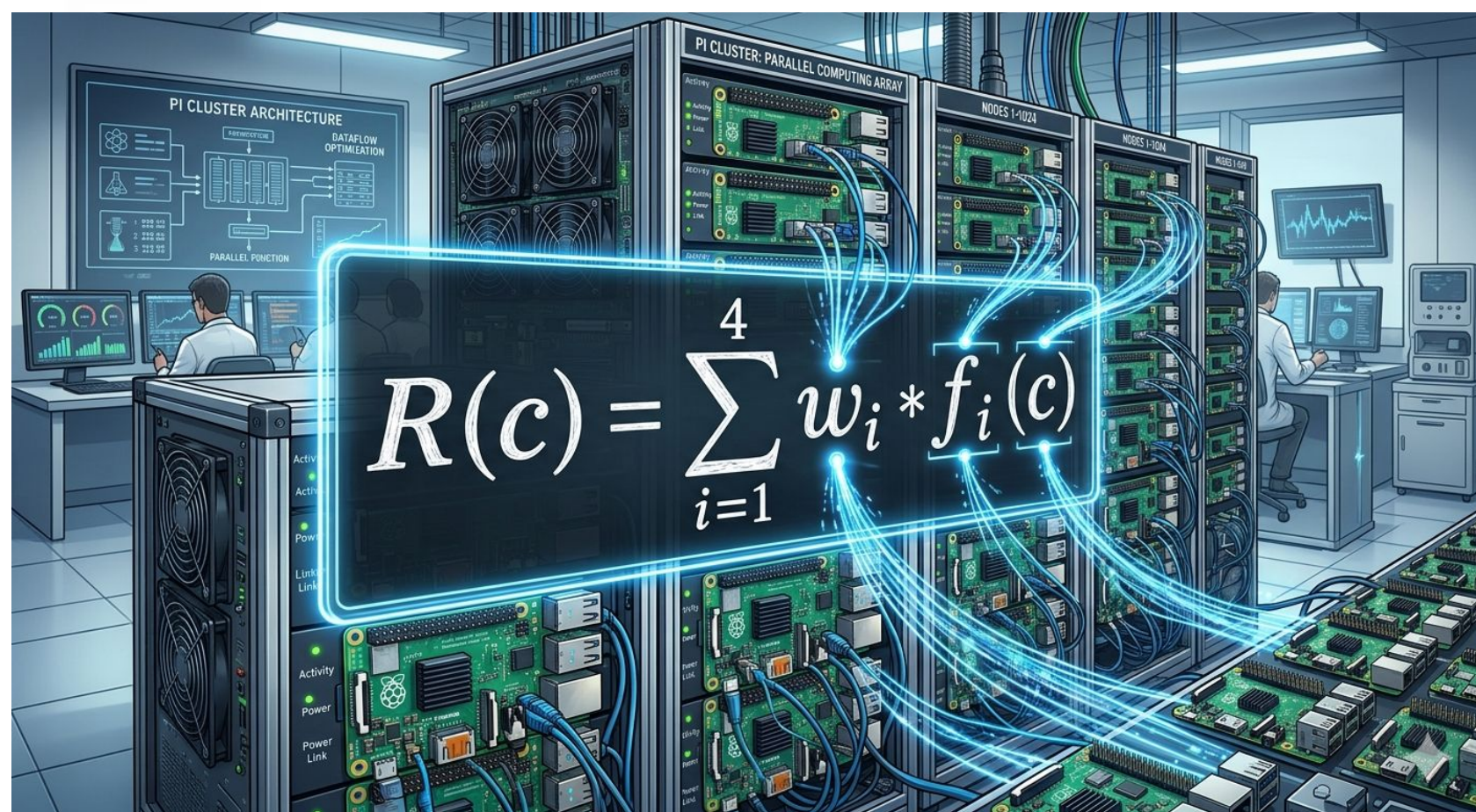
Este pipeline posee etapas donde la más importante es Feature extraction. En esta etapa se extraen los principales features que determinan un ranking de acuerdo a una función de pesos.

Finalmente se evalúa el ranking producido de acuerdo a medidas de precisión como NDCG y bias algorítmico con herramientas como AIF360.



Lineas de Investigación y Desarrollo

Línea 1: Cómputo Paralelo



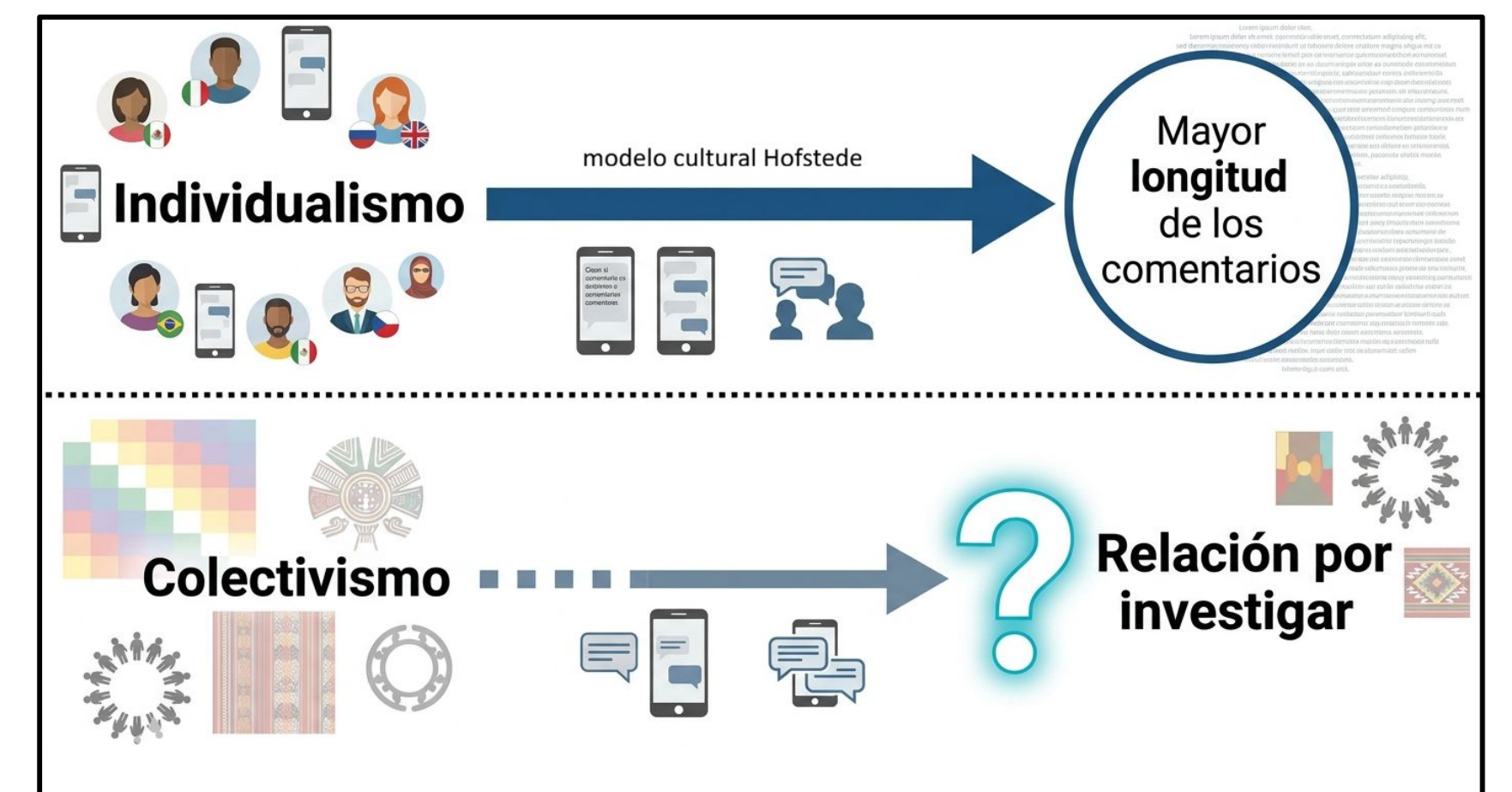
El ranking $R(c)$ de un comentario c , se calcula con una función que asigna un peso w_i a cada atributo $f_i(c)$ del comentario c . El objetivo es encontrar los mejores pesos w_i con herramientas de alto rendimiento y cómputo paralelo como un cluster Raspberry Pi.

Línea 2: Eliminación de Atributos



En trabajo previo se ha demostrado que la Entropía de Shannon se puede usar para reemplazar con éxito al atributo Revlen (longitud del comentario). El objetivo es determinar qué otros atributos se pueden eliminar utilizando la Entropía.

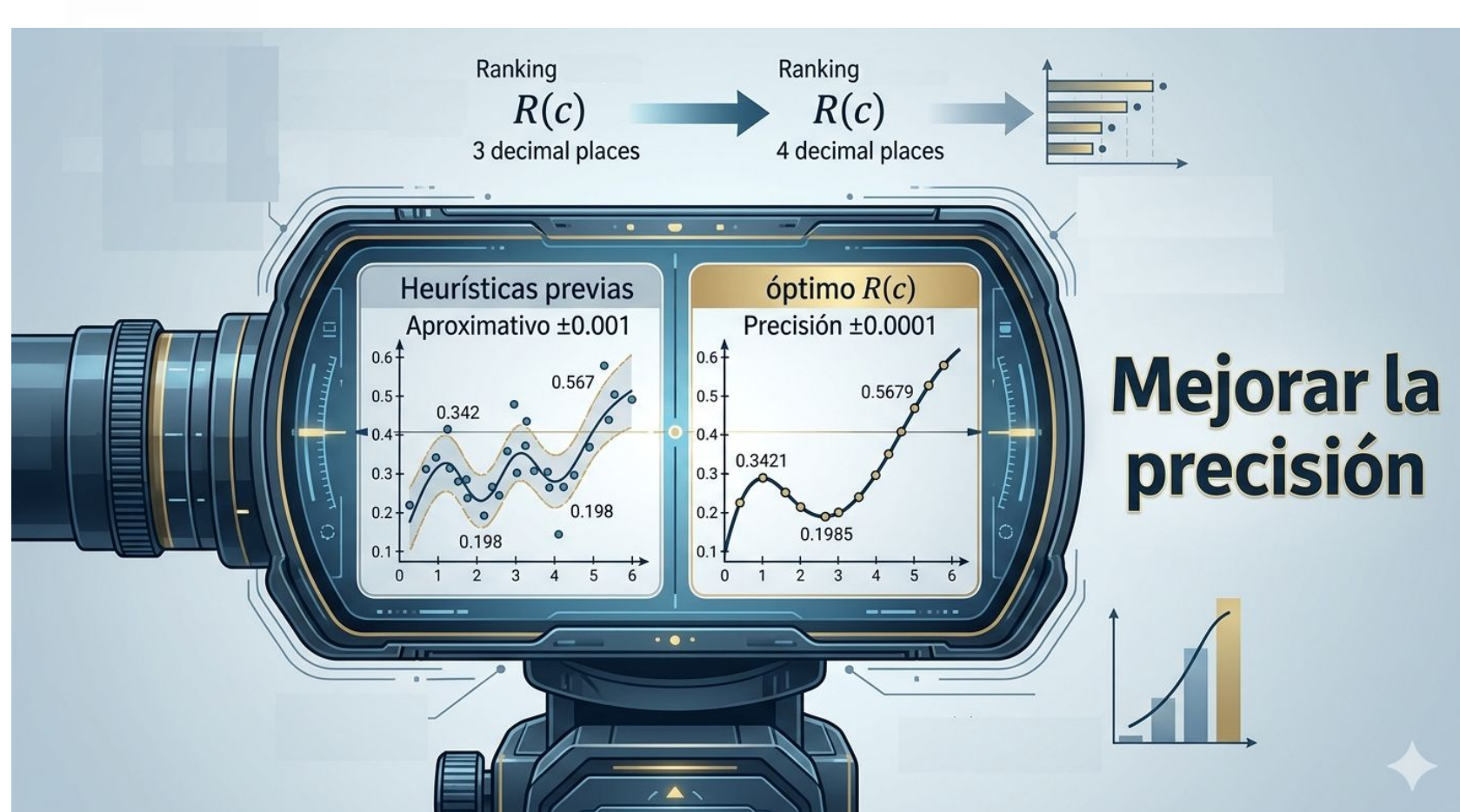
Línea 3: Idiomas Indígenas



En trabajo previo se ha demostrado que países con alto individualismo tienden a escribir comentarios de mayor longitud. El objetivo es encontrar una conexión similar entre el colectivismo de las culturas indígenas y los atributos de un comentario.

Resultados Obtenidos y Esperados

Línea 1



Se espera aumentar la precisión de tres a cuatro dígitos en los pesos de la función de ranking $R(c)$. Esta mejora permitirá la obtención de resultados óptimos en términos de la métrica NDCG contra resultados de trabajos previos hechos de manera aproximativa y con heurísticas.

Línea 2



Se espera eliminar los atributos que son más costosos computacionalmente de obtener, esto permitirá eliminar las librerías de aprendizaje automático requeridas para trabajar con atributos como sentimiento y categoría. Con esta simplificación se podrá portar el pipeline a infraestructuras heterogéneas como un cluster Raspberry Pi.

Línea 3



Se ha demostrado que países con alto individualismo tienden a escribir comentarios de mayor longitud lo que se traduce en un ranking sesgado o con bias algorítmico. El objetivo es encontrar una conexión similar entre el colectivismo de las culturas indígenas y los atributos de un comentario.

Formación de RRHH

Un miembro del equipo de la línea de investigación posee formación a nivel de doctorado, siendo especialista en Teoría de la Complejidad e Ingeniería de Requerimientos. El resto de los investigadores está compuesto por una graduada de la Licenciatura en Sistemas y por un estudiante de la Tecnicatura en Informática de la UNGS.

Bibliografía

- Guzmán E., Fischer R., Walczuch R: Does Culture Matter? Impact of Individualism and Uncertainty Avoidance on App Reviews. ICSE-SEIS (2021).
- Guzmán E., M. Ibrahim, and M. Glinz. Prioritizing user feedback from Twitter: A survey report. In Int. Workshop on Crowd Sourcing in Software Engineering, p. 21–24, (2017).
- Rojas Paredes A., Mareco B.. Shannon Entropy is better Feature than Category and Sentiment in User Feedback Processing. XXX Congreso Argentino de Ciencias de la Computación - CACIC (2024).