



CLASIFICACIÓN DE BIOENSAYOS DE PUBCHEM: DE TF-IDF/LDA A BERT, Y EL PROBLEMA DEL *DATA LEAKAGE*

Patricia A. Quispe^{1,2}, Ignacio E. León², Waldo Hasperué¹

¹Instituto de Investigación en Informática LIDI (UNLP-CIC PBA), Facultad de Informática,
UNLP, Calle 50 y 120, La Plata, Argentina.

²Centro de Química Inorgánica (CEQUINOR, UNLP-CONICET-CIC PBA), Facultad de
Ciencias Exactas, UNLP, Blvd. 120 N.º 1465, La Plata, Argentina.

Correo electrónico: quispepatricia@quimica.unlp.edu.ar

En *drug discovery*, bases de datos masivas como PubChem son una fuente fundamental de información sobre compuestos con actividad biológica. Sin embargo, la heterogeneidad en la documentación de los bioensayos dificulta determinar su tipo (bioquímico, celular, *in vivo*, etc.) y puede favorecer la propagación de errores hacia otras bases de datos alimentadas por ellas. En este contexto, los ensayos bioquímicos “cell free” sobre proteínas purificadas son de particular interés, porque permiten estudiar directamente la interacción entre fármacos y blancos moleculares asociados a procesos patológicos. Por eso, este trabajo se focalizó en su identificación y caracterización como objetivo principal.

Se ejecutó un pipeline en dos etapas. Primero, se comparó un enfoque clásico basado en TF-IDF+LDA con regresión logística frente al *fine-tuning* de PubMedBERT, utilizando pérdida ponderada por clase, F1-macro y distintas estrategias de *pooling* (CLS, mean, max y combinado). También se evaluó el enriquecimiento mediante *Bag of Ontology*, incorporando términos de la BioAssay Ontology (BAO) [1] mediante *matching* léxico. Aunque TF-IDF+LDA mostró métricas internas favorables, el análisis de aproximadamente 31.000 registros evidenció una mejor separación de los grupos sin enriquecimiento ontológico, debido a la presencia de términos compartidos entre clases en el vocabulario de la BAO.

PubMedBERT alcanzó inicialmente un F1-macro=1,0, pero la revisión reveló una fuga de información: en el 92,1% de los casos, la clase aparecía literalmente o casi literalmente en el texto (99,9% *Biochemical*, 98,8% *Cell-based*, 100% *In vivo*), asociada a las convenciones de nomenclatura de PubChem. Tras enmascarar estas menciones, el desempeño disminuyó a valores más realistas y se evaluó el modelo sobre 31.230 registros externos *in vitro*. Persistieron ambigüedades genuinas en descripciones con información experimental mixta que un clasificador de etiqueta única no logró resolver. Finalmente, el modelo se aplicó a la descarga completa de PubChem, identificándose 189.155 bioensayos sobre proteínas purificadas. Su comparación con las anotaciones *Biochemical* de ChEMBL33 permitió detectar inconsistencias de curación, incluyendo registros clasificados como proteína purificada que empleaban extractos crudos. En conjunto, el pipeline demostró utilidad para el enriquecimiento y control de calidad de datos de bioactividad a gran escala. Como perspectiva, se plantea complementar la clasificación mediante etiquetado asistido por LLM y un esquema multilabel basado en la BAO para representar adecuadamente los casos de información experimental mixta.

REFERENCIAS:

1. Vempati, U. D.; Przydzial, M. J.; Chung, C.; Abeyruwan, S.; Mir, A.; Sakurai, K.; Schürer, S. C. *PLoS One*, **2012**, 7 (11), e49198. DOI: [10.1371/journal.pone.0049198](https://doi.org/10.1371/journal.pone.0049198)