Implementation of BERT for Automatic Extraction of Named Entities and Semantic Visualization on Technologies in a Corpus of Scientific Articles

Javic Camilo Rojas Hurtado[1] ; Yeimy Vanessa Lopez Terreros[2] ; Diana M. Cardona-Román[3]
[1,2,3] University of Los Llanos, Colombia, javic.rojas@unillanos.edu.co, yeimy.lopez@unillanos.edu.co, dcardona@unillanos.edu.co

Abstract– ***Automated analysis of scientific literature represents an opportunity to discover relevant knowledge in a specific domain. The main challenge of Natural Language Processing (NLP) is the morphological descriptions of technical language and their scattered location, which hinders systematic processing. This work proposes the design and implementation of an NLP method such as bidirectional encoder representation from transformers (BERT) for the automatic analysis and visualization of information extracted from scientific articles on artificial intelligence technologies. Techniques such as Named Entity Recognition (NER) and semantic representations are used, supported by pre-trained models from the scientific domain. The work was approached from the construction of a specialized corpus, text processing, fine-tuning of the BERT and SciBERT models, and evaluation using standard metrics. In the first experiment, the best configuration obtained in fine-tuning was SciBERT_lrst1, which achieved an F1-score of 0.8262, with a learning rate of 5e-05, weight decay of 0.1, 100 warmup steps, and a linear scheduler. The second experiment maintained the same hyperparameters and the best model, BERT_lrst1, obtained an F1-score of 0.7573. It can be observed that the incorporation of warmup steps and a higher level of regularization promote training stability and improve generalization capacity, which is particularly relevant in scenarios with limited corpus size. SciBERT's specialization in scientific texts gives it an additional advantage in NER tasks in specialized domains such as the one analyzed in this study.***
Keywords-- Natural Language Processing (NLP), Named Entity Recognition (NER), Semantic Representation, Data Visualization.

Implementación del modelo de lenguaje BERT para la extracción automática de entidades nombradas y visualización semántica sobre tecnologías en un corpus de artículos científicos

Javic Camilo Rojas Hurtado[1] ; Yeimy Vanessa Lopez Terreros[2] ; Diana M. Cardona-Román[3]
[1,2,3] *Universidad de Los Llanos, Colombia,* javic.rojas@unillanos.edu.co, yeimy.lopez@*unillanos.edu.co, dcardona@unillanos.edu.co*

Resumen– ***El análisis automatizado de literatura científica representa una oportunidad para descubrir conocimiento relevante de un dominio específico. El principal desafío del Procesamiento de Lenguaje Natural (PLN) son las descripciones morfológicas del lenguaje técnico y su ubicación dispersa, lo que dificulta el procesamiento sistemático. Este trabajo propone el diseño e implementación de un método de PLN como la representación de codificadores bidireccionales a partir de transformadores (BERT) para el análisis automático y visualización de información extraída de artículos científicos sobre tecnologías de inteligencia artificial. Se emplean técnicas como Reconocimiento de Entidades Nombradas (NER) y representaciones semánticas, apoyadas en modelos preentrenados del dominio científico. El trabajo se abordó desde la construcción de un corpus especializado, el procesamiento de texto, fine-tuning de los modelos BERT y SciBERT y la evaluación mediante métricas estándar. En el primer experimento, la mejor configuración obtenida en el fine-tuning fue SciBERT_lrst1, que alcanzó un F1-score de 0,8262, con learning rate de 5e-05, weight decay de 0.1, 100 warmup steps y un scheduler lineal. El segundo experimento mantuvo los mismos hiperparámetros y el mejor modelo BERT_lrst1, obtuvo un F1-score de 0,7573. Se observa que la incorporación de warmup steps y un mayor nivel de regularización favorecen la estabilidad del entrenamiento y mejoran la capacidad de generalización, aspecto particularmente relevante en escenarios con corpus de tamaño limitado. La especialización de SciBERT en textos científicos le otorga una ventaja adicional en tareas de NER en dominios especializados como el analizado en este estudio. Palabras clave- Procesamiento de Lenguaje Natural (PLN), Reconocimiento de Entidades Nombradas (NER), Representación semántica, Visualización de Datos.***

I. Introducción

El crecimiento sostenido de la producción científica ha generado grandes volúmenes de información textual especializada cuya exploración manual resulta compleja. En consecuencia, el Procesamiento de Lenguaje Natural (PLN) ha demostrado su utilidad en la estructuración y análisis de información científica y biomédica mediante técnicas como Reconocimiento de Entidades Nombradas (NER) y extracción de conocimiento. Sin embargo, la mayoría de los desarrollos se han centrado en medicina humana, evidenciando una brecha en el dominio veterinario e inclusive en el dominio tecnológico o computacional.

En este contexto, se propone adaptar modelos de lenguaje preentrenados basados en arquitecturas Transformer al análisis de literatura sobre tecnologías de inteligencia artificial. El trabajo se desarrolla en fases progresivas: una construcción de un conjunto de datos con anotaciones y entrenamiento en el dominio tecnológico, aplicación del mejor modelo a nuevos documentos y visualización semántica de procesamiento.

Los modelos BERT y variantes especializadas como BioBERT [1], ClinicalBERT [2] y SciBERT [3] han mostrado mejoras significativas en tareas como NER y extracción de relaciones, aunque su entrenamiento requiere recursos computacionales significativos. Por ello, se plantea evaluar estrategias de ajuste fino aplicables en entornos académicos. El desarrollo se centra principalmente en literatura del dominio tecnológico en inteligencia artificial; sin embargo, el sistema se diseña con la capacidad de extenderse posteriormente a otros escenarios aplicados como la medicina veterinaria.

La pregunta resuelta en este artículo es ¿Cómo se puede realizar fine-tuning en el modelo BERT y SciBERT para el reconocimiento de entidades nombradas sobre tecnologías en un corpus de artículos científicos, evaluando su rendimiento en precisión y cobertura?

Se espera que esta propuesta contribuya a optimizar el aprovechamiento del conocimiento científico disponible, fortalecer la aplicación del PLN en contextos biomédicos especializados mediante la integración de extracción automática de información y técnicas de visualización interactiva.

II. Marco Teórico

A. Procesamiento de Lenguaje Natural (PLN)

El Procesamiento de Lenguaje Natural (PLN) constituye una de las principales líneas de la Inteligencia Artificial orientadas a la extracción automática de información a partir de texto no estructurado [4]. El PLN ha permitido estructurar literatura científica y registros clínicos mediante tareas como clasificación de texto, Reconocimiento de Entidades

Nombradas (NER) y representación de relaciones semánticas [5], [6].

Estas tareas resultan especialmente relevantes cuando se trabaja con literatura altamente especializada, donde el conocimiento se encuentra distribuido en documentos extensos y con terminología técnica compleja.

B. *Arquitectura general de un sistema de PLN*

La arquitectura general de un sistema de PLN comprende etapas de preprocesamiento, representación y modelado. Como se muestra en la Fig. 1, el flujo inicia con documentos en texto plano que son transformados mediante procesos de normalización y posteriormente representados en forma numérica para su análisis mediante modelos de aprendizaje automático.

En enfoques tradicionales, la representación textual se basa en técnicas como TF-IDF [7]. Sin embargo, modelos basados en arquitecturas Transformer han demostrado un desempeño superior al capturar dependencias contextuales complejas [8].

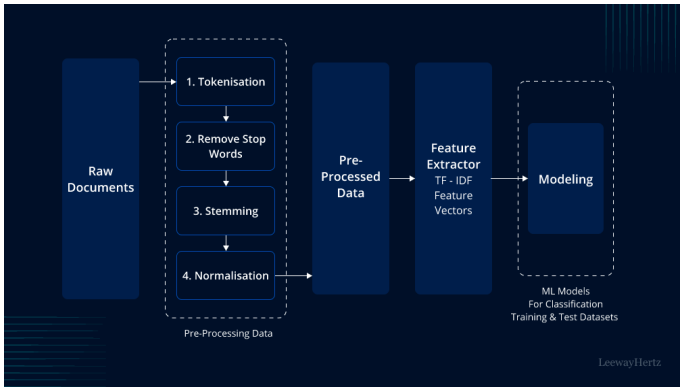

Fig. 1 Arquitectura general clásica del PLN. Fuente: Adaptado de [9].

C. *Preprocesamiento del texto*

El preprocesamiento constituye una etapa fundamental para garantizar la consistencia estructural del corpus. Como se ilustra en la Fig. 2, este proceso incluye segmentación, tokenización y normalización léxica, etapas que permiten transformar texto natural en unidades procesables por modelos computacionales [10]–[12].

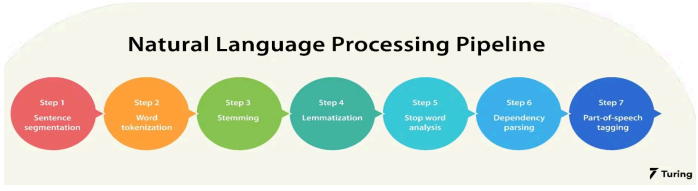

Fig. 2 Flujo general del preprocesamiento en PLN. Fuente: Adaptado de [13].

D. *Modelos preentrenados en PLN biomédico*

El principal avance reciente en PLN ha sido la introducción de modelos basados en Transformer, particularmente BERT [5], cuya arquitectura bidireccional permite capturar contexto semántico completo dentro de una oración.

En el dominio biomédico, variantes especializadas como BioBERT [1], ClinicalBERT [2] y SciBERT [3] han sido entrenadas sobre literatura científica y registros clínicos, logrando mejoras significativas en tareas como NER. Pero hasta el momento, no hay evidencia de un modelo de lenguaje entrenado de manera exclusiva en el dominio de las tecnologías.

La Fig. 3 presenta el esquema general de aplicación de un modelo BERT que aprende de un dominio específico, por ejemplo SciBERT o BioBERT para la extracción de NER, donde el texto tokenizado es procesado mediante capas de atención para generar representaciones contextuales (*embeddings*) que posteriormente son clasificadas.

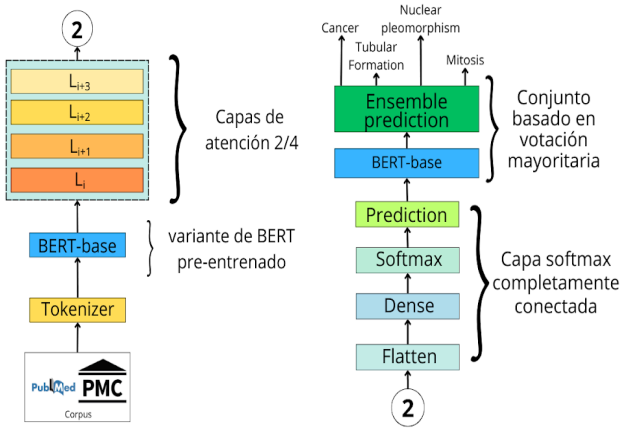

Fig. 3 Esquema simplificado de aplicación de BioBERT para NER. Fuente: Elaboración propia basada en [1].

En el presente trabajo, estos modelos constituyen la base para el desarrollo del pipeline metodológico inicialmente validado en un dominio tecnológico.

E. *Reconocimiento de Entidades Nombradas*

El NER permite identificar automáticamente menciones textuales pertenecientes a categorías predefinidas [1], [14]. Por ejemplo lugares, personas, organizaciones, fechas, entre otros. En el contexto de la tecnología estas categorías podrían ser modelos, técnicas, métricas, arquitecturas, entre otras.

F. *Representaciones semánticas y visualización*

Los modelos Transformer generan embeddings contextuales que permiten medir similitud semántica, agrupar conceptos y detectar patrones conceptuales [5], [1], [15]. Para facilitar la interpretación de estos resultados, se emplean técnicas de visualización como reducción de dimensionalidad (por ejemplo, t-SNE [16]), mapas de coocurrencia y grafos semánticos [17]. La integración de PLN con visualización interactiva fortalece la interpretación de resultados y facilita la comunicación de hallazgos a investigadores sin formación técnica compleja.

III. Metodología

A) Descripción del corpus

El corpus está conformado por 124 artículos científicos en idioma inglés enfocados en tecnologías del área de computación o de inteligencia artificial y extraídos mediante ecuaciones de búsqueda especializadas. Los documentos fueron seleccionados de bases de datos académicas indexadas como Scopus y ScienceDirect, priorizando publicaciones recientes para garantizar actualidad y relevancia científica.

Los artículos, en formato digital, son convertidos a texto plano para su procesamiento automático. Durante esta etapa se eliminan secciones no relevantes para el análisis semántico (referencias, tablas, figuras y metadatos editoriales), conservando exclusivamente el contenido técnico central. A partir de este conjunto documental, se llevó a cabo un proceso manual de anotación que permitió generar aproximadamente 2.000 anotaciones en siete categorías.

B) Diseño metodológico y fases de desarrollo

Fase 1. Preparación del corpus

Se realiza la normalización lingüística del contenido de los artículos científicos en texto plano, incluyendo segmentación de oraciones, tokenización y lematización, con el fin de garantizar consistencia estructural y calidad textual para el análisis computacional.

Fase 2. Modelado y extracción de información

Utiliza modelos de lenguaje preentrenados orientados al dominio científico (por ejemplo, SciBERT). El modelo es ajustado mediante *fine-tuning* sobre el corpus específico con exploración de hiperparámetros para la tarea de NER y los embedding de representaciones semánticas. Es importante señalar que el análisis se realiza sobre literatura del dominio tecnológico en inteligencia artificial.

Fase 3. Evaluación y ajuste del prototipo

El desempeño del sistema se evalúa utilizando métricas estándar en PLN (Precision, Recall y F1-score), combinando análisis cuantitativo con revisión cualitativa de las entidades y representaciones semánticas identificadas. Esta fase permite realizar ajustes iterativos y analizar el impacto del *fine-tuning* en el rendimiento del modelo.

Fase 4. Visualización y comunicación de resultados

Finalmente, los resultados se integran en un entorno de visualización interactiva desarrollado en Python, donde se representan entidades, similitud semántica y patrones relevantes identificados en la literatura. Este componente facilita la exploración estructurada del conocimiento extraído y su comunicación a investigadores del dominio.

C) Diseño experimental

El diseño experimental contempla la construcción y partición del corpus anotado en subconjuntos de entrenamiento y evaluación, se optó por un esquema 70/15/15, donde el 70 % destinado al entrenamiento del modelo, y el 15 % restante asignado tanto al conjunto de validación como al de prueba respectivamente. La aplicación inicial del modelo preentrenado sin ajuste, su posterior *fine-tuning* sobre el dominio específico y la comparación de desempeño mediante métricas cuantitativas.

IV. Resultados

A) *Construcción y anotación del corpus*

Para la construcción del corpus de literatura científica se llevó a cabo una etapa previa de definición ontológica, en colaboración con un experto en el dominio de estudio con formación de doctorado y con la revisión de los ingenieros de sistemas del semillero de investigación Adalab de la Universidad de los Llanos. El objetivo fue identificar las entidades clave del campo, garantizando que fueran semánticamente independientes entre sí y que no presentaran solapamiento conceptual, con el fin de minimizar ambigüedades durante el proceso de anotación. Las entidades definidas fueron: **MODEL, TECHNIQUE, METRIC, APPLICATION, DATASET, ARCHITECTURE** y **TECHNOLOGY**.

Una vez establecidas las entidades, se identificaron los términos representativos de cada una y se construyó un diccionario léxico para la anotación automática de entidades nombradas en el conjunto de datos.

Para la construcción del dataset, se consideraron los 124 artículos del corpus previamente recolectados. El volumen del corpus es acotado debido a que la mayoría de los artículos encontrados se encuentran detrás de muros de pago; por razones de accesibilidad económica, se optó exclusivamente por publicaciones de acceso abierto.

Con los artículos recopilados, se procedió a una etapa de preprocesamiento que consistió en la eliminación de secciones irrelevantes y caracteres de ruido, seguida de una segmentación del texto en fragmentos (*chunks*) de 250 tokens, unidad suficiente para el procesamiento eficiente por parte del modelo. Posteriormente, se ejecutó el proceso de anotación automática utilizando el diccionario previamente definido, obteniendo un dataset final de 1921 chunks con anotaciones.

Como se puede observar en la Tabla 1, el conjunto de datos presenta un notable desbalanceo de clases: la entidad MODEL concentra la mayor cantidad de anotaciones con 4.364, mientras que ARCHITECTURE registra el menor número con apenas 84, lo que representa una diferencia superior a 4.000 anotaciones. Este desbalanceo será analizado en detalle en la sección de resultados del modelo óptimo.

Tabla 1. Distribución total de anotaciones por entidad en el conjunto de datos completo.

Entidad	Cantidad de anotaciones
MODEL	4364
METRIC	2683
TECHNOLOGY	2306
TECHNIQUE	2090
APPLICATION	589
DATASET	227
ARCHITECTURE	84

B) *Ajuste y evaluación inicial de modelos de lenguaje*

La experimentación incluyó modelos preentrenados basados en arquitectura Transformer, específicamente BERT y SciBERT. Aunque en el marco teórico se mencionan variantes especializadas como BioBERT y ClinicalBERT, éstas se encuentran principalmente orientadas al dominio biomédico y clínico. En esta etapa del proyecto, el enfoque se centra en el análisis de literatura tecnológica y de inteligencia artificial, por lo que se prioriza el uso de SciBERT, ya que fue preentrenado sobre artículos científicos que incluyen ampliamente el dominio de computación. Esta característica permite una mejor representación semántica de términos técnicos relacionados con modelos, arquitecturas y técnicas de inteligencia artificial.

Dado que el corpus anotado disponible es de tamaño limitado, se implementó una estrategia de optimización de hiperparámetros inspirada en el principio de la validación cruzada (*cross-validation*). Este enfoque consiste en evaluar de forma iterativa distintas combinaciones de hiperparámetros con el fin de identificar la configuración que maximiza el rendimiento del modelo. En cada iteración se comparan los resultados obtenidos y se retiene la combinación con mejor desempeño, descartando las configuraciones subóptimas. Los hiperparámetros explorados en este proceso fueron: la tasa de aprendizaje (*learning rate*), la regularización mediante decaimiento de pesos (*weight decay*), el número de pasos de calentamiento (*warmup steps*) y el tipo de planificador de tasa de aprendizaje (*learning rate scheduler*).

Los resultados experimentales evidencian que el desempeño de los modelos es altamente sensible a la configuración de hiperparámetros. En el caso de SciBERT, como se muestra en la Fig. 4, la mejor configuración correspondió a *SciBERT_lrst1*, alcanzando un F1-score de 0,8262, con una tasa de aprendizaje de 5e-05, *weight decay* de 0.1, 100 *warmup steps* y un *scheduler* lineal. Se observa que la incorporación de *warmup steps* y un mayor nivel de regularización favorecen la estabilidad del entrenamiento y mejoran la capacidad de generalización, aspecto particularmente relevante en escenarios con corpus de tamaño limitado.

Model-name	learning-rate	weight_decay	warmup_steps	lr_scheduler_type	Score (F1)
Scibert_vlr1	2E-05	0.01	0	lineal	0,7066
Scibert_vlr2	1E-05	0.01	0	lineal	0,6175
Scibert_vlr3	3E-05	0.01	0	lineal	0,7648
Scibert_vlr4	5E-05	0.01	0	lineal	0,7939
Scibert_vwd1	5E-05	0.1	0	lineal	0,7944
Scibert_vwd2	5E-05	0.001	0	lineal	0,7944
Scibert_ws1	5E-05	0.1	100	lineal	0,8071
Scibert_ws2	5E-05	0.1	300	lineal	0,7434
Scibert_ws3	5E-05	0.1	500	lineal	0,6973
Scibert_lrst1	5E-05	0.1	100	lineal	0,8262
Scibert_lrst2	5E-05	0.1	100	cosine	0,8076

Fig. 4 Desempeño del modelo SciBERT bajo diferentes configuraciones de hiperparámetros en un escenario de datos limitados.

Model-name	learning-rate	weight_decay	warmup_steps	lr_scheduler_type	Score (F1)
Bert_vlr1	2E-05	0.01	0	linear	0,6186
Bert_vlr2	1E-05	0.01	0	linear	0,4244
Bert_vlr3	3E-05	0.01	0	linear	0,6827
Bert_vlr4	5E-05	0.01	0	linear	0,7386
Bert_vwd1	5E-05	0.1	0	linear	0,7479
Bert_vwd2	5E-05	0.001	0	linear	0,7201
Bert_ws1	5E-05	0.1	100	linear	0,7428
Bert_ws2	5E-05	0.1	300	linear	0,6824
Bert_ws3	5E-05	0,1	500	linear	0,5421
Bert_lrst1	5E-05	0,1	100	linear	0,7573
Bert_lrst2	5E-05	0,1	100	cosine	0,7403

Fig. 5 Desempeño del modelo BERT bajo diferentes configuraciones de hiperparámetros en un escenario de datos limitados.

Por su parte, los resultados obtenidos con BERT (ver Fig. 5) muestran un comportamiento similar en cuanto a la sensibilidad de los hiperparámetros, aunque con un desempeño general inferior al de SciBERT. La mejor configuración fue *BERT_lrst1*, con un F1-score de 0,7573, utilizando la misma combinación óptima de hiperparámetros (learning rate de 5e-05, *weight decay* de 0.1, 100 *warmup steps* y *scheduler* lineal). Esto sugiere que, aunque el ajuste de hiperparámetros mejora significativamente el rendimiento en ambos modelos, la especialización de SciBERT en textos científicos le otorga una ventaja adicional en tareas de reconocimiento de entidades en dominios especializados.

Con el fin de analizar en detalle el comportamiento del modelo bajo la configuración óptima seleccionada, la Fig. 6 presenta la evolución de la función de pérdida durante el proceso de entrenamiento y validación. Se observa una disminución progresiva y estable del *loss* a lo largo de las épocas, lo que indica una convergencia adecuada del modelo y una correcta adaptación al corpus utilizado.

De manera complementaria, la Fig. 7 muestra la evolución de las métricas de precisión y recall. Ambas presentan una tendencia ascendente hasta estabilizarse en las últimas etapas del entrenamiento, evidenciando un aprendizaje consistente de los patrones lingüísticos asociados a las entidades nombradas. La estabilización conjunta de estas métricas respalda el F1-score final alcanzado (0,8262) y sugiere un equilibrio adecuado entre la reducción de falsos positivos y falsos negativos.

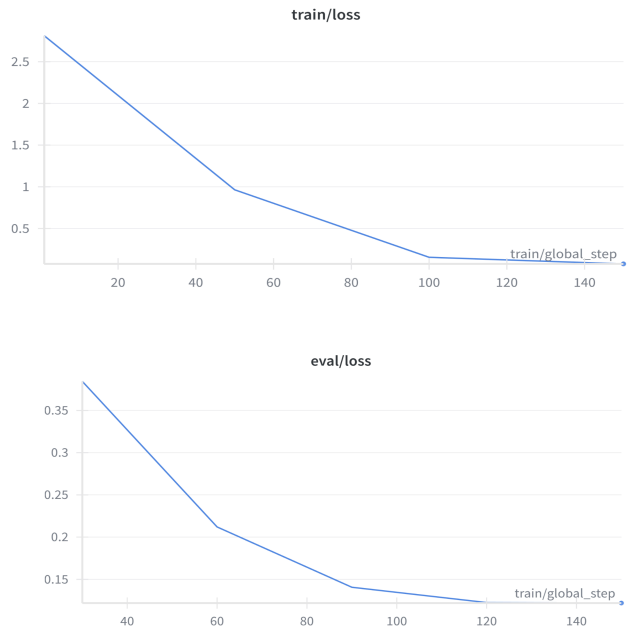

Fig. 6 Evolución de la función de pérdida en entrenamiento (a) y validación (b) durante el ajuste fino de SciBERT para la tarea de reconocimiento de entidades nombradas.

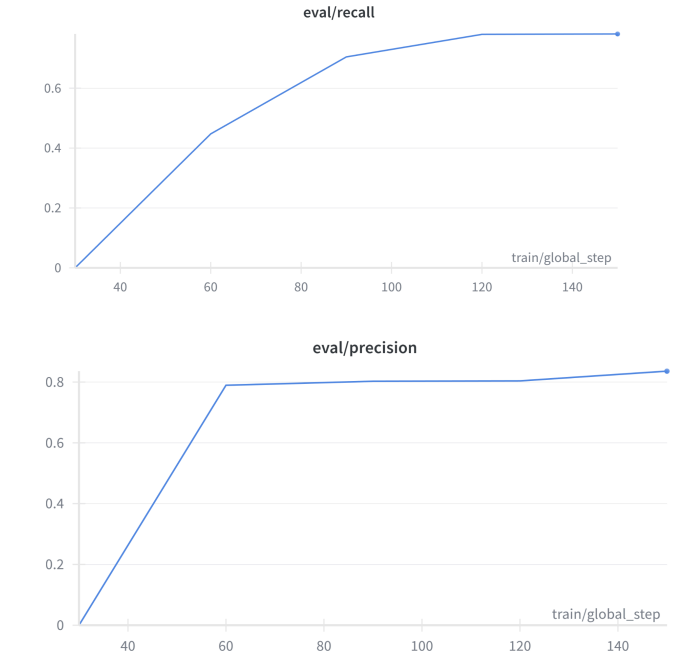

Fig. 7 Comportamiento de las métricas de evaluación (precision y recall) a lo largo de las épocas durante el ajuste fino de SciBERT.

En conjunto, estos resultados confirman que la combinación de una tasa de aprendizaje controlada, regularización mediante *weight decay* y la incorporación de *warmup steps* favorece la estabilidad del ajuste fino de SciBERT. Como se observa en la Figura 6, las curvas de pérdida de entrenamiento y validación presentan una tendencia decreciente sostenida sin señales de sobreajuste (*overfitting*), lo que indica que el modelo logró aprender patrones generalizables a partir del corpus disponible. Complementariamente, la Figura 7 muestra que las métricas de *precision* y *recall* alcanzan una convergencia estable a partir de los 80 pasos globales aproximadamente, estabilizándose en valores superiores a 0,7 y 0,8 respectivamente, lo que refleja un balance adecuado entre la capacidad del modelo para detectar entidades relevantes y la exactitud de sus predicciones.

En términos de desempeño global, los resultados obtenidos indican que SciBERT es capaz de reconocer entidades nombradas en textos científicos especializados de manera competente, aprovechando su preentrenamiento en literatura científica para capturar mejor el contexto semántico propio del dominio. Sin embargo, la confiabilidad práctica del modelo no puede asumirse de forma homogénea, puesto que el rendimiento final está condicionado en gran medida por la distribución de anotaciones en el corpus de entrenamiento: entidades con mayor representación tienden a ser aprendidas con mayor precisión, mientras que aquellas con menor volumen de ejemplos pueden presentar un desempeño más limitado. Este comportamiento diferenciado por clase se analiza en detalle a continuación.

Tabla 2. Resultados de F1-Score por entidad

Entity	F1-Score
APPLICATION	0,7609
ARCHITECTURE	0,5556
DATASET	0,6237
METRIC	0,8997
MODEL	0,8882
TECHNIQUE	0,7419
TECHNOLOGY	0,8411

En la Tabla 2 se presentan los resultados del F1-Score por entidad, donde es posible evidenciar el impacto del desbalanceo de clases en el desempeño del modelo. La entidad MODEL, que concentraba el mayor número de anotaciones en el conjunto de datos, obtuvo el segundo F1-Score más alto con 0,8882, lo que refleja la ventaja que representa disponer de mayor volumen de ejemplos de entrenamiento. En contraste, la entidad ARCHITECTURE, que contaba con la cantidad más reducida de anotaciones, registró el F1-Score más bajo de todas las entidades con 0,5556, evidenciando una diferencia de rendimiento significativa respecto a las clases mejor representadas. Estos resultados confirman que el desbalanceo de clases tuvo una influencia directa en la capacidad del modelo para generalizar sobre entidades con escasa representación en el corpus.

C) *Implementación del prototipo web de visualización*

Previo a la implementación de la visualización, se realizó un estudio comparativo de técnicas de reducción de dimensionalidad para la proyección de embeddings, incluyendo PCA (Principal Component Analysis) y t-SNE (t-distributed Stochastic Neighbor Embedding). Estas pruebas permitieron establecer una base técnica sólida para la visualización de entidades. En este contexto, PCA fue utilizado como una etapa de preprocesamiento para reducir la varianza y estabilizar los datos, mientras que t-SNE se empleó posteriormente para la proyección final en dos dimensiones, preservando la estructura local de los embeddings y permitiendo una mejor separación visual de las entidades semánticamente relacionadas.

De manera paralela al ajuste de los modelos, se desarrolló un prototipo de servidor web basado en Django, implementado en un entorno local de investigación, que permite la visualización interactiva de los resultados del procesamiento semántico. Como se observa en las Fig. 8 y 9, este prototipo corresponde a una prueba de concepto funcional que permite verificar la integración entre el modelo de reconocimiento de entidades y el módulo de visualización.

El prototipo posibilita la carga de artículos científicos en formato PDF o TXT (en inglés), la extracción automática de entidades mediante el modelo ajustado y su representación gráfica a partir de los embeddings proyectados en un espacio bidimensional utilizando PCA como etapa inicial de reducción y t-SNE como técnica de proyección final. Esta combinación permite obtener una visualización más estructurada, donde PCA reduce la complejidad del espacio vectorial y t-SNE mejora la separación de clusters semánticos en la representación final.

La visualización fue implementada con ECharts, permitiendo diferenciar categorías como TECHNIQUE (amarillo), METRIC (verde), MODEL (azul) y TECHNOLOGY (gris), además de mostrar dinámicamente los fragmentos textuales originales asociados a cada entidad seleccionada y el conteo total de entidades identificadas en cada documento.

Como se observa en la Fig. 9, esta proyección corresponde a la carga de un artículo con enfoque tecnológico y muestra una separación semántica entre categorías, donde los agrupamientos observados corresponden a entidades que comparten características semánticas similares, mientras que las entidades más alejadas presentan diferencias contextuales más marcadas.

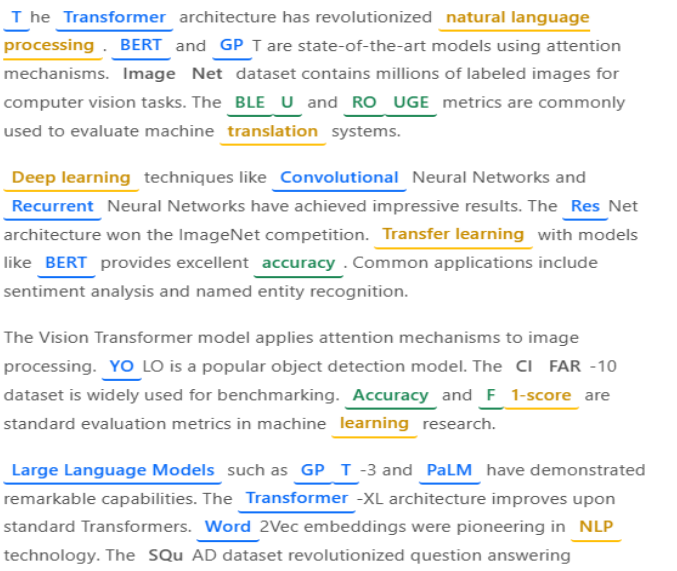

Fig. 8 Visualización semántica de entidades extraídas mediante SciBERT

No obstante, también se identifican algunas limitaciones, como la fragmentación de entidades compuestas. Por ejemplo, términos técnicos como cross validation pueden aparecer divididos en “cross” y “validation”, lo que indica que el modelo no siempre reconstruye adecuadamente expresiones multi-palabra. Este comportamiento está asociado al esquema de tokenización y al proceso de agregación de entidades, así como a la necesidad de reforzar el entrenamiento con mayor cantidad de ejemplos de términos compuestos frecuentes en literatura científica.

En la Fig. 8 se presenta la visualización semántica de las entidades extraídas mediante SciBERT, la cual permite observar de forma global la distribución de las categorías identificadas.

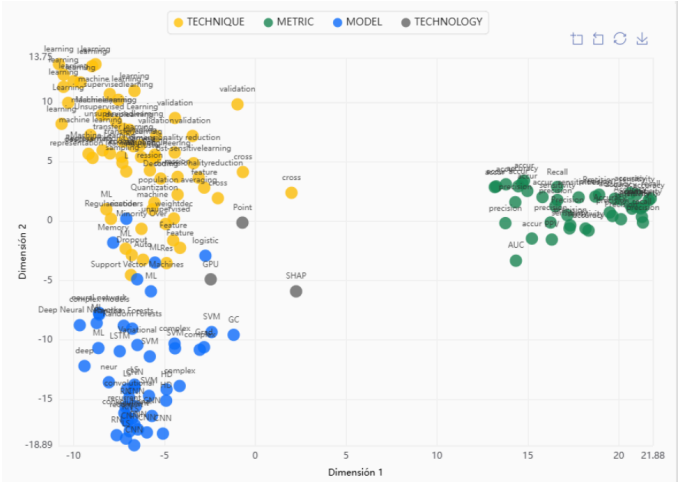

Fig. 9 Visualización de las entidades proyectados con PCA de un artículo científico del dominio tecnológico

Actualmente, el sistema permite la extracción automática de entidades conforme a las categorías previamente definidas,

así como el procesamiento de nuevos documentos cargados por el usuario. Además, integra un módulo de visualización que proyecta los embeddings asociados a las entidades detectadas, facilitando el análisis exploratorio de su organización semántica. El sistema también reporta el número total de entidades identificadas por documento, proporcionando una visión cuantitativa inicial del procesamiento realizado.

El prototipo web se encuentra funcional y ha sido validado en un entorno local. Sin embargo, aún no ha sido desplegado en infraestructura de servidor ni liberado como plataforma de acceso externo. En esta etapa, el sistema debe considerarse como una implementación experimental orientada a validación técnica y análisis metodológico.

V. Discusión

La construcción de un corpus anotado de calidad implicó un proceso manual cuidadoso, ya que las categorías definidas (MODEL, TECHNIQUE, METRIC, APPLICATION, DATASET, ARCHITECTURE y TECHNOLOGY) presentan fronteras semánticas que pueden solaparse. Por ejemplo, ciertos términos pueden ser interpretados como técnica o arquitectura dependiendo del contexto textual, lo que introduce ambigüedad durante la anotación. El tamaño del corpus puede afectar la capacidad de generalización del modelo frente a variaciones léxicas, estructuras sintácticas complejas o entidades poco frecuentes.

Desde el punto de vista computacional, el desempeño de los modelos (BERT y SciBERT) puede verse comprometido por un volumen de datos relativamente limitado. Aunque los modelos preentrenados poseen una sólida base semántica, la adaptación a un conjunto específico de entidades requiere un balance adecuado entre capacidad de generalización y especialización, evitando sobreajuste.

Asimismo, la integración entre el procesamiento semántico y la visualización interactiva representó un desafío técnico, especialmente en la proyección de embeddings a espacios bidimensionales, manteniendo coherencia entre la representación gráfica y el contexto textual original.

VI. Conclusiones

El presente trabajo propone el desarrollo de un método basado en Procesamiento de Lenguaje Natural para la extracción automática de entidades y la visualización semántica de información contenida en artículos científicos.

Se ha construido un corpus anotado de 1921 chunks etiquetado. Se han realizado pruebas de ajuste fino y análisis comparativo con modelos BERT y SciBERT para la tarea de NER, y se ha implementado un prototipo web en Django, validado en un entorno local como prueba de concepto, que permite cargar documentos, extraer entidades automáticamente y visualizar embeddings junto con los textos asociados como se visualiza en las Figuras 8 y 9.

El modelo de NER ha sido entrenado y ajustado bajo diferentes configuraciones de hiperparámetros, alcanzando resultados preliminares satisfactorios en términos de consistencia estructural y diferenciación entre categorías. La evolución estable de las métricas durante el entrenamiento sugiere una adecuada adaptación al dominio específico.

El sistema presenta limitaciones asociadas principalmente a la escala de los datos, el desbalance de clases en el corpus y la madurez del despliegue tecnológico. En particular, algunas categorías cuentan con un número reducido de ejemplos anotados, lo que puede afectar el rendimiento del modelo en entidades menos representadas. Asimismo, la presencia de datos limitados restringe la capacidad de generalización del sistema en escenarios más amplios o dominios diferentes al utilizado en este estudio.

Adicionalmente, se identificaron dificultades propias del dominio, donde ciertos términos compuestos o con caracteres especiales pueden generar ruido en la extracción de entidades, lo que evidencia la necesidad de estrategias más robustas de normalización y filtrado de texto. Por otro lado, aunque la visualización de embeddings constituye una herramienta útil para el análisis exploratorio, la interpretación de las proyecciones bidimensionales depende del método de reducción de dimensionalidad empleado, lo que puede introducir variaciones en la percepción de proximidad semántica.

El impacto potencial de esta propuesta radica en la creación de una herramienta que facilite la estructuración y comprensión de literatura científica, especialmente para investigadores junior que requieren apoyo en la interpretación de conceptos técnicos complejos.

Como trabajo futuro se propone la ampliación del corpus de datos, incorporando más artículos tecnológicos relacionados con inteligencia artificial aplicada a diagnóstico veterinario, con el fin de mejorar la robustez de los modelos y reducir posibles sesgos asociados al desbalance de clases. Se propone además reforzar la calidad de las entidades mediante estrategias de filtrado más consistentes y la incorporación de más ejemplos en categorías con baja representación. Finalmente, se plantea la validación del sistema con expertos del dominio para fortalecer la evaluación del prototipo en escenarios reales de uso.

VII. Agradecimientos

Los autores agradecen al proyecto código C01-01-2025-003 de la DGI / Unillanos, así como al semillero Automatic Data-Driven Analytics Laboratory (Adalab) y al Laboratorio de Software e Infraestructura Computacional Especializada de la Universidad de los Llanos por proveer los recursos de infraestructura tecnológica para el entrenamiento de los modelos.

Referencias

[1] J. Lee et al., “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

[2] E. Alsentzer et al., “Publicly Available Clinical BERT Embeddings,” 2019.
[3] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text,” EMNLP, 2019.
[4] J. Jiang et al., “A Survey on Natural Language Processing: Recent Advances and Future Directions,” Artificial Intelligence Review, 2023.
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of NAACL-HLT 2019, pp. 4171–4186, 2019.
[6] I. Tenney, D. Das, and E. Pavlick, “BERT Rediscovers the Classical NLP Pipeline,” ACL, 2019.
[7] K. Ayesha, D. Gupta, and A. Arora, “Text mining techniques using natural language processing,” Journal of Physics: Conference Series, 2022.
[8] A. Vaswani et al., “Attention is All You Need,” NeurIPS, 2017.
[9] A. Takyar and A. Takyar, “Natural Language Processing: A Comprehensive Overview,” 2023.
[10] S. Singh, “Natural Language Processing for Information Extraction,” arXiv:1807.02383, 2018.
[11] J. C. Obando, J. A. Pulido, and J. A. Gómez, “Procesamiento del Lenguaje Natural…,” Revista Ciencia y Tecnología, 2020.
[12] R. Torre Morenos, “Análisis de fractalidad en textos…,” 2021.
[13] Natural Language Processing Functionality in AI, 2022.
[14] S. Zhao et al., “A Comprehensive Survey of Entity Recognition and Normalization…,” Briefings in Bioinformatics, 2021.
[15] G. Gonzalez-Hernandez et al., “Capturing the patient’s perspective…,” Journal of the American Medical Informatics Association (JAMIA), 2021.
[16] L. Van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” Journal of Machine Learning Research (JMLR), 2008.
[17] J. Heer and B. Shneiderman, “Interactive Dynamics for Visual Analysis,” Communications of the ACM, vol. 55, no. 4, pp. 45–54, 2012.
[18] K. B. Cohen and L. Hunter, “Getting started in text mining,” PLoS Computational Biology, vol. 4, no. 1, e20, 2008.
[19] P. D. Stetson, S. B. Johnson, M. Scotch, and G. Hripcsak, “Evaluation of a Natural Language Processing System for Identifying and Encoding Findings in Chest Radiology Reports,” AMIA Annual Symposium Proceedings, pp. 712–716, 2002.
[20] F. Zhu et al., “Biomedical text mining and its applications in cancer research,” Journal of Biomedical Informatics, vol. 46, no. 2, pp. 200–211, 2013.
[21] National Library of Medicine, “Medical Subject Headings (MeSH),” 2021.
[22] Q. Xie et al., “Medical foundation large language models for comprehensive text analysis and beyond,” npj Digital Medicine, vol. 8, 141, 2025.
[23] S. M. Morales Segovia and J. P. Lozada Ortiz, “Neoplasias de las glándulas mamarias en caninos: prevalencia y factores de riesgo en Latinoamérica,” UCiencia, vol. 1, 2025.