# Etiquetado automático de la densidad semántica en el discurso académico de ingeniería en español mediante modelos Transformer

## *Automatic Labelling of Semantic Density in Spanish Engineering Academic Discourse Using Transformer Models*

**Resumen:** Este trabajo presenta un estudio experimental sobre la automatización del etiquetado de la densidad semántica (Semantic Density, SD) en el marco de la Legitimation Code Theory (LCT), aplicado al discurso académico de ingeniería en español. Se propone un sistema de clasificación supervisado basado en modelos Transformer (BETO, DistilBETO, RoBERTa, mBERT) y en IBETO, una versión adaptada de BETO entrenada con textos técnicos en español y ajustada mediante fine-tuning a un corpus de 1.569 oraciones anotadas en cuatro niveles de SD. Los resultados evidencian la viabilidad del etiquetado automático, con mBERT obteniendo el mayor F1 Macro (0,72) y valores de precisión superiores al 70 %, mientras que IBETO mostró un rendimiento comparable y estable tras su adaptación al dominio técnico. Las principales confusiones se observan entre categorías adyacentes, reflejando la naturaleza gradual de la densidad semántica. El estudio destaca el potencial de los modelos de lenguaje para analizar la complejidad conceptual del discurso técnico, subrayando la necesidad de fortalecer recursos lingüísticos en español para reducir la dependencia de corpus anglófonos.
**Palabras clave:** procesamiento del lenguaje natural, modelos Transformer, clasificación automática, español técnico.

**Abstract:** This study presents an experimental evaluation of the automation of semantic density (SD) tagging within the framework of Legitimation Code Theory (LCT), applied to academic discourse in engineering written in Spanish. A supervised classification system is proposed based on Transformer models (BETO, DistilBETO, RoBERTa, and mBERT), together with IBETO, an adapted version of BETO trained on technical Spanish texts and fine-tuned on a corpus of 1,569 sentences manually annotated across four levels of semantic density (SD). The results demonstrate the feasibility of automatic tagging, with mBERT achieving the highest F1 Macro (0.72) and precision values above 70 %, while IBETO showed comparable and stable performance following its adaptation to the technical domain. Most misclassifications occurred between adjacent categories, reflecting the gradual nature of semantic density. The study highlights the potential of deep language models for analyzing conceptual complexity in technical discourse and underscores the importance of developing Spanish-specific linguistic resources to reduce dependence on English-centered corpora and models.
**Keywords:** natural language processing, Transformer models, automatic classification, technical Spanish.

### *1 Introducción*

En el ámbito educativo, identificar el grado de complejidad del discurso docente resulta esencial para comprender cómo se construye y transmite el conocimiento disciplinar. La Legitimation Code Theory (LCT), desarrollada por Karl Maton (2014), ofrece un marco teórico para analizar cómo el conocimiento se estructura y comunica a través del lenguaje. Dentro de su dimensión semántica, introduce

los conceptos de densidad semántica (Semantic Density, SD) y gravedad semántica (Semantic Gravity, SG), que reflejan respectivamente el grado de condensación conceptual y la dependencia contextual del discurso. Estas categorías resultan especialmente útiles en la enseñanza de ingeniería, donde el lenguaje oscila entre la abstracción teórica y la aplicación práctica.

Tradicionalmente, el etiquetado de la densidad semántica se ha realizado de forma manual mediante protocolos sistemáticos, pero sigue siendo un proceso laborioso y de difícil escalabilidad. Este trabajo evalúa la viabilidad de automatizar el etiquetado de SD en el discurso académico de ingeniería en español mediante modelos de aprendizaje profundo basados en la arquitectura Transformer, con el fin de mejorar la coherencia y eficiencia del proceso y posibilitar análisis de mayor alcance.

Desde el punto de vista computacional, el etiquetado automático se formula en este estudio como una tarea de clasificación supervisada multiclase. Para ello, se elaboró un corpus de 1.569 oraciones procedentes de clases universitarias de ingeniería en Telecomunicación, etiquetadas manualmente en cuatro niveles de densidad semántica (SD++, SD+, SD– y SD--). A partir de este corpus, se entrenaron y evaluaron varios modelos Transformer preentrenados en español (BETO, DistilBETO, RoBERTa y mBERT) junto con IBETO, una versión adaptada específicamente para este proyecto a partir de BETO, reentrenada con textos técnicos y científicos en español. Se aplicaron técnicas de ajuste fino (fine-tuning) y regularización para optimizar el rendimiento de los modelos. Los resultados muestran que IBETO, al incorporar conocimiento lingüístico especializado, ofrece el mejor desempeño (F1 Macro $= 0{,}72$; precisión $> 70\,\%$), lo que fortalece la relevancia de disponer de modelos ajustados al dominio y al registro técnico del español.

En las secciones siguientes se presentan los fundamentos teóricos, la metodología aplicada y los resultados del estudio, así como las implicaciones para el análisis automatizado del discurso técnico y educativo.

## *2 Marco teórico y antecedentes*

La Legitimation Code Theory (LCT) (Maton, 2014; Maton, Hood & Shay, 2016) es una teoría sociológica del conocimiento que describe los principios que estructuran la producción y transmisión del saber en distintos campos disciplinares. Su dimensión semántica distingue entre densidad semántica (semantic density SD), entendida como el grado de condensación conceptual del discurso, y gravedad semántica (semantic gravity SG), vinculada a la dependencia del contexto. En los entornos educativos, la alternancia entre alta y baja SD y SG se asocia con la capacidad docente para traducir conceptos abstractos en ejemplos concretos y viceversa.

En los últimos años, la densidad semántica (SD) se ha consolidado como una herramienta analítica para el estudio del discurso académico y pedagógico en diversas áreas del conocimiento, desde la formación docente hasta la enseñanza STEM (Argüelles-Álvarez & Morton, 2023; Blackie, Adendorff & Mouton, 2023; Maton & Doran, 2017; Maton, Martin & Doran, 2021) aunque mayoritariamente (si no completamente) mediante procedimientos manuales. Hasta donde alcanza nuestro conocimiento, no existen corpus anotados de densidad semántica disponibles públicamente. Los estudios previos en LCT se han basado en análisis manuales aplicados a datos específicos de cada investigación, sin generar datasets reutilizables para tareas computacionales. Aunque contamos con anotaciones previas en otros contextos educativos, estas no fueron diseñadas como corpus estructurados para aprendizaje automático.

El auge del procesamiento del lenguaje natural (PLN) y de las arquitecturas Transformer (Vaswani et al., 2017) ha abierto nuevas posibilidades para automatizar este tipo de análisis, empleando modelos como BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) o ALBERT (Lan et al., 2020). Estos modelos aprenden representaciones contextuales del lenguaje y han demostrado una elevada eficacia en tareas de clasificación y análisis semántico.

En el ámbito hispanohablante, el desarrollo de modelos como BETO (Cañete et al., 2023) ha marcado un hito en la adaptación de las arquitecturas Transformer al español. Siguiendo esta línea, diversos trabajos recientes han explorado la extensión y especialización de modelos en dominios técnicos o científicos (Beltagy, Lo & Cohan, 2019; Gutiérrez-Fandiño et al., 2022), lo que demuestra el potencial de los enfoques de preentrenamiento continuo y fine-tuning en contextos específi-

cos. Tal como muestran de la Rosa et al. (2022) en el preentrenamiento eficiente de un modelo español, el muestreo de calidad de los datos puede reducir considerablemente los requisitos computacionales sin comprometer el rendimiento.

El desarrollo de modelos en español ha sido impulsado recientemente por iniciativas como MarIA (Gutiérrez-Fandiño et al., 2022), que amplían la cobertura del lenguaje técnico y científico.

IBETO se obtuvo mediante una fase de adaptación al dominio (continual pretraining) a partir de BETO (dccuchile/bert-base-spanish-wwm-cased), utilizando Masked Language Modeling (MLM) sobre un corpus compuesto por textos científicos en español original representativos del discurso técnico universitario y textos técnicos en español (libros y manuales académicos de matemáticas, programación, redes y servicios). Esta fase se entrenó durante 3 épocas (masking 15 %, max length 512, batch size 4, weight decay 0.01). Posteriormente, se realizó fine-tuning supervisado para la clasificación de densidad semántica sobre el corpus anotado (1.569 oraciones), aplicando validación cruzada (K=3) y pesos de clase para mitigar el desbalanceo.

Aunque existen estudios sobre clasificación del nivel académico del discurso o complejidad léxica, estos se han desarrollado en marcos y con objetivos distintos. En el ámbito del español académico, se han propuesto medidas de sofisticación y diversidad léxica para caracterizar el perfil colocacional de textos especializados (Guzzi & Alonso Ramos, 2023). Desde una perspectiva computacional y multilingüe, herramientas como MultiAzterTest permiten evaluar automáticamente distintos niveles de complejidad lingüística para la estimación de la legibilidad (Bengoetxea & Gonzalez-Dios, 2021). Asimismo, en el contexto de la adquisición de segundas lenguas, se han desarrollado modelos para alinear la complejidad lingüística con niveles de dificultad basados en el CEFR (Zhang & Lu, 2025). No obstante, la automatización del etiquetado de densidad semántica en el marco específico de la LCT constituye una línea de investigación aún no explorada.

## *3 Metodología*

El objetivo metodológico del estudio es entrenar y evaluar distintos modelos Transformer en español para clasificar automáticamente oraciones de acuerdo con su nivel de densidad semántica. Para ello, se elaboró un corpus etiquetado manualmente, se diseñó un flujo de procesamiento de texto y se realizaron experimentos comparativos con varios modelos preentrenados.

### 3.1 Corpus y etiquetado manual

El corpus está compuesto por 1.569 oraciones extraídas de grabaciones de clases de ingeniería en el ámbito de la telecomunicación impartidas en español. Se grabaron bloques de contenido completos (diez horas de grabación de media por asignatura) de tres asignaturas del programa común de los grados en Ingeniería de Telecomunicación en las áreas de electrónica (sistemas lineales y las funciones y transformadas relacionadas, teoría de circuitos eléctricos, circuitos electrónicos), programación (uso y programación de los ordenadores, sistemas operativos, bases de datos y programas informáticos con aplicación en ingeniería) y física y matemáticas aplicadas (conceptos básicos sobre las leyes generales de la mecánica, termodinámica, campos y ondas y electromagnetismo). El texto transcrito se dividió en oraciones, se segmentó y normalizó (minúsculas, eliminación de puntuación y caracteres no alfabéticos). Posteriormente, se aplicó un etiquetado manual basado en una rúbrica para el etiquetado, denominada "Translation Device" en LCT (Maton & Chen, 2016). El proceso de elaboración del TD es a partir de los datos, se definen los niveles de SD y se proporciona un ejemplo real extraído del corpus (Tabla 1).

El proceso de etiquetado fue realizado por tres anotadores entrenados y validados mediante pruebas de concordancia entre anotadores (índice $\kappa = 0{,}81$). Las discrepancias se resolvieron por consenso. Como resultado del etiquetado manual, se aprecia un claro desbalance en la distribución de las clases del corpus (SD++ =169; SD+ =599; SD- =465; SD- - =336), cuyas implicaciones se abordan con mayor detalle en el marco tecnológico. Aunque esta asimetría resulta coherente y esperable desde un punto de vista lingüístico, constituye un desafío para el modelo, pues la escasez de ejemplos en las categorías menos frecuentes puede afectar negativamente su rendimiento y estabilidad durante el entrenamiento.

Una vez finalizado el análisis, se decidió

| Etiqueta | Descripción | Ejemplo |
|---|---|---|
| SD++ | Enunciados con alta condensación conceptual, donde se articulan principios, relaciones o categorías propias del campo disciplinar. El significado se concentra en pocos elementos lingüísticos que remiten a conceptos teóricos amplios o complejos. | La función exponencial y logarítmica son unas funciones que se llaman inversas. |
| SD+ | Enunciados que movilizan conceptos disciplinares pero en un contexto de explicación o aplicación práctica. El significado técnico está presente, aunque acompañado de ejemplos. | Esto tendrá una capacidad equivalente, todos los condensadores que he pintado aquí valen lo mismo, valen cuatro microfaradios, ¿vale? |
| SD- | Enunciados que describen acciones, procesos o elementos concretos sin desarrollar relaciones conceptuales. | Vamos a hacer, vamos a ir revisando todas las características que hemos dicho que nos interesan de la programación orientada a objetos. |
| SD- - | Enunciados sin contenido conceptual o técnico. Su función es interactiva, organizativa o expresiva dentro del discurso docente. No contribuyen directamente a la construcción del conocimiento disciplinar. | Pues venga, intentad pensar, os doy ventaja. |

Tabla 1: Translation device para el etiquetado manual de la SD del corpus.

abordar el desbalanceo mediante la aplicación de pesos de clase inversamente proporcionales a la frecuencia de cada etiqueta durante el entrenamiento, lo que permite compensar la disparidad sin modificar los datos originales. En conjunto, aunque el corpus es limitado, se considera adecuado como primer acercamiento experimental, capaz de ofrecer una base empírica sólida para evaluar la viabilidad del etiquetado automático de la densidad semántica en español y orientar investigaciones futuras con conjuntos ampliados y balanceados.

Más allá de su uso experimental, este corpus constituye una herramienta de análisis del discurso pedagógico en ingeniería. Su diseño permite examinar cómo los docentes alternan entre registros más y menos complejos, configurando representaciones semánticas que reflejan la naturaleza híbrida del conocimiento técnico. En este sentido, el proyecto contribuye a vincular la lingüística computacional con la investigación educativa, al ofrecer un recurso que puede emplearse tanto para el entrenamiento de modelos como para el estudio de la comunicación disciplinar.

### 3.2 Preprocesamiento de los datos

El preprocesamiento de los datos incluyó la tokenización mediante el tokenizer específico de cada modelo (Bird et al., 2009; Manning & Schütze, 1999; Jurafsky & Martin, 2023), el truncado o padding de las secuencias según la longitud máxima permitida por cada arquitectura, la conversión de las etiquetas a formato numérico (0 a 3) y la división del corpus en conjuntos de entrenamiento (80 %), validación (10 %) y prueba (10 %). Esta segmentación del conjunto de datos responde a una práctica estándar en el entrenamiento y evaluación de modelos de machine learning, particularmente en arquitecturas basadas en transformers como BERT o GPT [1]. Todas las pruebas se realizaron en Python 3.10 uti-

[1] Aunque las proporciones específicas (como 80/10/10) no siempre son discutidas en artículos teóricos, sí hay trabajos que exploran la importancia de separar conjuntos para entrenamiento y validación en modelos complejos (Bai et al. 2021). En apuntes de prácticas de NLP del Departamento de Lingüística Computacional de Cambridge (https://www.cl.cam.ac.uk/teaching/1819/NLP/) se describe explícitamente el uso de un split 80 % entrenamiento, 10 % validación y 10 % prueba como forma estándar de configurar un experimento para comparar modelos y ajustar parámetros antes de reportar resultados finales.

lizando las librerías Transformers (HuggingFace) y PyTorch, y se ejecutaron sobre una GPU NVIDIA Tesla T4 en Google Colab, que ofreció los recursos computacionales necesarios para el ajuste fino (fine-tuning) de los modelos.

### 3.3 Modelos y configuración de entrenamiento

En este estudio se evaluaron distintos modelos basados en la arquitectura Transformer (Vaswani et al., 2017; Wolf et al., 2020; Devlin et al., 2019), con variaciones en su corpus y alcance lingüístico. La selección de los modelos Transformer se realizó con el objetivo de comparar diferentes enfoques dentro del ecosistema de modelos preentrenados en español.

En primer lugar, se incluye BETO como modelo monolingüe entrenado exclusivamente en español, ampliamente utilizado en tareas de PLN. BETO está entrenado desde cero en español, a partir de textos de Wikipedia y OpenSubtitles; aunque no especializado en lenguaje científico, muestra una buena capacidad para identificar tecnicismos. Se evaluó mBERT como modelo multilingüe, con el fin de analizar el impacto de un preentrenamiento no específico del idioma ni del dominio técnico. Esta versión multilingüe está entrenada en 104 idiomas, lo que le otorga versatilidad, pero un rendimiento moderado en dominios técnicos y específicos del español. Se incorporaron variantes como RoBERTa y DistilBETO para contrastar arquitecturas optimizadas y versiones comprimidas orientadas a eficiencia computacional. RoBERTa en español implementa una optimización de BERT al eliminar el objetivo de predicción de la siguiente frase y ampliar el corpus de entrenamiento, pudiendo superar a BETO en determinados contextos. DistilBETO es una versión comprimida del modelo, que reduce tamaño y tiempo de inferencia con una leve pérdida de precisión, adecuada para entornos con recursos limitados. Respecto a XLNet, aunque supera a BERT en algunas tareas generales, no ha sido entrenado específicamente en español ni con textos técnicos, por lo que su desempeño en este dominio podría ser más limitado sin una adaptación adicional. Finalmente, se consideró IBETO como modelo adaptado específicamente al dominio técnico mediante preentrenamiento adicional. Esta selección permite analizar no solo el rendimiento absoluto, sino también el efecto del idioma, la especialización del dominio y la complejidad del modelo en la tarea de clasificación de densidad semántica.

Cada modelo se ajustó (fine-tuned) durante un máximo de 15 épocas con las siguientes configuraciones:

- Batch size: 32
- Learning rate: 2e-5
- Optimizador AdamW
- Early stopping tras 3 épocas sin mejora de pérdida de validación
- Métrica principal: F1 Macro

### 3.4 Evaluación y métricas

Se emplearon tres métricas estándar: Precisión global (accuracy) que se refiere al porcentaje de oraciones clasificadas correctamente; F1 Macro o media armónica de precisión y exhaustividad por clase, ponderada igualmente. Matriz de confusión o análisis cualitativo de errores entre categorías contiguas. La elección de K=3 responde al equilibrio entre estabilidad estadística y tamaño efectivo del conjunto de entrenamiento. Dado que el corpus consta de 1.569 instancias y presenta desbalance entre clases, aumentar el número de folds habría reducido el número de muestras disponibles para entrenamiento en cada iteración, lo que podría afectar la capacidad de generalización del modelo en un escenario de bajo recurso. Asimismo, al tratarse de modelos Transformer de gran capacidad, se optó por una configuración que garantizara viabilidad experimental y consistencia en los resultados sin comprometer el tamaño del conjunto de entrenamiento.

El pipeline completo, desde la recopilación de datos hasta la evaluación, fue diseñado para facilitar su replicabilidad (Reimers & Gurevych, 2019; Wolf et al., 2020). Se documentaron todas las configuraciones y se creó un repositorio interno[2] con el código, los pesos del modelo y las instrucciones de uso. Esto permitirá futuras extensiones del sistema a otros dominios y su integración en herramientas de análisis de discurso educativo.

[2]Este estudio forma parte de un proyecto de tesis doctoral actualmente en desarrollo. Por este motivo, el corpus completo anotado no se encuentra disponible públicamente en esta fase. No obstante, tenemos la intención de hacer públicos los datos y recursos asociados una vez finalizado el proyecto de investigación.

## *4 Resultados y discusión*

Los resultados obtenidos en los experimentos apuntan a la viabilidad de automatizar el etiquetado de densidad semántica en español mediante modelos Transformer. Todos los modelos superaron el umbral del 60 % de precisión global, con diferencias relevantes entre los modelos generalistas y los adaptados al dominio técnico.

### 4.1 Rendimiento comparativo de los modelos

Antes de analizar los resultados de entrenamiento, se definen las principales métricas de evaluación utilizadas. El Train Loss y el Validation Loss reflejan el error promedio del modelo sobre los conjuntos de entrenamiento y validación, respectivamente; su evolución permite detectar sobreajuste o aprendizaje insuficiente. La Test Accuracy mide la proporción de aciertos sobre los datos de prueba y sirve como indicador general de rendimiento, aunque en contextos desbalanceados puede ofrecer una visión parcial. Por ello, se emplean también métricas basadas en precisión y exhaustividad: el F1 Macro otorga igual peso a todas ellas y el F1 Weighted pondera según la frecuencia de cada categoría. Estas dos últimas métricas permiten evaluar de forma más equilibrada el desempeño de los modelos en un corpus con distribución desigual de etiquetas.

En conjunto, estas medidas ofrecen una evaluación complementaria que combina indicadores clásicos de ajuste y generalización con métricas robustas ante el desbalanceo de clases, adecuadas para la tarea de etiquetado abordada.

### 4.2 Análisis comparativo del entrenamiento de los modelos

Las métricas de entrenamiento permiten evaluar la capacidad de aprendizaje, estabilidad y generalización de los modelos. Las Figuras 3 a 8, incluidas en el Anexo 1, muestran la evolución del Train Loss, Validation Loss, Accuracy y F1 para cada arquitectura (BETO, mBERT, RoBERTa, DistilBETO, XLNet e IBETO), complementando el análisis cuantitativo.

Los modelos BETO, mBERT e IBETO destacan por su rendimiento y estabilidad. mBERT logra la mejor capacidad de ajuste (F1 Macro = 0.72; precisión ≈ 77 %), mientras que IBETO (entrenado desde cero en español) muestra un comportamiento estable y una coherencia interclase consistente con su adaptación al dominio técnico (F1 Macro = 0.72; precisión ≈ 73 %). BETO mantiene un equilibrio notable entre aprendizaje y generalización. En contraste, RoBERTa y DistilBETO presentan fluctuaciones y sobreajuste, y XLNet, aunque consistente, alcanza precisiones inferiores (≈ 51 %).

En conjunto, los resultados sugieren que los modelos basados en BERT, especialmente IBETO, mBERT y BETO, son los más adecuados para el etiquetado automático de densidad semántica en español técnico, evidenciando la importancia de adaptar los modelos al dominio y a la lengua de trabajo.

La Figura 1, muestra la evolución del valor F1 Macro a lo largo de las épocas de entrenamiento, mientras que la Figura 2 presenta la evolución de la precisión (accuracy) para los distintos modelos evaluados.

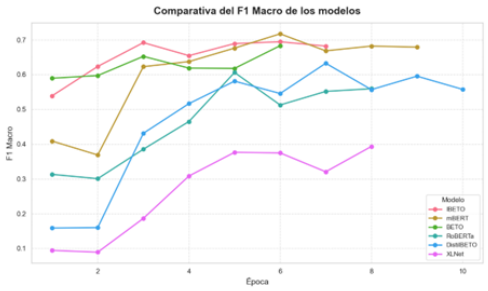


Figura 1: Evolución del F1 Macro por modelo durante el entrenamiento.

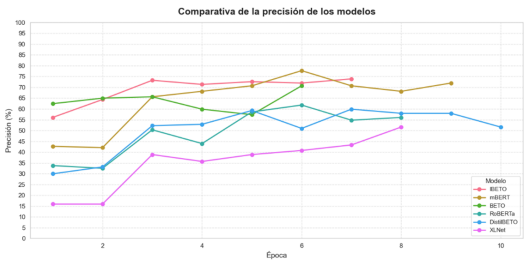


Figura 2: Evolución de la precisión ( %) de los modelos.

El análisis comparativo revela diferencias notables en el comportamiento y rendimiento de los modelos. El modelo mBERT alcanza el valor más alto de F1 Macro (0,72) y la mayor precisión pico (77 %), situándose como el más eficaz en esta tarea. Le sigue IBETO, con un rendimiento muy próximo (0,72 de F1 Macro y 74 % de precisión), y BETO, que mantie-

ne resultados consistentes (0,68 de F1 Macro y 70 % de precisión). En el rango inferior se encuentran RoBERTa, DistilBETO y XLNet, cuyos valores de F1 Macro y precisión se mantienen claramente por debajo de los tres modelos anteriores, lo que sugiere una menor adecuación a la tarea de etiquetado semántico en español técnico.

Un aspecto relevante es la estabilidad del entrenamiento. BETO destaca por su comportamiento más uniforme a lo largo de las épocas, seguido de IBETO, que presenta una progresión predecible y sin oscilaciones abruptas. En contraste, mBERT, pese a su rendimiento máximo, muestra mayor variabilidad entre épocas, lo que podría indicar un ajuste más sensible al tamaño del corpus o a la configuración de hiperparámetros.

Asimismo, los modelos con mejor desempeño global presentan tendencias de mejora sostenidas: mBERT muestra un incremento acumulado del 75,6 %, IBETO del 27,8 % y BETO del 15,3 % en sus respectivas métricas de rendimiento. Estos resultados sugieren que los tres modelos han logrado una convergencia adecuada, aunque con distintos grados de estabilidad y generalización.

En síntesis, IBETO, mBERT y BETO emergen como los modelos más adecuados para la tarea de etiquetado automático de densidad semántica en textos de ingeniería. Como indican Agerri y Agirre (2023) en su evaluación de modelos de lenguaje en español, las versiones multilingües no siempre superan a los monolingües cuando el dominio es técnico. mBERT destaca por su rendimiento máximo, IBETO combina precisión y coherencia en el dominio técnico y BETO ofrece la mayor estabilidad operativa durante el entrenamiento. Esta comparación reafirma la posible viabilidad del enfoque propuesto y orienta las futuras optimizaciones hacia configuraciones híbridas y corpus ampliados en español técnico. Dado que el estudio se realizó sobre un corpus limitado (1.569 oraciones) y con validación cruzada, pequeñas variaciones en la partición de los datos pueden influir en las métricas finales. Por ello, las diferencias observadas entre mBERT e IBETO no permiten afirmar una superioridad absoluta de un modelo sobre otro, sino más bien tendencias de desempeño dentro de un margen comparable.

El análisis de la matriz de confusión revela que las confusiones más frecuentes se producen entre las clases SD+ y SD++, seguidas de SD– y SD--. Este patrón refleja que las fronteras entre niveles contiguos son graduales y dependen del contexto discursivo.

### 4.3 Interpretación lingüística de los resultados

El análisis de activaciones internas del modelo IBETO muestra que la atención se concentra en los sustantivos técnicos, las nominalizaciones y los verbos de definición, mientras que los pronombres y conectores reciben pesos menores. Esta distribución es coherente con la noción de densidad semántica alta (SD++) descrita en la LCT (Maton & Doran, 2017): los modelos reconocen como informativamente centrales los elementos que condensan significado disciplinar. Por el contrario, las frases de baja densidad (SD--) contienen más verbos de acción, interjecciones o comentarios metadiscursivos, características que la red aprende a asociar a registros no técnicos.

El desempeño superior de IBETO frente a modelos generalistas sugiere que la representación de los conceptos técnicos requiere vocabularios ajustados y entrenamiento en dominios específicos. Esto es especialmente relevante en español, donde la terminología de ingeniería combina préstamos, siglas y nominalizaciones no siempre bien representadas en los corpus generales.

El hecho de que los modelos logren identificar patrones de densidad semántica sugiere que son capaces de reproducir parcialmente los procesos humanos de clasificación conceptual. No obstante, esta "comprensión" es estadística: el modelo detecta correlaciones léxicas y sintácticas que coinciden con categorías teóricas, sin captar su fundamento epistémico. Esta distancia entre representación algorítmica y significado disciplinar revela tanto el potencial como el límite del aprendizaje automático aplicado al lenguaje del conocimiento.

### 4.4 Predicción del tamaño del dataset para una precisión objetivo

El corpus utilizado en este estudio presenta un tamaño limitado, lo que restringe la capacidad de generalización de los modelos. Se realizó una estimación del volumen de datos necesario para alcanzar una precisión del 90 %, con el fin de establecer un objetivo de crecimiento realista para futuras fases del

proyecto. Los modelos seleccionados para este análisis predictivo fueron mBERT e IBETO, dado su rendimiento superior en el estudio comparativo previo y su idoneidad para tareas de etiquetado de la densidad semántica en español técnico.

Se generaron subconjuntos del corpus original mediante muestreo aleatorio, creando seis conjuntos de entrenamiento con tamaños de 100, 200, 500, 1.000, 1.200 y 1.569 oraciones. Un script en Python automatizó el proceso, garantizando que las muestras fueran disjuntas y representativas. Cada modelo fue entrenado con los distintos subconjuntos, registrando para cada uno la época con mejor rendimiento en las métricas Test Accuracy y F1 Macro.

| **Nº de oraciones** | **Test Accuracy (%)** | **F1 Macro** |
|---|---|---|
| 100 | 20.00 | 0.49 |
| 200 | 35.00 | 0.65 |
| 500 | 60.00 | 0.56 |
| 1000 | 64.00 | 0.63 |
| 1200 | 65.83 | 0.67 |
| 1569 | 77.71 | 0.72 |

Tabla 2: Resultados del modelo mBERT por tamaño del corpus.

| **Nº de oraciones** | **Test Accuracy (%)** | **F1 Macro** |
|---|---|---|
| 100 | 40.00 | 0.34 |
| 200 | 50.00 | 0.67 |
| 500 | 64.00 | 0.67 |
| 1000 | 66.00 | 0.66 |
| 1200 | 71.67 | 0.62 |
| 1569 | 73.25 | 0.69 |

Tabla 3: Resultados del modelo IBETO por tamaño del corpus.

Con los datos obtenidos se analizó la relación entre el tamaño del corpus y la precisión alcanzada, ajustando diferentes funciones matemáticas para determinar el patrón de crecimiento que mejor describiera el comportamiento de cada modelo. El criterio de selección fue el coeficiente de determinación ($R^2$), que mide el grado de correlación entre los datos experimentales y la función ajustada.

El análisis de correlación mostró que el comportamiento de mBERT sigue una tendencia logarítmica, con un coeficiente de determinación $R^2 = 0.96$, lo que indica un ajuste excelente (Figura 9 en el Anexo 2). Este resultado es coherente con las curvas de aprendizaje típicas de los modelos de machine learning, donde las mejoras iniciales son pronunciadas y se estabilizan a medida que aumenta el volumen de datos. Según la proyección obtenida, serían necesarias aproximadamente 3.309 oraciones para alcanzar una precisión del 90 %. Aunque esta estimación presenta cierto margen de variabilidad, se considera un objetivo alcanzable a medio plazo mediante la expansión progresiva del corpus.

Aplicando el mismo procedimiento, el modelo IBETO también mostró un ajuste logarítmico, con un $R^2 = 0.97$, lo que sugiere una correlación muy fuerte entre el tamaño del corpus y la precisión lograda (Figura 10 en el Anexo 2). La proyección indica que IBETO requeriría un corpus de aproximadamente 6.079 oraciones para alcanzar la precisión objetivo del 90 %. Este resultado es coherente con el mayor grado de especialización del modelo, que demanda un volumen más amplio de ejemplos para generalizar correctamente las estructuras semánticas del dominio técnico.

El análisis predictivo sugiere que el tamaño actual del corpus, aunque suficiente para un estudio exploratorio como este, debe incrementarse significativamente para alcanzar niveles de precisión próximos al 90 %. Los modelos evaluados siguen patrones logarítmicos típicos del aprendizaje automático, con ganancias rápidas en las fases iniciales y mejoras más lentas conforme aumenta el volumen de datos. En términos comparativos, mBERT presenta una trayectoria más eficiente, requiriendo menos ejemplos para aproximarse a la precisión objetivo, mientras que IBETO, pese a su mayor especialización, demanda un corpus más extenso. Estas estimaciones orientan las fases futuras del proyecto, priorizando la ampliación gradual del conjunto de datos y la validación periódica del modelo durante el proceso de etiquetado. En conjunto, los resultados respaldan la validez del enfoque adoptado y establecen un referente cuantitativo claro para la expansión del corpus y la optimización de los modelos Transformer en tareas de etiquetado de la densidad semántica en español.

## 5 Aplicaciones y limitaciones

La automatización del etiquetado de la densidad semántica (SD) en el marco de la Legitimation Code Theory (LCT) permite abordar tareas que antes resultaban inviables por su alta carga manual. Entre las aplicaciones más directas destacan el análisis de materiales educativos para detectar zonas de alta o baja densidad semántica en clases o manuales docentes, la evaluación del discurso académico en textos de estudiantes o investigadores, y la minería de conocimiento en corpus técnicos extensos. A medio plazo, estos modelos podrían integrarse en sistemas de tutoría inteligente o plataformas de aprendizaje adaptativo, ajustando la complejidad de los contenidos al nivel de comprensión del usuario.

La principal limitación del estudio es el tamaño del corpus, que restringe la capacidad de generalización del modelo. Aunque el etiquetado manual se realizó con protocolos validados, conserva cierto grado interpretativo. Futuras investigaciones deberían ampliar el corpus (Gutiérrez-Fandiño et al., 2022; Rajpurkar et al., 2018), incorporar más áreas de ingeniería y explorar modelos instruccionales o generativos (LLMs) que integren un contexto discursivo más amplio. Asimismo, conviene realizar evaluaciones de transferencia hacia otros dominios para analizar la robustez del modelo y su capacidad de adaptación.

El estudio sugiere además que el español técnico presenta particularidades lingüísticas como nominalizaciones, estructuras pasivas o conectores discursivos, que justifican la creación de corpus especializados y de estrategias de fine-tuning sensibles a la variedad lingüística. En este sentido, la aparición de modelos de lenguaje especializados en español, como BETO (Cañete et al., 2020), junto con los avances en su adaptación a dominios técnicos y científicos, supone un paso decisivo hacia una infraestructura lingüística en español más representativa del conocimiento científico. En esta línea se enmarca el desarrollo de IBETO, modelo adaptado en el presente trabajo a partir de BETO con textos técnicos en español.

## 6 Conclusiones

Este trabajo ha presentado la evaluación experimental de un sistema de etiquetado automático de la densidad semántica en textos de ingeniería en español, combinando un corpus anotado manualmente con modelos Transformer preentrenados. Los resultados muestran que es posible clasificar automáticamente, con un grado elevado de precisión, los niveles de densidad semántica previamente definidos en el marco de la LCT. En términos cuantitativos, mBERT obtuvo el valor más alto de F1 Macro y accuracy. No obstante, la diferencia respecto a IBETO fue moderada y se mantiene dentro de un margen reducido considerando el tamaño y desbalance del corpus. Considerando que IBETO fue adaptado previamente con textos técnicos en español, su desempeño estable y su alineación con el dominio especializado lo posicionan como una base especialmente adecuada para futuras ampliaciones del corpus y mejoras del sistema. Los análisis de atención y error sugieren que los modelos identifican patrones léxicos y sintácticos asociados a la densidad semántica, validando su aplicabilidad al discurso académico y técnico.

En el futuro, la ampliación del corpus y la exploración de nuevas arquitecturas (DeBERTa, GPT en español) permitirán aplicar este sistema al análisis de clases completas o materiales educativos digitales. A largo plazo, la integración de modelos de lenguaje en entornos de aprendizaje podría mejorar la comprensión y comunicación del conocimiento técnico en español, aportando una herramienta de apoyo a la docencia y la investigación.

Finalmente, este estudio subraya que la automatización del análisis del lenguaje no solo incrementa la eficiencia, sino que ofrece una nueva perspectiva sobre cómo se codifica y transmite el conocimiento. Cada avance en PLN implica equilibrar la precisión estadística con el significado humano. El verdadero desafío no reside únicamente en entrenar modelos más exactos, sino en diseñar sistemas capaces de captar las estructuras simbólicas que sustentan la comunicación del saber.

## 7 Referencias

Agerri, R. & Agirre E. (2023). Lessons learned from the evaluation of Spanish Language Models. Procesamiento del Lenguaje Natural, 70, 157-170. https://doi.org/10.26342/2023-70-13

Argüelles-Álvarez, I. & Morton, T. (2023). Using legitimation code theory to investigate English medium lecturers' knowledge-building practices. Journal of English for

Academic Purposes, 65, 101285. https://doi.org/10.1016/j.jeap.2023.101285.

Bai, Y., Chen, M., Zhou, P., Zhao, T., Lee, J. D., Kakade, S., Wang, H., & Xiong, C. (2021). How important is the train-validation split in meta-learning? arXiv. https://arxiv.org/abs/2010.05843

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3615–3620). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Bengoetxea, K., & Gonzalez-Dios, I. (2021). MultiAzterTest: A multilingual analyzer on multiple levels of language for readability assessment. arXiv. https://arxiv.org/abs/2109.04870

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. Sebastopol, CA: O'Reilly Media.

Blackie, M., Adendorff, H. & Mouton, M. (eds.) (2023). Enhancing Science Education: Exploring Knowledge Practices with Legitimation Code Theory. Routledge.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2023). Spanish pre-trained BERT model and evaluation data. arXiv. https://arxiv.org/abs/2308.02976

de la Rosa, J., Ponferrada, E. G., Villegas, P., González de Prado Salas, P., Romero, M., & Grandury, M. (2022). BERTIN: Efficient Pre-Training of a Spanish Language Model via Perplexity Sampling. Procesamiento del Lenguaje Natural, 68, 13-23. https://doi.org/10.48550/arXiv.2207.06814

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Pio-Carrino, C., Armentano-Oller, C., Rodríguez-Penagos, C., Gonzalez-Agirre, A., & Villegas, M. (2022). MarIA: Spanish language models. Procesamiento del Lenguaje Natural, 68, 39–60. https://doi.org/10.26342/2022-68-3

Guzzi, E., & Alonso Ramos, M. (2023). Sofisticación y diversidad como medidas de complejidad léxica para determinar el perfil colocacional de textos académicos en español. Revista Signos. Estudios De Lingüística, 56(112). Recuperado a partir de https://revistasignos.cl/index.php/signos/article/view/872

Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Pearson.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In Proceedings of the International Conference on Learning Representations (ICLR 2020).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D.,Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Manning, C. & Schütze, H. (1999). Foundations of statistical natural language processing. The MIT Press.

Maton, K. (2014). Knowledge and Knowers: Towards a Realist Sociology of Education. Routledge.

Maton, K. & Chen, R. T-H. (2016) LCT in qualitative research: Creating a translation device for studying constructivist pedagogy, in K. Maton, S. Hood & S. Shay (eds) Knowledge-building: Educational studies in Legitimation Code Theory. London: Routledge.

Maton, K., Hood, S. & Shay, S. (eds.) (2016) Knowledge-building: Educational Studies in Legitimation Code Theory. Routledge.

Maton, K., & Doran, Y. J. (2017). Semantic density: A translation device for revealing complexity of knowledge practices in discourse, Part 1 - Wording. Onomázein, (Número Especial II), 46–76. https://doi.org/10.7764/onomazein.ne2.03

Maton, K., Martin, J. R. & Doran, Y. J. (eds) (2021) Teaching Science: Knowledge, Language, Pedagogy. Routledge.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 784-789.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019) (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017/2023). Attention is all you need (arXiv:1706.03762v7). arXiv. https://arxiv.org/abs/1706.03762

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zhang, X., & Lu, X. (2025). Aligning linguistic complexity with the difficulty of English texts for L2 learners based on CEFR levels. Studies in Second Language Acquisition, 47, 5, December 2025, 1407 – 1434. https://doi.org/10.1017/S0272263125101125

## *A Anexo 1. Resultados visuales de entrenamiento de los modelos*

Se muestran en este anexo las representaciones gráficas que acompañan al análisis de los resultados experimentales descritos en la sección de resultados. Estas figuras permiten observar de manera visual el comportamiento de los distintos modelos a lo largo del proceso de entrenamiento, aportando una visión complementaria al análisis cuantitativo. Las Figuras 3 a 8 muestran la evolución de las principales métricas de rendimiento —Train Loss, Validation Loss, Accuracy y F1— para cada una de las arquitecturas evaluadas (BETO, mBERT, RoBERTa, DistilBETO, XLNet e IBETO). Su interpretación conjunta permite apreciar las diferencias en estabilidad, velocidad de convergencia y capacidad de generalización de los modelos, ofreciendo así una comprensión más completa de su desempeño en la tarea de etiquetado automático de densidad semántica.

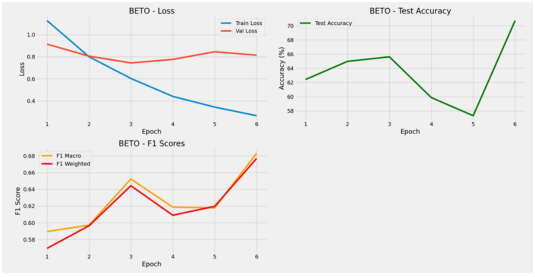


Figura 3: BETO

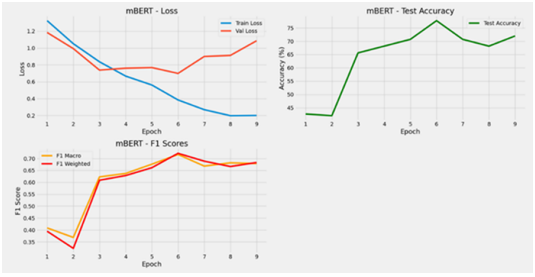


Figura 4: mBERT

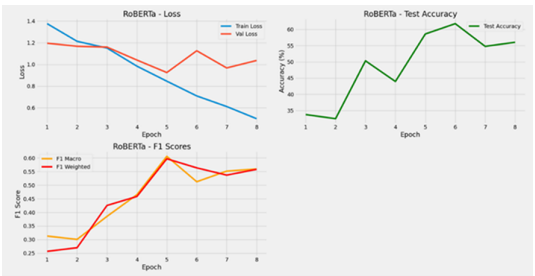


Figura 5: RoBERTa

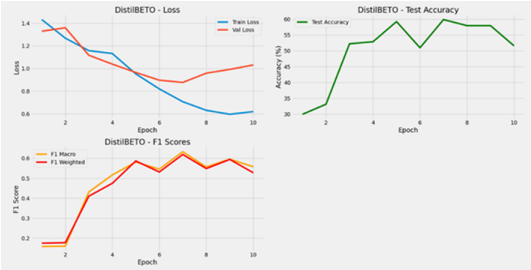


Figura 6: DistilBETO

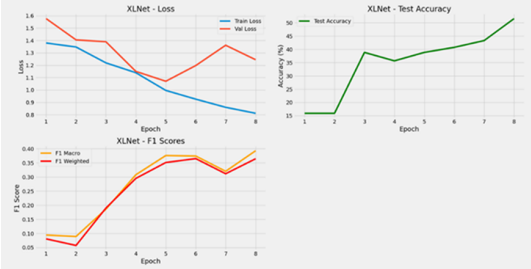


Figura 7: XLNeT

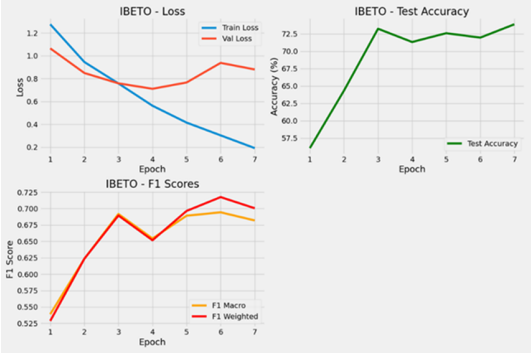


Figura 8: IBETO

# *B Anexo 2. Análisis de correlación y proyección del rendimiento en función del tamaño del corpus*

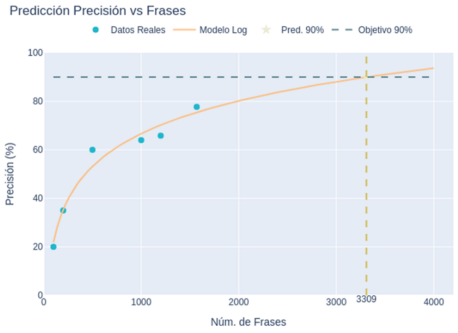


Figura 9: Relación entre tamaño del corpus y precisión para mBERT.

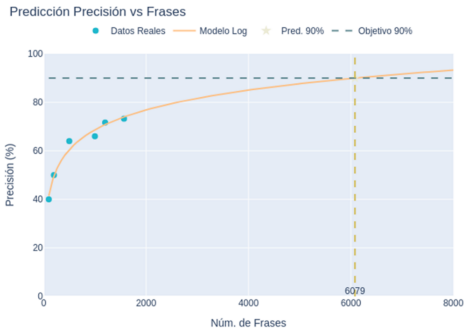


Figura 10: Relación entre tamaño del corpus y precisión para IBETO