2 DE JULIO DE 2025
FACULTAD DE INFORMÁTICA. UNIVERSIDAD DE MURCIA
GRADO EN INGENIERÍA INFORMÁTICA
TRABAJO FIN DE GRADO

Creación de perfiles de usuario a partir de sus contribuciones textuales en Youtube

---

Autor: Raúl Martínez Campos

Tutor: Rafael Valencia García

Cotutor: José Antonio García Díaz

Índice

Índice de figuras

Resumen

Este trabajo aborda la detección automática de la ideología política en emisores digitales mediante el análisis de transcripciones de vídeos publicados en YouTube. Para ello, se ha construido un corpus original compuesto por más de 19.000 vídeos procedentes de 21 canales distintos, clasificados según su orientación ideológica en cuatro categorías: izquierda, izquierda moderada, derecha moderada y derecha. A partir de este conjunto de datos, se agruparon vídeos en clústeres ideológicos y se entrenaron diversos modelos de lenguaje preentrenados —como BETO, RoBERTa-BNE, DistilBERT y mBERT— para realizar la tarea de clasificación.

Los experimentos realizados revelan un alto rendimiento en escenarios donde los datos de entrenamiento y prueba comparten canales, con valores de precisión y macro F1 superiores al 0.98. Sin embargo, al introducir una separación estricta por canal entre los conjuntos, el rendimiento cae por debajo del azar, lo que evidencia una fuerte dependencia de los modelos respecto a señales estilísticas o léxicas específicas de cada emisor. Esta hipótesis se refuerza mediante una evaluación externa con el corpus anotado *PoliticES*, que confirma el correcto funcionamiento de los modelos en un entorno controlado y con anotaciones más coherentes a nivel discursivo.

Además del análisis cuantitativo, se ha llevado a cabo una exploración cualitativa del corpus, que ha puesto de manifiesto varios factores que dificultan la clasificación, como la ambigüedad ideológica de algunos vídeos, la inclusión de contenido neutro o poco informativo, y la inconsistencia en la asignación ideológica en algunos casos, al haberse realizado a nivel de canal en lugar de por vídeo. En consecuencia, aunque las arquitecturas utilizadas demuestran un gran potencial técnico, se concluye que el rendimiento práctico en tareas como esta depende en gran medida de la calidad, definición y consistencia del corpus. Como línea de mejora prioritaria, se propone refinar el proceso de anotación mediante la asignación de etiquetas ideológicas a nivel individual de vídeo, lo que permitiría reducir la ambigüedad y mejorar la capacidad de generalización de los sistemas desarrollados.

Extended Abstract

The growing digitalization of political discourse has profoundly and irreversibly transformed contemporary communication dynamics. In recent decades, traditional media have gradually relinquished their central role to digital platforms that allow a broad array of communicators—from official political parties to independent YouTubers—to broadcast their messages globally and continuously. YouTube, in particular, has solidified its role as a major platform for political communication not only due to its reach, but also thanks to its flexibility in hosting a wide variety of formats, tones, and rhetorical strategies. This democratization of political expression raises important questions for computational approaches to language analysis.

This shift in the ways political content is produced and consumed presents new challenges and opportunities for automated language analysis. Among them is the possibility of identifying the ideological orientation of content creators operating in these spaces—not through manual observation or editorial analysis, but via computational tools that work directly with linguistic data. In this context, the question of whether it is feasible to automatically detect the ideology of video content creators based on their textual transcripts—generated by YouTube itself—becomes particularly relevant.

This study addresses that question from an empirical perspective, combining natural language processing techniques and machine learning on a text corpus derived from Spanish-language YouTube videos. The main objective is to assess whether the speeches contain sufficiently stable linguistic patterns to enable automatic classification of the speaker's ideology. To this end, pretrained language models such as BETO and RoBERTa-BNE are applied to textual representations built from the transcripts, aiming to classify them into one of four ideological categories.

To carry out this analysis, a custom corpus was designed and constructed, comprising 19,730 Spanish-language videos extracted from 21 manually selected YouTube channels. Special effort was made to ensure a representative diversity of speaker types. Therefore, the corpus includes not only political parties and media institutions, but also individual content creators whose presence in public debate has grown significantly in recent years. This variety captures different discursive styles: from the formal language of traditional media to the more spontaneous and direct expressions typical of YouTubers. Including these profiles makes it possible to study how ideology manifests at different levels of formality and in various registers of political language. Each channel was categorized into one of four ideological groups: left, center-left, center-right, or right, based on its editorial stance, political affiliations, and qualitative content analysis. The final corpus is evenly distributed across the four ideological classes: right (5,101 videos), center-right (5,102), center-left (5,093), and left (4,434). This balance ensures that models are not biased toward a dominant class and enables the use of metrics like macro F1-score to fairly evaluate class-wise performance.

Data collection was carried out using the YouTube Data API and automated tools for retrieving platform-generated transcripts. For each video, relevant metadata (title, date, hashtags) was extracted, along with the full Spanish transcript. Since the transcripts are auto-generated, various filters and preprocessing steps were applied to remove

non-linguistic annotations, normalize the text, and control input length. Videos without available transcripts were discarded, which required exploring a large number of videos per channel, even reaching back to videos from 2014. An inherent limitation of the corpus is the quality of automatic transcripts. These may contain segmentation, punctuation, and speech recognition errors—especially in videos with background noise, regional accents, or interruptions. Such errors can introduce semantic noise that affects both textual representation and the model training process.

An additional key methodological limitation is that all ideological labels were assigned at the channel level, meaning each video inherits the ideology of its source without individualized analysis. While this simplifies corpus construction, it ignores the internal diversity of many channels, which may feature content of varying orientation or neutral character. It also overlooks potential shifts in editorial line or political positioning over time—a particularly relevant phenomenon in highly polarized contexts or electoral periods.

To reduce bias and ensure greater representativeness, data were grouped into ideologically homogeneous clusters, and predictions were made at the cluster level. For each cluster, predictions were made individually on each transcript, and the final ideological class was assigned based on the mode of those predictions. This aggregation approach helps mitigate the impact of outliers and yields more stable group-level classifications. Two main corpus partitioning strategies were explored: one that allows mixing videos from the same channel in both training and test sets, and a stricter one that fully separates channels between the two sets, forcing the model to generalize beyond known senders.

The resulting texts, composed of the video title and its transcript, were used as input for a series of Transformer-based language models. Initially, the inclusion of video descriptions and hashtags was also considered, as these elements are part of the metadata generated by content creators. However, preliminary analysis showed that these fields were often redundant, uninformative, or repeated across videos from the same channel, with many descriptions merely copying the title or adding generic promotional lines, and hashtags offering little ideological value. As a result, only the title and full transcript were retained as input features. Each input was tokenized with a maximum limit of 512 tokens, applying truncation when necessary and prioritizing the inclusion of the title. During fine-tuning, models were trained with a learning rate of 2e-5, a batch size of 16, and a maximum of five epochs, following common empirical configurations. These decisions balanced training stability, computational cost, and predictive performance, ensuring all models were compared under equivalent experimental conditions. Pretrained models such as BETO (a Spanish-optimized version of BERT), DistilBERT (a lightweight version of BERT offering a good trade-off between performance and efficiency), RoBERTa-BNE (trained on large journalistic and institutional corpora), and the multilingual mBERT were used. These were fine-tuned and evaluated using the Hugging Face Transformers library, respecting tokenization and maximum length requirements. The goal was to compare their performance in the multiclass ideological classification task, identifying not only which model performed best, but also under which conditions the results were most reliable or generalizable.

Before proceeding to evaluation, key metrics were defined to compare model performance: accuracy, macro F1-score, precision, and recall, along with confusion matrices. Macro average was prioritized to compensate for potential class imbalances. These metrics offer a comprehensive view of system behavior, both overall and by individual ideological class.

The experiments revealed a marked duality in model behavior depending on the corpus partition strategy. In favorable scenarios—where mixing videos from the same channel across training and test sets is allowed—models achieved outstanding performance. Experiments with different cluster sizes (5, 10, and 20 transcripts) were conducted to assess model stability with more aggregated information per sample. In all cases, models like BETO, DistilBETO, and RoBERTa-BNE achieved accuracies and macro F1-scores above 0.98 in the four-class ideological classification task. These results confirm the strong capability of modern language models to capture statistical patterns in political discourse, provided the evaluation environment does not demand strong structural generalization.

However, this capacity drops drastically when a strict channel separation is introduced, requiring models to confront unseen senders during testing. In this more realistic and demanding configuration, all models experienced sharp performance declines, with accuracies below 15 % and similarly low macro F1-scores, yielding results worse than random guessing. This contrast suggests that models are not learning to identify ideologies based on universal linguistic features, but rather relying on channel-specific signals such as discursive style, recurring terms, or particular topics. In other words, the system can recognize "who" is speaking, but not necessarily "what" ideology they represent in a broader context.

The confusion matrices reflect this fragility. In the favorable scenario, all four ideological classes were correctly identified in more than 96 % of cases, with only marginal errors. Most misclassifications occurred between adjacent ideological categories, such as "left" and "center-left," or "right" and "center-right," which is expected due to the semantic proximity of their discourses. In contrast, under the channel-separation scenario, the RoBERTa-BNE model showed a clear loss of discriminative power: the "center-right" class was never recognized, and both "center-left" and "left" were misclassified as "center-right" in over 23 % and 16 % of cases, respectively. This behavior indicates that the model fails to identify general ideological patterns and instead relies on superficial features, particularly confusing the moderate positions on the political spectrum.

To confirm that the models were properly configured and that the poor results observed earlier were not due to technical errors, a cross-evaluation was performed using an external dataset: *PoliticES*, a collection of Spanish-language tweet clusters manually labeled according to political ideology and used as a benchmark in the IberLEF 2023 task. The RoBERTa-BNE model, trained on this corpus, achieved a macro F1-score of 0.64, a value comparable to those reported by top-performing systems in the benchmark. Moreover, the resulting confusion matrix shows a reasonably balanced distribution across classes, with particularly strong performance in identifying center-left and center-right discourse. Most misclassifications occurred between ideologically

adjacent classes, suggesting that the model was able to learn general and transferable ideological features when provided with coherent annotations and a well-structured dataset. This reinforces the idea that the limitations observed in the YouTube corpus are primarily due to data quality issues, rather than flaws in the architecture or training of the models used.

Additionally, an alternative scenario was evaluated: a model trained on *PoliticES* was directly applied to the YouTube corpus. While overall performance was modest, it exceeded the random baseline, achieving an accuracy of 45 % and a macro F1-score of 0.41. Notably, the model was able to correctly identify "center-left" content with a recall of 0.98, although performance on other classes was more erratic. The confusion matrix revealed some class imbalance and confusion between ideologically adjacent positions, but the results nevertheless indicate partial transferability of learned ideological patterns. This finding reinforces the hypothesis that, in the absence of clear structure and reliable labeling, even state-of-the-art language models struggle to robustly extract generalizable ideological information from noisy or inconsistently annotated corpora.

To further understand the causes of the observed model limitations, a qualitative analysis of the training corpus was conducted, adopting two complementary approaches: on one hand, lexical distribution by ideological class was studied after applying linguistic cleaning filters; on the other, a manual inspection of content and communicative contexts in a sample of videos was performed.

The lexical analysis revealed that even after applying cleaning techniques—such as removing stop words using tools like NLTK—and using discriminative metrics like mutual information (`mutual_info_classif`), many of the most frequent words lacked differentiating value across ideological classes. Functional terms like "si" or "gracias" appeared uniformly across all speeches, while potentially meaningful terms—such as "vox," "podemos," or "palestina"—showed ambiguous usage patterns: they were used both in supportive and critical contexts, complicating automatic interpretation without considering semantic context. This ambiguity highlights the need for models that can capture not just term presence but also their pragmatic orientation in discourse, requiring more contextual approaches and precise input segmentation.

Additionally, an asymmetric distribution of discriminative vocabulary was identified: the extreme ideological classes (left and right) concentrated a higher number of distinctive terms, while the moderate classes were associated with more neutral or ambiguous vocabulary. This imbalance introduces structural bias into the models, which tend to overclassify toward the extremes, weakening the system's ability to detect more nuanced discourse.

The manual video review revealed multiple sources of ambiguity. For example, there were cases where media outlets of one ideology fully reproduced statements from opposing figures without adding their own editorial framing. Other videos contained essentially neutral content—such as personal interviews, cultural events, institutional news, or anecdotal information—that, although produced by ideologically labeled channels, lacked sufficient clues to infer a clear political orientation. These situations compromise label consistency and, therefore, the effectiveness of the models trained on them. Tagging all videos by channel ideology, while convenient, ignores the internal

heterogeneity of many channels and exacerbates the problems already described. It becomes one of the main bottlenecks in developing truly accurate and generalizable systems.

Therefore, the study concludes that while current models for ideological classification possess great technical potential, their performance in realistic scenarios depends critically on the quality, definition, and consistency of the input data. As a priority improvement, the need to individually label videos—rather than assuming a homogeneous ideology at the channel level—is emphasized. This change would better capture discursive variability and reduce label noise, laying the groundwork for more robust future applications in automated political analysis.

1. Introducción

La transformación digital ha modificado significativamente los modos de producción, difusión y consumo de información política. En las últimas décadas, el auge de las redes sociales y las plataformas de contenido abierto, como Twitter o YouTube, ha permitido que millones de usuarios expresen sus opiniones y participen activamente en la construcción del discurso político. Este cambio ha multiplicado las fuentes de información política, descentralizando la comunicación y dando lugar a nuevas formas de interacción ideológica en el espacio público.

Frente a este nuevo ecosistema comunicativo, la capacidad de analizar automáticamente el contenido generado por los usuarios se ha convertido en una herramienta clave para investigadores, medios y organizaciones. En particular, la *detección de la ideología política* a partir del lenguaje escrito se ha consolidado como un reto relevante en el campo del *Procesamiento de Lenguaje Natural* (PLN). Esta investigación busca inferir de manera automática la orientación ideológica implícita en textos.

El interés por esta problemática no solo es técnico, sino también social y político. Comprender cómo se manifiestan las posturas ideológicas a través del lenguaje puede ayudar a detectar dinámicas de polarización, estudiar la evolución del discurso político, combatir la desinformación o diseñar sistemas de recomendación más transparentes. Sin embargo, el lenguaje político presenta múltiples desafíos: es a menudo ambiguo, estratégico e influido por el contexto cultural y discursivo. Además, plataformas como YouTube introducen complejidades añadidas, como la variabilidad temática, la ausencia de etiquetas explícitas y la diversidad lingüística de sus contenidos.

Para afrontar estos retos, este trabajo se apoya en avances recientes del aprendizaje profundo, en particular en modelos basados en la arquitectura *Transformer*, que han revolucionado el tratamiento del lenguaje gracias a su capacidad para capturar dependencias contextuales de largo alcance y representar el significado de forma más precisa. El objetivo principal es evaluar la eficacia de estos modelos para la detección de la ideología política en textos en español, tomando como unidad de análisis textual el conjunto de publicaciones realizadas por los canales de YouTube.

El enfoque adoptado combina varias etapas: desde la recolección y limpieza de datos, hasta el diseño de representaciones agregadas de los canales, pasando por la experimentación con diferentes variantes del modelo BERT adaptadas al español. La investigación se centra en un corpus contemporáneo anotado con etiquetas ideológicas, lo que permite analizar el rendimiento de los modelos en un entorno realista y desafiante.

La estructura de este documento es la siguiente:

- En el **capítulo 2** se presenta el estado del arte, que incluye antecedentes en la clasificación de textos ideológicos y los desarrollos más recientes en PLN y modelos basados en *Transformers*.
- El **capítulo 3** describe los objetivos del trabajo y la metodología empleada, abordando la selección del corpus, las técnicas de preprocesamiento y el diseño experimental.

- En el **capítulo 4** se detalla el desarrollo del sistema junto con los resultados obtenidos, analizando el rendimiento de los modelos propuestos y sus principales limitaciones.
- Finalmente, el **capítulo 5** recoge las conclusiones del estudio y plantea posibles líneas de investigación futura para mejorar la detección automática de ideología política en entornos digitales.

2. Estado del arte

La detección automática de la ideología política a partir de texto es una tarea compleja y relevante que ha cobrado especial importancia con la proliferación de contenido generado por usuarios en plataformas digitales. A medida que la comunicación política se ha desplazado hacia entornos abiertos como Twitter, YouTube o foros públicos, se abre la posibilidad de aplicar técnicas de análisis del lenguaje para inferir la orientación ideológica de los emisores [6]. Este tipo de análisis resulta útil en contextos como el estudio del discurso político, la investigación sociológica, la recomendación personalizada de contenidos o la caracterización de comunidades digitales.

Esta problemática se sitúa en la intersección entre el *Procesamiento de Lenguaje Natural* (PLN), la lingüística computacional y la ciencia política. En particular, el reto no solo reside en analizar el contenido explícito de los mensajes, sino también en capturar patrones latentes, sesgos implícitos y estructuras discursivas que revelan afiliaciones ideológicas [19].

Además, el estudio de la ideología en plataformas como YouTube introduce nuevos factores: la informalidad del lenguaje, la escasez de etiquetas, la variabilidad temática y el papel del canal como unidad de agregación. Por ello, este trabajo se apoya en enfoques recientes de aprendizaje profundo y modelado lingüístico contextual para representar usuarios en función del texto que producen, con el objetivo final de clasificar perfiles ideológicos y descubrir similitudes entre canales.

2.1. Procesamiento de Lenguaje Natural y clasificación de textos

El *Procesamiento de Lenguaje Natural* (PLN) es una rama de la inteligencia artificial que permite a las máquinas comprender y manipular el lenguaje humano [15]. Dentro de las tareas más relevantes del PLN se encuentra la *clasificación de texto*, que consiste en asignar a un documento una categoría específica en función de su contenido. Esta técnica ha sido ampliamente utilizada en dominios como la detección de spam, el análisis de sentimientos, la clasificación temática o la detección de ideología política.

En los enfoques iniciales, el texto era representado mediante modelos sencillos como *bag-of-words* (BoW) o TF-IDF, que trataban el texto como una bolsa de palabras sin considerar el orden ni el contexto. Estos vectores de características alimentaban algoritmos clásicos de aprendizaje automático como *Naive Bayes*, máquinas de vectores de soporte (SVM) o regresión logística. A pesar de su simplicidad, lograban buenos resultados en entornos formales como editoriales o discursos parlamentarios [19].

No obstante, estos métodos presentaban limitaciones importantes: no capturaban la semántica del lenguaje, ignoraban relaciones contextuales y requerían una fuerte ingeniería de características. Para superar estas barreras, se introdujeron los *word embeddings*, como Word2Vec [16] y GloVe [18], que representan las palabras como vectores densos en espacios semánticos continuos, permitiendo modelar relaciones como analogías y similitudes semánticas.

Posteriormente, el avance del aprendizaje profundo introdujo el uso de redes neuro-

nales, en particular redes recurrentes (RNN) y sus variantes como LSTM [13] y GRU [5]. Estas arquitecturas permitieron modelar el lenguaje de forma secuencial, capturando dependencias a corto y largo plazo. Su aplicación en la clasificación de textos políticos aportó mejoras al considerar el contexto de aparición de las palabras, una dimensión crucial en discursos con carga ideológica [14].

La detección automática de ideología política se sitúa como una tarea especialmente desafiante dentro de la clasificación textual. Implica inferir la postura ideológica de un autor o mensaje a partir del lenguaje, incluso cuando no se mencionan explícitamente partidos, etiquetas o ideologías. En redes sociales, como Twitter o YouTube, esta tarea se complica aún más por la informalidad, la ironía, los códigos culturales y la ambigüedad estratégica en el lenguaje utilizado [1].

En este contexto, la representación semántica y contextual del texto es clave. El discurso político puede estar marcado por sutiles diferencias léxicas (como el uso de "redistribución" frente a "libertad") o por estructuras discursivas complejas que reflejan ideologías subyacentes. Por ello, los métodos tradicionales han sido progresivamente reemplazados por enfoques más robustos que permiten una comprensión más profunda del lenguaje, y que serán abordados en el siguiente apartado dedicado a la arquitectura Transformer.

Finalmente, en el ámbito hispanohablante, se ha avanzado significativamente en la disponibilidad de modelos entrenados específicamente para el español y de corpus anotados con etiquetas ideológicas, como PoliticES [10]. Esta tarea buscó extraer, entre otros rasgos personales, la ideología política de los usuarios de redes sociales, basándose en sus publicaciones.

Estas limitaciones abren paso al uso de arquitecturas más sofisticadas, como los Transformers, que abordaremos a continuación.

2.2. Transformers

En los últimos años, los modelos basados en la arquitectura *Transformer* [23] han transformado radicalmente el campo del procesamiento de lenguaje natural. Los Transformers introdujeron el mecanismo de *self-attention*, que permite modelar relaciones contextuales entre todas las palabras de una secuencia sin necesidad de estructuras recurrentes o convolucionales. Este cambio estructural permitió mejorar el rendimiento, paralelizar el entrenamiento y capturar dependencias de largo alcance de manera más eficiente.

Para comprender su funcionamiento, en la Figura 1 se muestra la arquitectura original propuesta por Vaswani et al. [23], en la que se distinguen dos bloques principales: el **codificador** (izquierda) y el **decodificador** (derecha). Aunque algunos modelos modernos, como BERT, utilizan únicamente el codificador, otros como GPT [3] o T5 [20] emplean toda la arquitectura, incluyendo ambos componentes.

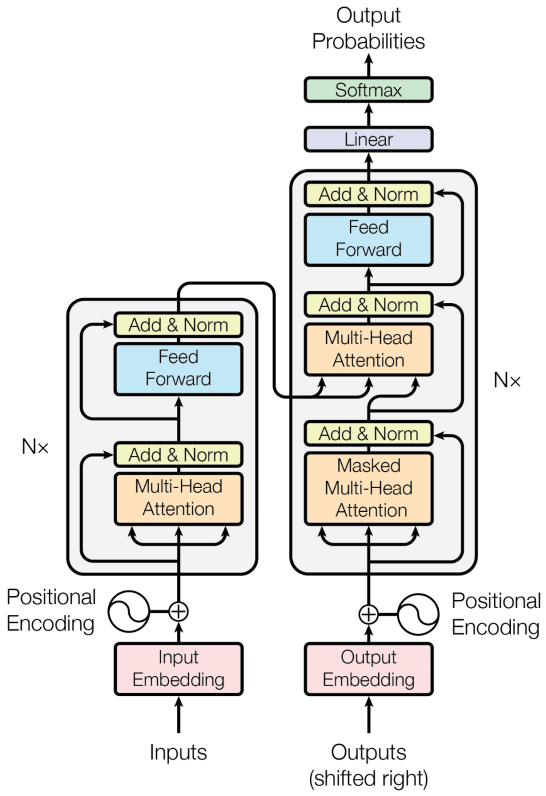

Figura 1: Arquitectura del modelo Transformer [23].

Cada bloque del codificador contiene un mecanismo de **atención multi-cabeza** (*Multi-Head Attention*), que permite al modelo enfocarse simultáneamente en distintas partes de la secuencia de entrada. Luego, se aplica una red *feed-forward* totalmente conectada a cada posición de forma independiente, seguida por una operación de normalización (*Add & Norm*) y conexiones residuales.

El **decodificador** incorpora una atención enmascarada para evitar que el modelo acceda a posiciones futuras durante la generación de texto, así como una atención cruzada (*cross-attention*) sobre las salidas del codificador, que le permite integrar el contexto completo de la entrada.

A partir de esta arquitectura se han desarrollado múltiples modelos de lenguaje preentrenados, siendo uno de los más influyentes BERT (*Bidirectional Encoder Representations from Transformers*) [7]. BERT introdujo el preentrenamiento sobre grandes corpus textuales utilizando tareas de enmascaramiento de palabras (*masked language modeling*) y predicción de frases adyacentes. Posteriormente, se adapta a tareas específicas mediante *fine-tuning*, lo que ha permitido lograr resultados de estado del arte en clasificación de texto, reconocimiento de entidades, inferencia textual, entre otras.

En el ámbito hispanohablante, se han desarrollado variantes adaptadas al idioma. Uno de los primeros fue BETO [4], entrenado exclusivamente sobre textos en español extraídos de Wikipedia, prensa y redes sociales. Este modelo ha sido ampliamente adoptado en tareas de clasificación, especialmente en entornos donde el lenguaje

presenta características informales o variantes regionales. Otra alternativa robusta es RoBERTa-BNE [8], una versión ajustada y optimizada para el español, desarrollada por la Biblioteca Nacional de España. Su entrenamiento sobre corpora de alta calidad lo hace particularmente efectivo en tareas lingüísticas complejas y formales.

También se ha popularizado el uso de versiones ligeras como DistilBERT [22], una red compacta obtenida mediante técnicas de distilación de conocimiento a partir de BERT. Este modelo mantiene buena parte del rendimiento del original, pero con una reducción considerable en el tamaño y tiempo de inferencia, lo que lo convierte en una opción viable para aplicaciones con restricciones computacionales o necesidad de respuesta en tiempo real.

Los modelos basados en Transformers han demostrado ser especialmente útiles en tareas de detección ideológica, ya que permiten capturar matices sutiles del lenguaje político, como el uso estratégico de términos, la elección de marco semántico, la polarización discursiva o la evasión de etiquetas ideológicas explícitas. En particular, su capacidad para modelar el contexto completo de una oración resulta crucial cuando la ideología no se manifiesta directamente a través de palabras clave, sino mediante asociaciones implícitas o construcciones retóricas.

3. Objetivos y metodología

Tras explicar el desarrollo del *Procesamiento de Lenguaje Natural* a lo largo de la historia, pasamos a exponer los objetivos que pretenden conseguirse en este trabajo y la metodología empleada para alcanzarlos.

3.1. Objetivos

Este trabajo tiene como objetivo general investigar la viabilidad y efectividad de los modelos de lenguaje actuales, en particular aquellos basados en la arquitectura *Transformer*, para la detección automática de la ideología política a partir del análisis del lenguaje empleado en plataformas digitales. Para ello, se plantea el estudio de casos centrado en canales de YouTube en español, con el fin de evaluar la capacidad de los modelos para captar matices ideológicos presentes en un entorno informal, dinámico y poco estructurado.

De manera específica, el proyecto persigue dos objetivos principales:

- **Construcción de un corpus anotado con etiquetas ideológicas.** El primer objetivo consiste en la creación de un conjunto de datos que refleje la diversidad ideológica de los contenidos políticos presentes en YouTube. Este corpus no solo servirá como base para la experimentación, sino que también busca aportar un recurso útil a la comunidad investigadora, que puede ser incrementado y mejorado con el tiempo, dada la escasez actual de datos públicos en este ámbito específico.
- **Evaluación comparativa de modelos basados en Transformers.** El segundo objetivo se centra en la experimentación con distintos modelos de lenguaje preentrenados, adaptados o no al español, como BETO, RoBERTa-BNE, DistilBERT y mBERT. Se busca determinar su rendimiento en la tarea de clasificación ideológica, comparando su capacidad para representar usuarios a partir del texto que producen y su habilidad para captar diferencias ideológicas implícitas. Esta evaluación contempla distintas estrategias de representación textual y segmentación, así como un análisis detallado de los errores y limitaciones observadas.

Estos objetivos se articulan en torno a una doble contribución: por un lado, la generación de un nuevo recurso empírico específicamente diseñado para la detección de ideología en lengua española; por otro, la validación de enfoques avanzados de PLN en un dominio de creciente interés político y social. En conjunto, el proyecto busca avanzar en la comprensión de cómo el lenguaje revela afiliaciones ideológicas, y qué tan eficaces son las herramientas actuales para detectar esas señales en escenarios abiertos y ruidosos como YouTube.

Además, la visualización y el análisis del rendimiento alcanzado por los diferentes modelos permitirán identificar cuáles ofrecen mejores resultados, indagar en los factores que contribuyen a su eficacia y extraer conclusiones que orienten futuras líneas de trabajo. Esto incluye la posibilidad de seleccionar modelos especialmente prometedores, explorar ajustes arquitectónicos o de entrenamiento, incorporar nuevos datos que

enriquezcan el corpus existente o incluso evaluar arquitecturas que no fueron incluidas en esta primera fase experimental. De este modo, se abre la puerta a una profundización progresiva en el estudio de la ideología política en medios digitales, apoyada en evidencia empírica y en el conocimiento acumulado a lo largo de este trabajo.

3.2. Metodología

Para alcanzar los objetivos planteados, se ha seguido una metodología que combina la construcción de un corpus específico con la aplicación y evaluación de modelos de lenguaje avanzados. Este enfoque busca no solo explorar la capacidad de los modelos Transformer para detectar afiliaciones ideológicas implícitas en el lenguaje, sino también asegurar la validez empírica del experimento en un entorno realista y representativo como YouTube. La metodología abarca desde la recolección y anotación de datos hasta el diseño de las representaciones textuales, la configuración de los modelos y la evaluación sistemática de su rendimiento. Este procedimiento ayuda a garantizar la consistencia, la comparabilidad entre enfoques y la reproducibilidad del estudio.

3.2.1. Construcción y anotación del corpus

El primer paso metodológico consistió en la elaboración de un corpus adaptado a la tarea de detección de ideología política en lengua española.

El conjunto de datos empleado en este trabajo se compone de un total de 21 canales de YouTube en español, seleccionados manualmente con el objetivo de representar una diversidad ideológica y de formatos de comunicación política. Estos canales se han agrupado en cuatro categorías ideológicas: izquierda, izquierda moderada, derecha moderada y derecha. Esta clasificación refleja una segmentación comúnmente reconocida en el panorama político español actual [12] [21], en el que dichas etiquetas permiten distinguir entre posiciones progresistas, socialdemócratas, conservadoras o de derecha alternativa.

Además, se ha buscado un equilibrio entre distintos tipos de actores del ecosistema digital: medios de comunicación tradicionales con presencia en YouTube, partidos políticos con canales oficiales y creadores de contenido individuales, cuya presencia en el debate político es cada vez más significativa, y que emiten discursos desde perspectivas ideológicas claramente marcadas.

La asignación se ha realizado de forma manual, basándose en la afiliación partidaria explícita, el posicionamiento editorial de medios y el contenido discursivo dominante de los creadores individuales.

A continuación, se muestra un resumen de los canales utilizados, junto con su categorización ideológica:

Ideología	Canales
Izquierda	@ahora_podemos, @sumar_oficial, @eldiarioes, @publico_es, @canalredtv, @AlanBarrosoA
Izquierda moderada	@psoe, @LaVanguardia, @ElPlural_TV, @ondacero, @elpais
Derecha moderada	@partidopopular, @elmundo, @Elconfidencialtv, @LibertadDigital, @esRadiovideos
Derecha	voxespanatv, @Okdiariovideos, @juanrallo, @WallStreetWolverine, @ViOneMedia

Figura 2: Canales clasificados por ideología

Esta diversidad permite evaluar el rendimiento de los modelos de detección ideológica en un entorno representativo del espectro político español actual, que incluye tanto contenidos institucionales como discursos de carácter más informal o activista.

3.2.2. Preprocesamiento y representación textual

Una vez definidos los canales y sus etiquetas ideológicas, se procedió a la recopilación del contenido textual asociado a cada uno. Para ello, se utilizó la *YouTube Data API v3* [11] para recopilar metadatos de los vídeos y el paquete `youtube-transcript-api` [9] para extraer sus transcripciones automáticas. Estas herramientas facilitan la extracción sistemática de datos y garantizó la consistencia del corpus recopilado en términos de volumen y estructura.

A los textos extraídos se les aplicó un proceso de preprocesamiento orientado a mejorar la calidad y consistencia del corpus. Este proceso incluyó pasos como la eliminación de elementos no informativos, el filtrado de vídeos con transcripciones incompletas y la concatenación de los fragmentos en un único bloque de texto por canal. Este enfoque busca reflejar el estilo discursivo general de cada canal, en lugar de analizar vídeos de forma aislada.

El resultado de este procedimiento es un conjunto de documentos representativos de cada canal, limpios y listos para ser transformados en entradas vectorizadas por los modelos de lenguaje. Las decisiones sobre segmentación y representación contextualizada se detallan posteriormente en el capítulo de desarrollo.

3.2.3. Selección de modelos y configuración experimental

Tras tener preparado el corpus, fue necesario identificar cuáles serían los modelos que se evaluarán. Los modelos seleccionados son:

- **BETO** [4], un BERT monolingüe entrenado desde cero sobre grandes corpus en español.
- **RoBERTa-BNE** [8], una versión de RoBERTa entrenada por la Biblioteca Nacional de España sobre textos de alta calidad en español.

- **DistilBERT** [22], una variante compacta obtenida mediante distilación de BERT para mejorar eficiencia sin pérdida significativa de rendimiento.
- **mBERT** [7], el modelo multilingüe de Google, entrenado sobre más de 100 idiomas, incluido el español.

Todos los modelos se ajustaron mediante *fine-tuning* sobre el corpus anotado. Para ello, se emplearon técnicas estándar de aprendizaje supervisado, utilizando la clase ideológica como etiqueta de salida y aplicando funciones de pérdida categórica para optimizar la clasificación.

El entrenamiento se realizó utilizando la biblioteca Transformers de Hugging Face [24].

Durante el proceso de fine-tuning, todos los modelos fueron entrenados con los mismos hiperparámetros base: una tasa de aprendizaje de $2e{-}5$, un tamaño de lote de 16 y un máximo de cinco épocas. Estas configuraciones se seleccionaron por ser comunes en la literatura especializada y por ofrecer un buen compromiso entre eficiencia y precisión. Asimismo, permitieron garantizar que las comparaciones entre modelos no se vieran afectadas por diferencias en la configuración del entrenamiento.

Las métricas empleadas para evaluar el rendimiento incluyen la *accuracy* general, así como la *precision*, *recall* (sensibilidad) y *F1-score* por clase, calculadas mediante el informe de clasificación de `scikit-learn` [17]. La *accuracy* representa el porcentaje total de predicciones correctas, mientras que la *precision* indica qué proporción de las predicciones positivas realizadas para una clase fueron correctas. Por su parte, el *recall* mide la cantidad que el modelo es capaz de identificar de ese tipo. Finalmente, el *F1-score* combina precisión y cobertura en una sola medida armónica. Este conjunto de métricas permite valorar tanto la eficacia general del sistema como su comportamiento específico para cada orientación ideológica, proporcionando así una visión más detallada del desempeño del modelo. Además, se incluye la matriz de confusión como herramienta para el análisis cualitativo de los errores. Este conjunto de medidas permite valorar tanto la precisión global del sistema como su comportamiento con clases desbalanceadas.

4. Desarrollo y resultados experimentales

Este capítulo presenta el desarrollo práctico del sistema propuesto y los resultados obtenidos tras su evaluación empírica. A partir de la metodología descrita en el apartado anterior, se detallan las distintas fases del proceso de implementación: desde la construcción de representaciones textuales para los canales, pasando por el entrenamiento y ajuste de los modelos, hasta la obtención e interpretación de los resultados de clasificación ideológica.

4.1. Construcción del corpus

Esta sección detalla el proceso seguido para construir el corpus textual sobre el que se han entrenado y evaluado los modelos de detección ideológica. A partir del diseño metodológico descrito en el capítulo anterior, se implementó una estrategia práctica de recolección, limpieza, anotación y análisis de datos que dio lugar a un conjunto amplio y representativo de vídeos en español sobre temática política.

4.1.1. Selección de fuentes

La recolección de datos partió de una selección manual de **21 canales de YouTube** que generan contenido político o sociopolítico en español. Esta muestra se diseñó con el objetivo de cubrir diferentes posiciones ideológicas y tipos de emisores dentro del ecosistema comunicativo digital. Concretamente, se incluyeron:

- **Partidos políticos**, como `@psoe` o `voxespanatv`, con discursos institucionales.
- **Medios de comunicación tradicionales**, como `@elpais` o `@elmundo`, con enfoque editorial establecido.
- **Creadores de contenido independientes**, como `@WallStreetWolverine` o `@AlanBarrosoA`, que representan discursos personales más informales o activistas.

Cada canal fue etiquetado en una de las siguientes categorías ideológicas: *Izquierda*, *Izquierda moderada*, *Derecha moderada* o *Derecha*. Esta clasificación se basó en su afiliación partidaria, su línea editorial o el análisis cualitativo del contenido dominante. La distribución de los canales por clase se resume en la Figura 2 del capítulo anterior.

4.1.2. Extracción y estructuración de datos

La extracción de datos se realizó mediante scripts desarrollados en Python, que utilizaron la *YouTube Data API v3* [11] y el paquete `youtube-transcript-api` [9] para acceder a los metadatos que podrían resultar útiles a futuro y recuperar las **transcripciones automáticas generadas por la propia plataforma**. Para cada vídeo se almacenaron:

- ID del canal e ID del vídeo.
- Título, fecha de publicación y hashtags.

- Transcripción completa en español.
- Etiqueta ideológica, heredada del canal correspondiente.

Los vídeos recopilados abarcan un rango temporal que va desde el **6 de agosto de 2014** hasta el **4 de abril de 2025**. Esta amplia cobertura se debe a la disparidad en la frecuencia de publicación de los distintos canales, así como a la disponibilidad o no de transcripción para cada vídeo.

Durante este proceso, se descartaron centenares de vídeos por no contar con transcripción automática. Además, se controló la **longitud máxima de las transcripciones** (8.000 caracteres) para evitar un crecimiento desmedido del conjunto de datos. También se eliminaron anotaciones automáticas irrelevantes como *[Música]* o *[Aplausos]*, ya que no aportaban contenido útil para el análisis lingüístico.

4.1.3. Estadísticas del corpus final

Tras el filtrado, el corpus final contiene un total de **19.730 vídeos**. La siguiente tabla muestra la distribución por clase ideológica:

Ideología	**Número de vídeos**
Derecha	5.101
Derecha moderada	5.102
Izquierda moderada	5.093
Izquierda	4.434
Total	19.730

Figura 3: Distribución final del corpus por ideología

La distribución es razonablemente equilibrada, lo cual es crucial para evitar sesgos de entrenamiento en los modelos supervisados.

4.1.4. Ejemplos representativos

En la Figura 4 se recogen ejemplos reales de vídeos incluidos en el corpus. Se muestra el canal de origen, el título del vídeo y un fragmento de la transcripción asociada, representativos de cada una de las categorías ideológicas.

Ideología	Canal	Título del vídeo	Fragmento de transcripción
Derecha	`@Okdiariovideos`	"Zapatero y Montero blanquean a Chaves y a Griñán"	"...por lo que el partido socialista ha creído en la inocencia..."
Derecha moderada	`@Elconfidencialtv`	"El Gobierno estudia subir el IRPF a las rentas altas"	"...comprometidos a colaborar gobierno y unidos podemos estudian subir el IRPF..."
Izquierda moderada	`@psoe`	"Pedro Sánchez en Castellón"	"...porque andan un poco molestos estos de la de la derecha..."
Izquierda	`@sumar_oficial`	"lo que pasa cuando las normas que protegen a los trabajadores no se cumplen"	"...yo aconsejaría al dueño de rayanair que no insultara a los sindicatos..."

Figura 4: Ejemplos reales de transcripciones por ideología

4.1.5. Limitaciones del corpus

A pesar del esfuerzo por construir un corpus representativo y equilibrado, existen ciertas limitaciones que deben tenerse en cuenta:

- Las etiquetas ideológicas se asignan a nivel de canal, no de vídeo, lo cual puede ignorar la diversidad temática o tonal de ciertos contenidos.
- Las transcripciones automáticas pueden contener errores de segmentación, puntuación o interpretación de voz.
- Algunos canales han podido cambiar de línea editorial a lo largo del tiempo, introduciendo cierta variabilidad discursiva interna.

No obstante, el conjunto de datos resultante ofrece una base sólida para el entrenamiento y evaluación de modelos de clasificación ideológica en el contexto digital hispanohablante.

4.2. Entrenamiento de modelos

El objetivo principal del presente trabajo es evaluar la capacidad de distintos modelos de aprendizaje automático para detectar la ideología política de canales de YouTube a partir de sus transcripciones. Para ello, se entrenaron y compararon varios modelos sobre documentos representativos del discurso político de cada canal.

4.2.1. Formulación del problema

La tarea se plantea como un problema de clasificación multiclase supervisada, en el que cada ejemplo se asocia a una de las cuatro categorías ideológicas definidas: *Izquierda*, *Izquierda moderada*, *Derecha moderada* o *Derecha*. Se utilizó una división del corpus en conjuntos de entrenamiento (80 %) y prueba (20 %), asegurando una distribución proporcional de clases.

4.2.2. Agrupación de videos

Para representar mejor los discursos ideológicos sin depender exclusivamente de los canales individuales, se optó por agrupar vídeos de distintos canales en clusters que comparten la misma etiqueta ideológica.

Se siguieron dos estrategías para la creación de los clusters:

- **Agrupación balanceada por clústeres**: se intercalaron vídeos de distintos canales dentro de la misma ideología y se agruparon en clústeres de tamaño fijo (por ejemplo, 5 vídeos por clúster).
- **Agrupación por canal**: en otra variante experimental, se asignaron canales enteros a cada conjunto, impidiendo que los videos de un canal no puedan aparecen en prueba y test a la vez. Esta configuración evita que el modelo aprenda a distinguir ideologías basándose en señales específicas de un canal concreto, obligándolo a generalizar a partir del lenguaje común a cada corriente.

En ambos casos, las predicciones se realizaron de forma individual sobre cada transcripción que compone el clúster, y la etiqueta ideológica final se asignó mediante la moda de dichas predicciones. Esta estrategia de agregación permite suavizar posibles errores puntuales y aporta mayor robustez al proceso de clasificación a nivel de conjunto.

4.2.3. Construcción del input

Cada entrada del modelo se compone de un input construido a partir de los vídeos del conjunto de datos. Para cada vídeo, se extrajeron y concatenaron dos elementos: el título y la transcripción. Este texto combinado sirvió como unidad de entrada para los modelos de clasificación.

Inicialmente, se consideró incluir también la descripción del vídeo y los hashtags como parte del input. Sin embargo, durante el análisis preliminar se observó que estos campos presentaban un contenido redundante, poco informativo o repetido entre vídeos del mismo canal. En muchos casos, las descripciones no aportaban información nueva respecto al discurso, y los hashtags eran genéricos o irrelevantes desde el punto de vista ideológico. Por este motivo, se decidió omitirlos en la construcción final del conjunto de entrenamiento.

El texto resultante de cada entrada se procesó y tokenizó con el modelo correspondiente, respetando sus límites máximos de longitud (normalmente 512 tokens). En el caso de entradas más largas, se truncó el contenido priorizando el texto de los títulos.

4.2.4. Modelos evaluados

Se entrenaron y evaluaron distintos modelos de clasificación ideológica basados en arquitecturas de lenguaje preentrenadas, concretamente:

- **BETO (dccuchile/bert-base-spanish-wwm-cased)**: es la adaptación del modelo BERT al español, entrenado por la Universidad Católica de Chile sobre un corpus mixto de textos en español (Wikipedia, prensa, libros, etc.). Su entrenamiento incluye el uso de enmascaramiento de palabras completas (*whole word masking*), lo que lo hace especialmente eficaz para captar unidades léxicas completas en tareas de clasificación de texto. [4]
- **DistilBERT**: es una versión más ligera de BERT, entrenada mediante técnicas de distilación del conocimiento (*knowledge distillation*) para reducir el número de parámetros sin perder demasiada capacidad predictiva. Ofrece tiempos de entrenamiento más rápidos y menor uso de memoria, lo que lo convierte en una opción eficiente para experimentos exploratorios o entornos con recursos limitados [22].
- **BERT multilingüe (mBERT)**: se trata del modelo multilingüe original de Google, entrenado sobre Wikipedia en más de 100 idiomas simultáneamente. A pesar de no estar específicamente adaptado al español, su versatilidad permite evaluar su rendimiento como modelo base en tareas ideológicas, especialmente en comparación con versiones monolingües como BETO [7].
- **RoBERTa-BNE**: desarrollado por la Biblioteca Nacional de España en colaboración con la Universitat Politècnica de Catalunya, este modelo parte de la arquitectura RoBERTa y ha sido entrenado sobre un gran corpus de prensa y documentos institucionales en español. Su exposición a textos formales y editoriales lo hace especialmente adecuado para tareas de clasificación política o mediática [8].

Todos ellos fueron implementados utilizando la librería `Transformers` de Hugging Face [24].

Estos modelos se compararon en igualdad de condiciones, evaluando su capacidad para clasificar correctamente los documentos ideológicos generados a partir de los vídeos del corpus. Los resultados obtenidos se presentan en la siguiente sección.

4.3. Análisis de resultados

En esta sección se analizan los resultados obtenidos tras la evaluación de los modelos de clasificación ideológica entrenados sobre el corpus descrito. Se examina el rendimiento de las distintas arquitecturas considerando dos configuraciones del dataset: por un lado, la agrupación balanceada de vídeos en clústeres sin restricción por canal, y por otro, la partición estricta por canal, que impide la presencia simultánea de un mismo canal en los conjuntos de entrenamiento y prueba.

El objetivo es doble: por un lado, valorar la capacidad predictiva bruta de los modelos en un entorno favorable; y por otro, evaluar su habilidad para generalizar

a discursos ideológicos provenientes de emisores no vistos durante el entrenamiento. El análisis incluye métricas cuantitativas (precisión, recall, F1, accuracy) así como un estudio cualitativo de los patrones de confusión entre clases.

A continuación, se presentan los resultados de cada configuración, ordenados según la estrategia de clusters utilizada.

4.3.1. Evaluación con agrupación balanceada de clústeres

Una vez entrenados los distintos modelos sobre el conjunto de datos con una distribución balanceada de clústers ideológicos, se procedió a su evaluación utilizando particiones aleatorias del corpus, sin separación explícita por canal. Esta configuración busca analizar el rendimiento puro de los modelos sin introducir restricciones adicionales de generalización.

A continuación, se presentan los resultados obtenidos para cada uno de los modelos evaluados: `BETO` [4], `DistilBETO` [22], `mBERT` [7] y `RoBERTa-BNE` [8].

Modelo	**Tamaño clúster**	**Accuracy**	**Macro F1**
BETO	5	0.98	0.98
	10	1.00	1.00
	20	1.00	1.00
DistilBETO	5	0.98	0.98
	10	1.00	1.00
	20	1.00	1.00
mBERT	5	0.95	0.95
	10	0.99	1.00
	20	1.00	1.00
RoBERTa-BNE	5	0.99	0.99
	10	1.00	1.00
	20	1.00	1.00

Figura 5: Exactitud y macro F1-score para cada modelo según tamaño de clúster (agrupación balanceada)

La Tabla 5 recoge los valores de **accuracy** y **macro F1-score** obtenidos por los distintos modelos para tamaños de clúster 5, 10 y 20. Los resultados muestran un rendimiento excepcionalmente alto en todos los casos, con valores cercanos al 1.00 en casi todos los experimentos.

Particularmente, los modelos **RoBERTa-BNE**, **BETO** y **DistilBETO** alcanzan un *accuracy* y *macro F1* perfectos (1.00) cuando el tamaño del clúster es 10 o 20, mientras que con clústeres de tamaño 5 mantienen también un rendimiento sobresaliente, con valores por encima del 0.98. El modelo **mBERT**, si bien muestra una leve caída en el clúster de tamaño 5 (0.95), se comporta igualmente de forma notable en los tamaños superiores.

En lo que respecta a las matrices de confusión muestran un comportamiento prácticamente perfecto para tamaños de clúster reducidos, con una separación clara entre clases ideológicas y errores mínimos. A modo ilustrativo, la Figura 6 presenta la matriz correspondiente al modelo `bert-base-multilingual-cased` con clústeres de tamaño 10:

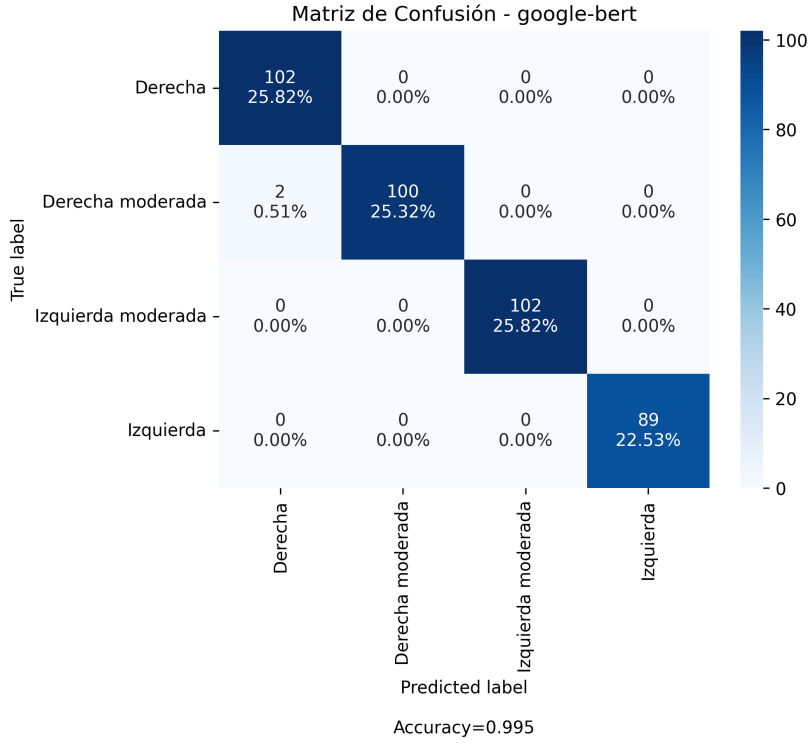

Figura 6: Matriz de confusión – Google BERT (clúster tamaño 10)

En este caso, únicamente se observa un error puntual en la clase *derecha moderada*, mientras que el resto de predicciones son correctas. Este patrón se repite en todos los modelos, con confusiones principalmente entre clases ideológicas contiguas, como *derecha* y *derecha moderada*, o entre *izquierda moderada* e *izquierda*, lo cual es esperable dada la proximidad discursiva entre dichas categorías.

Para clústeres de mayor tamaño, como se aprecia en la Figura 7, se incrementan ligeramente las confusiones, aunque la estructura general se mantiene:

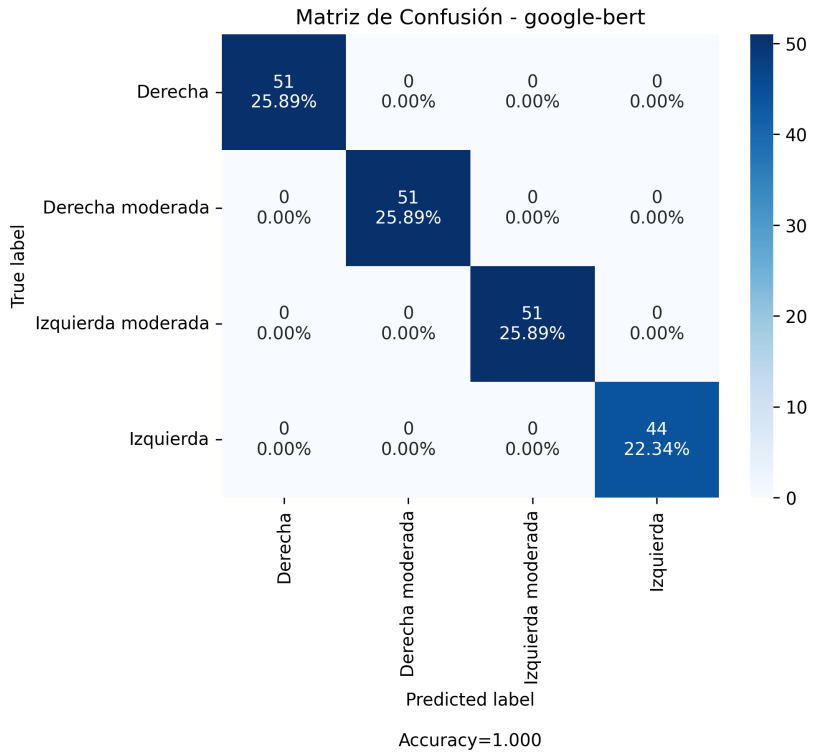

Figura 7: Matriz de confusión – Google BERT (clúster tamaño 20)

En este caso, aparecen errores dispersos en todas las clases, pero el modelo sigue siendo capaz de identificar correctamente la mayoría de ejemplos. Esto indica que, aunque el aumento del contenido introduce cierta ambigüedad o ruido, los modelos conservan una capacidad elevada de detección ideológica.

Si bien los resultados obtenidos con esta configuración muestran un rendimiento elevado en todos los modelos evaluados, es importante matizar estas cifras. Dado que los conjuntos de entrenamiento y prueba comparten vídeos de un mismo canal, es posible que los modelos estén captando patrones específicos del estilo discursivo, léxico o temático propios de ciertos emisores. Esto podría inflar artificialmente las métricas, al facilitar la identificación de la clase ideológica sin necesidad de generalizar más allá de las señales características de los canales.

Por tanto, aunque esta configuración sirve como una primera aproximación al rendimiento de los modelos, es necesario someterlos a una evaluación más exigente que mida su capacidad real de generalización. Para ello, se adopta una estrategia alternativa: separar completamente los canales entre los conjuntos de entrenamiento y prueba. De este modo, se obliga al sistema a predecir la ideología de emisores nunca vistos durante el entrenamiento, basándose únicamente en rasgos discursivos comunes a cada corriente ideológica.

En consecuencia, aunque estos resultados son prometedores, se hace necesario contrastarlos con otros escenarios experimentales más exigentes —como veremos en el siguiente apartado— para poder extraer conclusiones robustas sobre la generalización real de los modelos.

4.3.2. Evaluación con separación por canal

Con el fin de evaluar la capacidad de los modelos para generalizar más allá de los patrones específicos de cada canal, se adoptó una configuración experimental en la que los conjuntos de entrenamiento y prueba se construyen con canales completamente disjuntos. Esto significa que los modelos deben inferir la ideología de emisores no vistos previamente, basándose exclusivamente en regularidades discursivas propias de cada corriente ideológica.

En esta variante, se mantuvo el tamaño de clúster fijo en 20, dada la estabilidad observada en los experimentos anteriores. La Figura 8 muestra los resultados obtenidos por los cuatro modelos bajo esta configuración:

Modelo	**Accuracy**	**Macro F1**
BETO	0.09	0.13
DistilBETO	0.14	0.10
mBERT	0.08	0.09
RoBERTa-BNE	0.11	0.13

Figura 8: Rendimiento de los modelos con separación por canal

Como se observa en la Figura 8, los resultados presentan una caída pronunciada respecto a la evaluación previa. El modelo que alcanza mayor exactitud es `DistilBETO`, con un *accuracy* del 14 %, seguido de `RoBERTa-BNE` con un 11 %. Todos los modelos quedan por debajo del umbral del azar (25 % en una clasificación multiclase con cuatro clases), lo que sugiere que las predicciones no son mejores que una asignación aleatoria.

Los valores de *macro F1-score* refuerzan esta conclusión, situándose entre 0.09 y 0.13, lo que evidencia una escasa capacidad para capturar patrones generalizables más allá de los canales concretos del entrenamiento. Esta caída tan acusada pone de manifiesto que los modelos dependen en gran medida de características específicas del canal —como su estilo lingüístico o temáticas recurrentes— para inferir la ideología, en lugar de apoyarse en rasgos más generales del discurso político.

Este patrón de errores se puede observar con mayor claridad en las matrices de confusión. A modo ilustrativo, se presenta en la Figura 9 la matriz correspondiente al modelo `RoBERTa-BNE`:

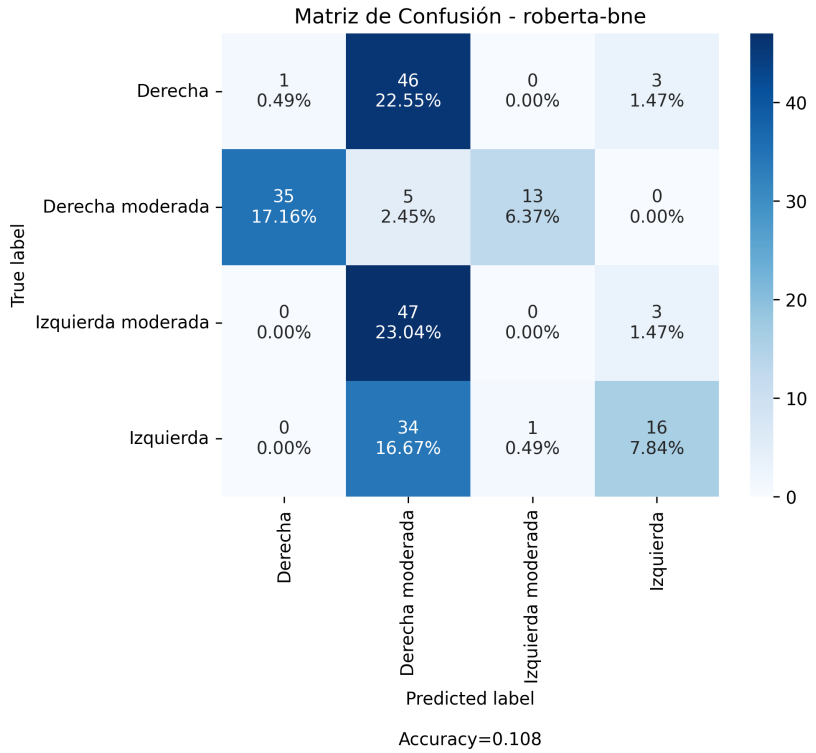

Figura 9: Matriz de confusión – RoBERTa-BNE con separación estricta por canal

Como se observa en la Figura 9, las predicciones muestran una fuerte tendencia a concentrarse en una única clase, en este caso *derecha moderada*, independientemente del valor real. Este patrón, que se repite en otros modelos, pone de manifiesto la incapacidad del sistema para diferenciar adecuadamente entre corrientes ideológicas cuando los emisores del conjunto de prueba no han sido vistos durante el entrenamiento.

Este tipo de comportamiento sugiere que el modelo no ha logrado capturar patrones ideológicos generalizables, sino que ha dependido de señales contextuales o estilísticas muy específicas de los canales presentes durante el entrenamiento. Al enfrentarse a nuevos emisores, esta dependencia le impide trasladar correctamente las representaciones aprendidas a discursos desconocidos, llevando a predicciones sesgadas hacia ciertas clases dominantes.

El fenómeno se acentúa por el hecho de que, en ausencia de ejemplos diversos para cada ideología en el conjunto de entrenamiento, el modelo podría haber internalizado asociaciones espurias entre ciertos estilos de comunicación y etiquetas ideológicas concretas. Esto resulta en una clasificación incorrecta de discursos ideológicamente variados bajo una categoría común, minando la utilidad práctica del sistema en escenarios reales.

Estos resultados remarcan la necesidad de validar los modelos en contextos donde se minimice el solapamiento entre emisores conocidos y desconocidos. Solo así es posible evaluar de forma fiable la robustez del sistema y su capacidad para extrapolar más allá de los datos sobre los que fue entrenado.

Con este objetivo, el siguiente apartado se centra en una evaluación adicional sobre

un corpus externo e independiente: **PoliticES**, un conjunto de clústeres de tweets en español agrupados según características ideológicas, presentado en la tarea IberLEF 2023 [10]. Esta validación permite comprobar que los modelos están correctamente configurados y operan de forma fiable en un entorno controlado y bien definido. Al utilizar un recurso anotado manualmente y ampliamente utilizado en la comunidad, se busca confirmar la integridad funcional de los sistemas antes de abordar análisis más complejos o extraer conclusiones generales sobre su rendimiento.

Sólo si los modelos demuestran un comportamiento coherente y competitivo sobre este corpus independiente, podrá considerarse que las deficiencias observadas anteriormente no se deben a fallos en su implementación. En caso contrario, cualquier análisis posterior del dataset original carecería de fundamento sólido. Por tanto, se pospone el estudio estructural del corpus de entrenamiento hasta haber verificado la validez de los modelos empleados.

4.3.3. Evaluación con corpus externo: *PoliticES*

Con el objetivo de confirmar que los modelos están adecuadamente configurados y su comportamiento no se debe a errores de implementación, se realizó una evaluación sobre un corpus: **PoliticES**, presentado en la tarea IberLEF 2023 [10]. Este recurso consiste en clústeres de tweets en español etiquetados según ideología política (izquierda, izquierda moderada, derecha moderada y derecha), y ha sido utilizado ampliamente como referencia en tareas de clasificación ideológica.

Al provenir de un dominio distinto —Twitter, en lugar de transcripciones de vídeo— y haber sido construido con una estrategia de agrupación diferente, PoliticES permite comprobar la validez funcional de los modelos fuera del entorno original de entrenamiento. Aquí no se busca una transferencia directa de conocimiento sobre el contenido, sino una verificación de que las arquitecturas y sus configuraciones son capaces de aprender patrones ideológicos generales cuando se les proporciona un corpus consistente y anotado manualmente.

Los resultados obtenidos con el modelo `RoBERTa-BNE` se muestran en la Figura 10:

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
Derecha	0.93	0.37	0.53	67
Derecha moderada	0.69	0.70	0.70	153
Izquierda moderada	0.63	0.92	0.75	210
Izquierda	0.90	0.44	0.59	117
Accuracy		0.69		
Macro avg	0.79	0.61	0.64	547
Weighted avg	0.74	0.69	0.67	547

Figura 10: Reporte de clasificación para RoBERTa-BNE sobre el corpus PoliticES

Como se puede observar, el modelo alcanza una **accuracy**, con un **macro F1-score de 0.64**, lo que representa un rendimiento a la par con los obtenidos en la tarea.

Además, se aprecia una capacidad razonable de distinguir entre clases, especialmente en el caso de las etiquetas *izquierda moderada* y *derecha moderada*, lo que sugiere que el modelo responde bien cuando se enfrenta a un corpus coherente y correctamente anotado.

La matriz de confusión correspondiente, mostrada en la Figura 11, confirma esta observación:

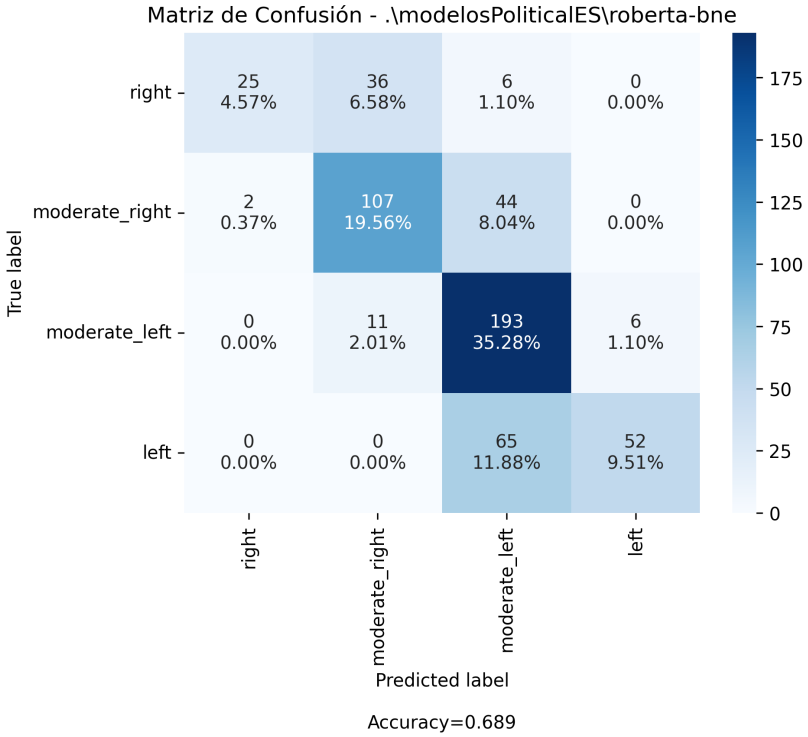

Figura 11: Matriz de confusión – RoBERTa-BNE sobre PoliticES

Si bien existe cierta confusión entre clases ideológicamente cercanas, las predicciones del modelo son mucho más balanceadas que en el caso anterior. Esto refuerza la idea de que los problemas de generalización identificados previamente no se deben a una configuración deficiente del sistema, sino posiblemente a deficiencias estructurales del corpus de entrenamiento original.

Como prueba adicional, se evaluó el comportamiento del modelo entrenado sobre *PoliticES* al aplicarse directamente sobre el corpus de YouTube, invirtiendo el enfoque. Los resultados obtenidos se presentan a continuación:

Clase	Precisión	Recall	F1-score	Soporte
Derecha	1.00	0.07	0.12	255
Derecha moderada	0.54	0.32	0.40	255
Izquierda moderada	0.35	0.98	0.51	255
Izquierda	0.99	0.44	0.61	222
Accuracy		0.45		
Macro avg	0.72	0.45	0.41	987
Weighted avg	0.71	0.45	0.40	987

Figura 12: Evaluación del modelo entrenado con *PoliticES* sobre el corpus de YouTube

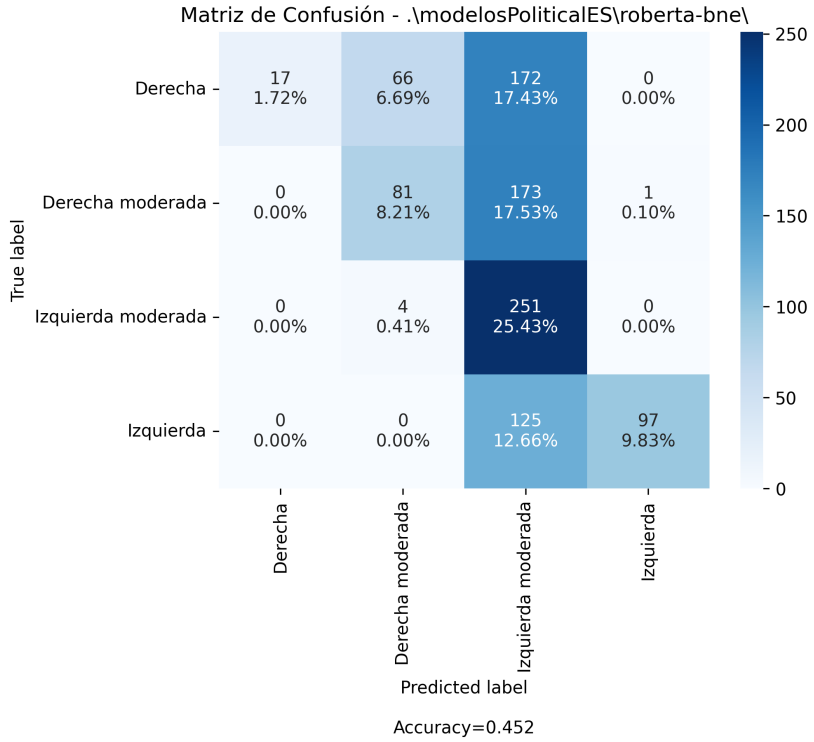

Figura 13: Matriz de confusión – modelo entrenado con *PoliticES*, evaluado sobre YouTube

Aunque el rendimiento disminuye notablemente (accuracy del 45 %), los valores siguen por encima del azar, lo que indica cierta transferencia parcial del conocimiento aprendido. Se observan aciertos notables en la clase *izquierda moderada* y una mayor capacidad para evitar la predicción única en una sola clase, como ocurría en los experimentos anteriores.

En conjunto, esta sección confirma que las arquitecturas y configuraciones utilizadas son funcionales. La baja capacidad de generalización observada en apartados anteriores usando el corpus de YouTube no parece deberse a deficiencias técnicas de los modelos, sino más bien a limitaciones propias del corpus. En consecuencia, el foco

debe desplazarse hacia un análisis crítico del conjunto de datos original para identificar posibles sesgos o problemas estructurales que comprometan su utilidad en tareas reales de clasificación ideológica.

4.4. Análisis cualitativo del corpus de entrenamiento

Tras evaluar el rendimiento de los modelos con diversas particiones y conjuntos de prueba, surge la necesidad de examinar más a fondo el contenido del corpus original. Con el objetivo de identificar posibles causas de los fallos observados, se realizó un análisis cualitativo desde dos enfoques complementarios: la distribución léxica por clase y una evaluación manual de vídeos específicos.

Distribución léxica por clase. Se calculó la frecuencia de aparición de palabras en los vídeos por clase ideológica, tras filtrar las *stopwords* en español utilizando `nltk` [2]. Los conteos se generaron con `CountVectorizer`, y las palabras fueron ordenadas según su puntuación en `mutual_info_classif` [17], como estimación de su capacidad discriminativa entre clases. A pesar del filtrado, persisten algunos términos con baja utilidad práctica, como *si* o *acompañaré*. No obstante, los valores más altos del ranking sí permiten identificar patrones léxicos ideológicamente marcados. Un detalle relevante es que la gran mayoría de las palabras con mayor puntuación pertenecen a las clases extremas; solo unas 20 de las 200 primeras están dominadas por clases moderadas, lo que sugiere una asimetría en la carga semántica de los discursos según el espectro político.

La Figura 14 muestra las 20 palabras más frecuentes en todo el corpus, junto con su distribución por ideología. Como puede verse, la mayoría de estos términos aparecen de forma recurrente en todas las clases, lo que dificulta su utilidad para la clasificación.

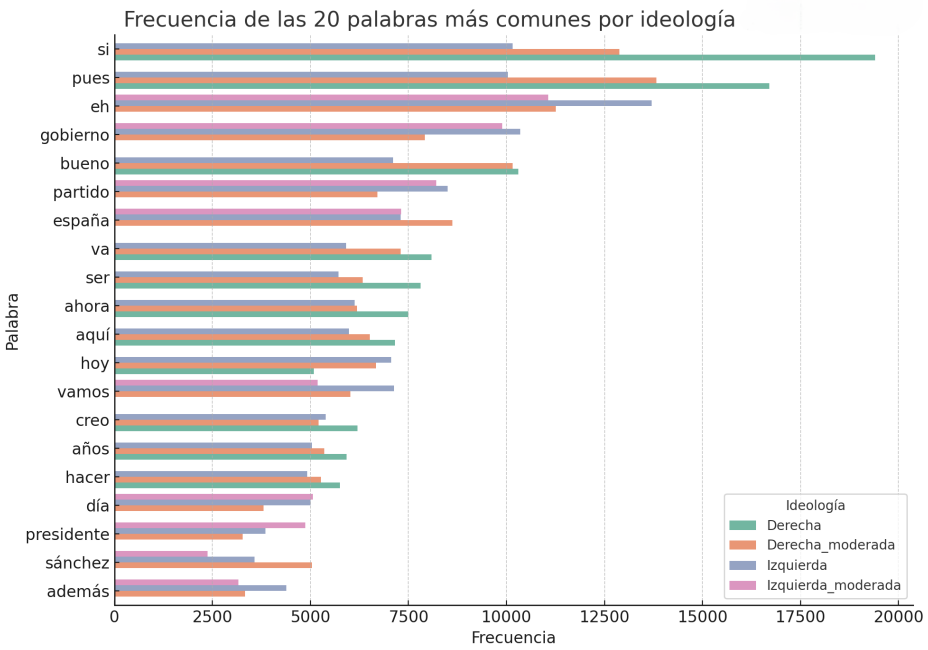

Figura 14: Frecuencia de las 20 palabras más comunes por ideología.

Por otro lado, la Figura 15 recoge las 20 palabras con mayor puntuación de discriminación (*score*), independientemente de su frecuencia absoluta. Esta selección revela términos que reflejan con mayor claridad una alineación ideológica específica, como *derechos*, *palestina* o *podemos* en el caso de la izquierda, y *vox* en el de la derecha. Aun así, se observa presencia cruzada de estos términos en clases opuestas, como ocurre con *vox*, que aparece tanto en discursos de derecha como en su crítica desde la izquierda, lo que genera ambigüedad si no se tiene en cuenta el contexto semántico.

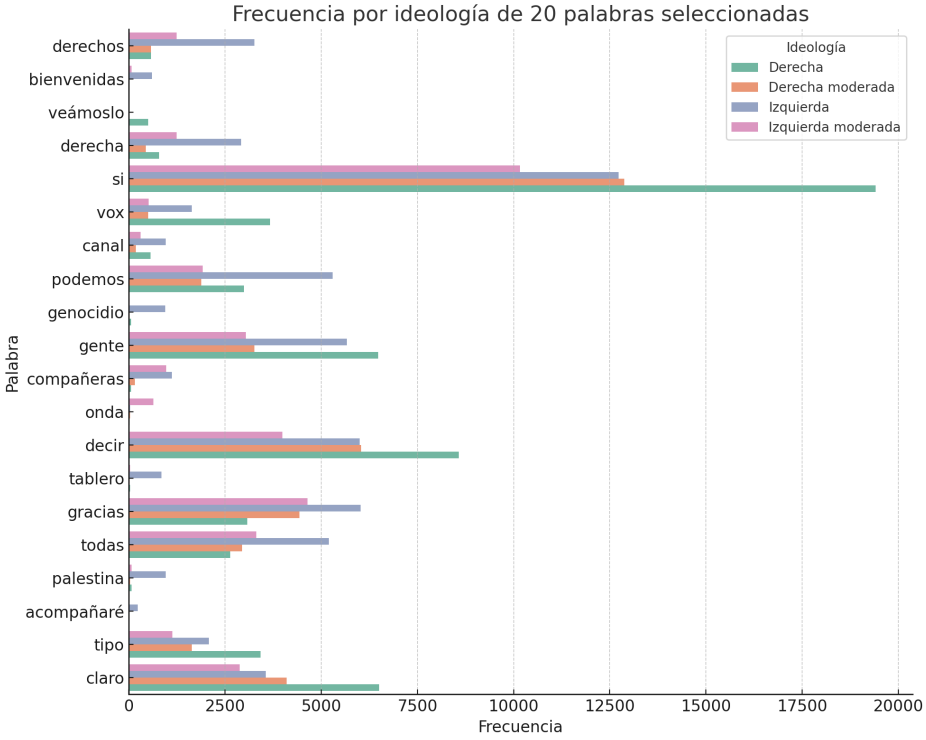

Figura 15: Frecuencia por ideología de las 20 palabras con mayor puntuación de discriminación.

Evaluación manual de vídeos. También se intentó determinar manualmente la ideología de una pequeña muestra de vídeos, con base en su contenido textual. En varios casos, se logró inferir la orientación política de forma razonable, aunque en ocasiones las diferencias eran sutiles o ambiguas, por ejemplo, cuando se trataban temas compartidos entre diferentes partidos. Se observaron vídeos donde medios de una ideología reproducen íntegramente declaraciones de figuras de signo contrario, sin aportar opinión editorial, lo que dificulta asociar el contenido a una clase concreta. Asimismo, se detectaron vídeos con escasa carga política —como noticias institucionales sobre los reyes, entrevistas centradas en aspectos personales del invitado o eventos como el Benidorm Fest—, donde resulta difícil extraer un tinte ideológico claro. Este tipo de ejemplos refuerzan la idea de que la dificultad de clasificación no radica únicamente en las limitaciones de los modelos, sino también en la propia naturaleza del corpus, que contiene elementos neutros, ambiguos o ajenos al debate político.

En conjunto, estos hallazgos apuntan a la existencia de ambigüedades semánticas, sesgos editoriales cruzados y una cierta laxitud en la definición de los contenidos por clase, lo que complica la tarea de clasificación. Esto refuerza la hipótesis de que parte de los problemas detectados en los modelos no se deben únicamente a su capacidad, sino también a la calidad, consistencia y definición del corpus de entrenamiento.

5. Conclusiones y vías futuro

Este trabajo ha explorado la viabilidad de detectar la ideología política de emisores digitales a partir del análisis automático de transcripciones de vídeos de YouTube. Para ello, se ha construido un corpus original, se han aplicado diversos modelos de lenguaje preentrenados, y se han diseñado diferentes configuraciones experimentales con el objetivo de evaluar tanto el rendimiento bruto como la capacidad de generalización de los sistemas.

Los resultados obtenidos muestran que los modelos evaluados —como BETO, RoBERTa-BNE o DistilBERT— son capaces de alcanzar un rendimiento excelente cuando el entorno de evaluación no impone restricciones de generalización, con valores de precisión y F1-score superiores al 0.98. Sin embargo, cuando se introduce una separación estricta por canal, la capacidad de los modelos se reduce drásticamente, situándose incluso por debajo del azar. Esto sugiere que gran parte del rendimiento observado en los escenarios más favorables se debe a la presencia de señales específicas de canal (estilo, términos recurrentes, tono) más que a la captación de patrones ideológicos generales.

La validación cruzada con el corpus externo *PoliticES* permite descartar errores de implementación o problemas estructurales en los modelos, y refuerza la hipótesis de que las limitaciones encontradas se deben, en gran medida, a la propia naturaleza del corpus original. A pesar del esfuerzo por construir un conjunto de datos balanceado y representativo, el hecho de asignar etiquetas ideológicas a nivel de canal —en lugar de vídeo—, junto con la presencia de contenido neutro, ambiguo o ajeno a la política, complica la tarea de clasificación automática.

El análisis cualitativo ha puesto de manifiesto varias fuentes de ruido y ambigüedad: discursos cruzados, falta de marcadores ideológicos explícitos, fragmentos informativos neutros y errores en las transcripciones automáticas. También se ha observado una distribución asimétrica del vocabulario discriminativo, más concentrado en las clases ideológicas extremas que en las moderadas, lo que introduce un sesgo adicional en el aprendizaje.

En conjunto, los resultados indican que, si bien los modelos de lenguaje actuales poseen una alta capacidad técnica para la clasificación ideológica, su rendimiento práctico depende críticamente de la calidad y la definición del corpus. Es por tanto fundamental prestar especial atención al diseño y anotación de los conjuntos de datos, especialmente en tareas sensibles como la detección de ideología política.

Por último, a la luz de las limitaciones detectadas durante el desarrollo del trabajo, se abren líneas claras para futuras mejoras. Una de ellas es la posibilidad de refinar el etiquetado, asignando la ideología no al canal completo, sino a cada vídeo de forma individual, lo cual permitiría capturar mejor la variabilidad interna de los emisores y reducir la ambigüedad en los datos. Complementariamente, sería deseable incorporar un proceso de validación humana sobre una muestra representativa del corpus, con el fin de asegurar la coherencia y calidad de las etiquetas utilizadas. Estos avances contribuirían a reforzar la solidez metodológica del sistema y a sentar las bases para aplicaciones más precisas en escenarios reales.

Bibliografía

[1] Pablo Barberá. «Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data». En: *Political Analysis* 23.1 (2015), págs. 76-91. URL: https://doi.org/10.1093/pan/mpu011.

[2] Steven Bird, Ewan Klein y Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009. URL: https://www.nltk.org/book/.

[3] Tom B. Brown et al. «Language Models are Few-Shot Learners». En: *Advances in Neural Information Processing Systems* 33 (2020), págs. 1877-1901. URL: https://arxiv.org/abs/2005.14165.

[4] José Cañete et al. «Spanish Pre-Trained BERT Model and Evaluation Data». En: *PML4DC at ICLR 2020*. 2020. URL: https://arxiv.org/abs/2308.02976.

[5] Kyunghyun Cho et al. «Learning phrase representations using RNN encoder-decoder for statistical machine translation». En: *arXiv preprint arXiv:1406.1078* (2014). URL: https://arxiv.org/abs/1406.1078.

[6] Michael D Conover et al. «Political polarization on Twitter». En: *ICWSM* (2011), págs. 89-96. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847.

[7] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *arXiv preprint arXiv:1810.04805* (2018). URL: https://arxiv.org/abs/1810.04805.

[8] Asier Gutiérrez Fandiño et al. «MarIA: Spanish Language Models». En: *Procesamiento del Lenguaje Natural* 68 (2022). ISSN: 1135-5948. DOI: 10.26342/2022-68-3. URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley.

[9] Johannes Filter. *youtube-transcript-api*. Python package for retrieving YouTube video transcripts. 2024. URL: https://github.com/jdepoix/youtube-transcript-api.

[10] José Antonio García-Díaz et al. *Overview of PoliticES at IberLEF 2023: Political Ideology Detection in Spanish Texts*. Sep. de 2023.

[11] Google Developers. *YouTube Data API v3*. https://developers.google.com/youtube/v3. https://developers.google.com/youtube/v3. 2023.

[12] Frederic Guerrero-Solé y Clara Virós i Martín. *Populismo de extrema derecha y redes sociales en España*. Communication Reports. Universitat Pompeu Fabra, jun. de 2023. URL: https://repositori.upf.edu/bitstream/handle/10230/57425/GuerreroSole_cr_popul.pdf.

[13] Sepp Hochreiter y Jürgen Schmidhuber. «Long short-term memory». En: *Neural computation* 9.8 (1997), págs. 1735-1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[14] Mohit Iyyer et al. «Political ideology detection using recursive neural networks». En: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, págs. 1113-1122. URL: https://aclanthology.org/P14-1105.

[15] Daniel Jurafsky y James H Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2000. URL: https://web.stanford.edu/~jurafsky/slp3/.

[16] Tomas Mikolov et al. «Efficient estimation of word representations in vector space». En: *arXiv preprint arXiv:1301.3781* (2013). URL: https://arxiv.org/abs/1301.3781.

[17] F. et al. Pedregosa. *Scikit-learn: Machine Learning in Python*. 2011. URL: https://jmlr.org/papers/v12/pedregosa11a.html.

[18] Jeffrey Pennington, Richard Socher y Christopher D Manning. «GloVe: Global Vectors for Word Representation». En: *EMNLP*. 2014, págs. 1532-1543. URL: https://aclanthology.org/D14-1162.

[19] Daniel Preoțiuc-Pietro et al. «Beyond binary labels: Political ideology prediction of Twitter users». En: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (2017), págs. 729-740. URL: https://aclanthology.org/P17-1068.

[20] Colin Raffel et al. «Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer». En: *Journal of Machine Learning Research* 21.140 (2020), págs. 1-67. URL: https://arxiv.org/abs/1910.10683.

[21] Luis Ramiro y Raul Gomez. «Radical-left populism during the Great Recession: Podemos and its competition with the established radical left». En: *Political Studies* 65 (2016). URL: https://journals.sagepub.com/doi/10.1177/0032321716647400.

[22] Victor Sanh et al. «DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter». En: *ArXiv* (2019). URL: https://arxiv.org/abs/1910.01108.

[23] Ashish Vaswani et al. «Attention is all you need». En: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. URL: https://arxiv.org/abs/1706.03762.

[24] Thomas Wolf et al. «Transformers: State-of-the-Art Natural Language Processing». En: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, págs. 38-45. URL: https://aclanthology.org/2020.emnlp-demos.6.

Anexos

Anexo 1 - Resultados adicionales

A continuación se muestran una serie de figuras generadas durante el proceso experimental. Aunque no fueron incluidas en el cuerpo principal del documento por cuestiones de espacio, complementan el análisis realizado y pueden resultar útiles para una interpretación más detallada de los resultados obtenidos.

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
	Tamaño de clúster: 5			
Derecha	0.97	1.00	0.99	205
Derecha moderada	0.97	0.95	0.96	205
Izquierda moderada	0.98	0.98	0.98	204
Izquierda	0.99	0.98	0.99	178
	Tamaño de clúster: 10			
Derecha	1.00	1.00	1.00	102
Derecha moderada	1.00	0.99	1.00	102
Izquierda moderada	1.00	1.00	1.00	102
Izquierda	0.99	1.00	0.99	89
	Tamaño de clúster: 20			
Derecha	1.00	1.00	1.00	51
Derecha moderada	1.00	1.00	1.00	51
Izquierda moderada	1.00	1.00	1.00	51
Izquierda	1.00	1.00	1.00	44

Figura 16: Reporte de clasificación para BETO

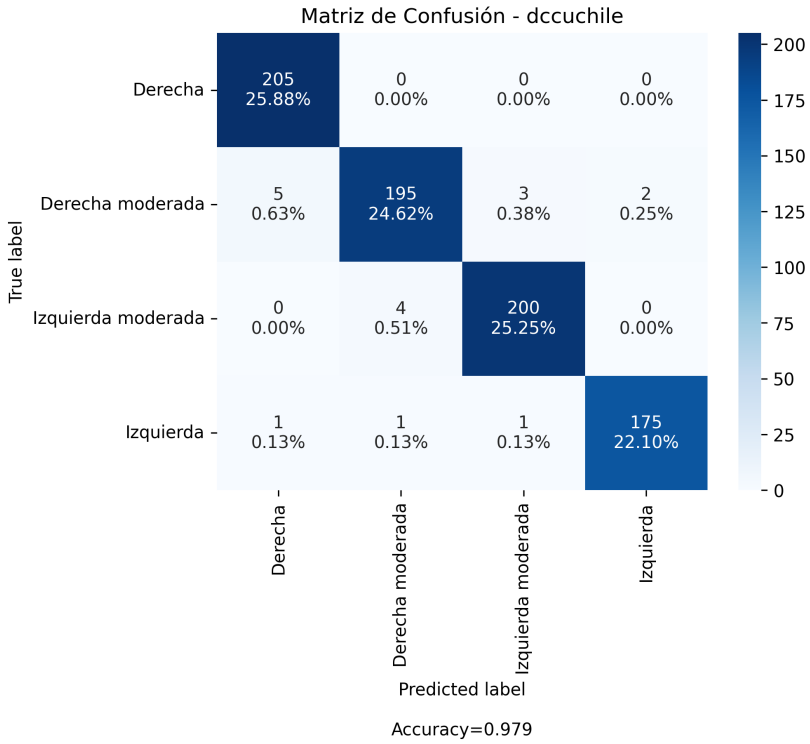

Figura 17: Matriz de confusión para BETO (clúster tamaño 5)

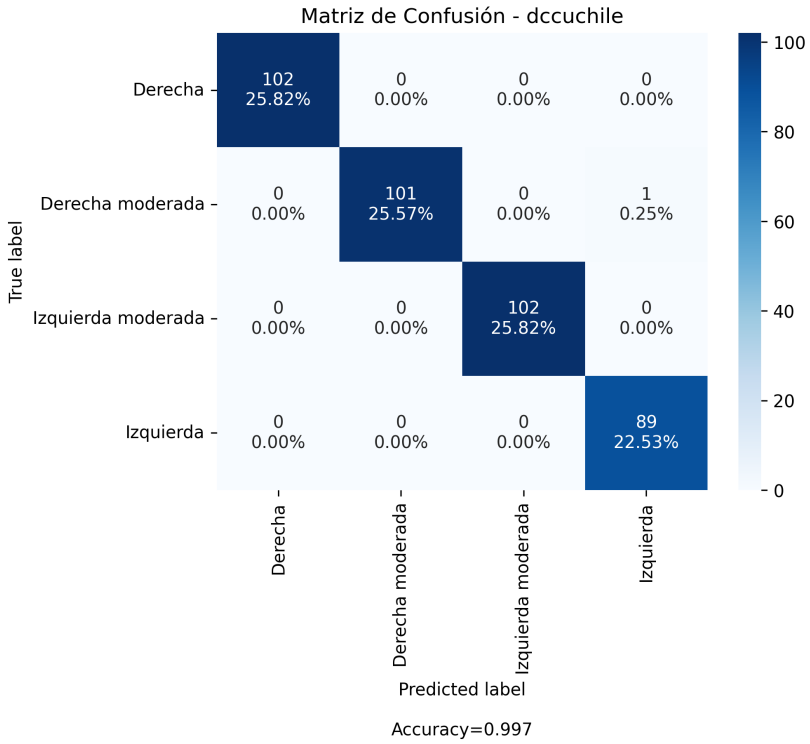

Figura 18: Matriz de confusión para BETO (clúster tamaño 10)

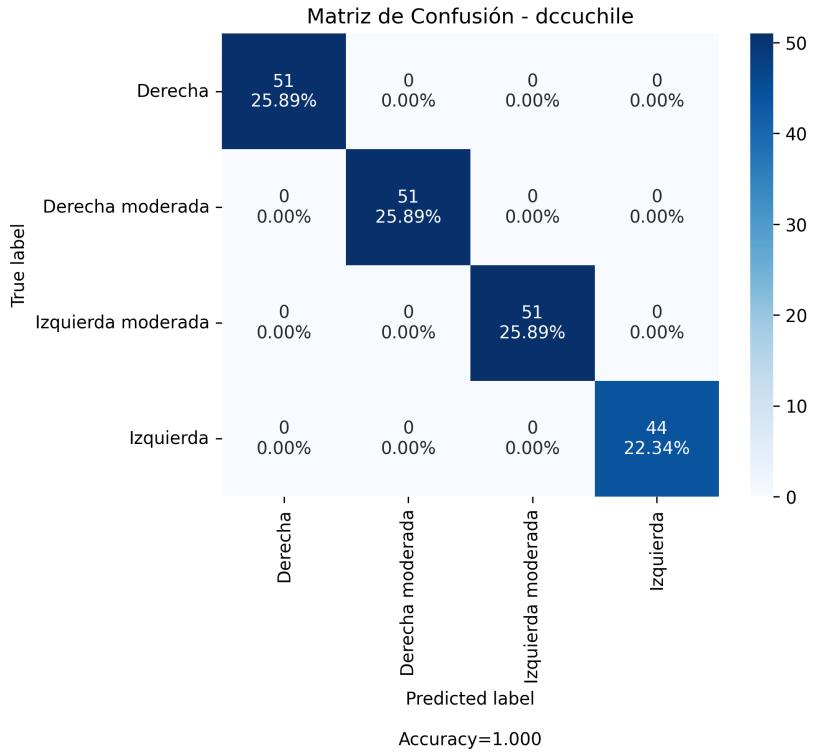

Figura 19: Matriz de confusión para BETO (clúster tamaño 20)

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
	Tamaño de clúster: 5			
Derecha	0.95	1.00	0.97	205
Derecha moderada	0.98	0.95	0.97	205
Izquierda moderada	0.99	0.98	0.98	204
Izquierda	0.99	0.98	0.99	178
	Tamaño de clúster: 10			
Derecha	0.99	1.00	1.00	102
Derecha moderada	1.00	0.99	1.00	102
Izquierda moderada	1.00	1.00	1.00	102
Izquierda	1.00	1.00	1.00	89
	Tamaño de clúster: 20			
Derecha	1.00	1.00	1.00	51
Derecha moderada	1.00	1.00	1.00	51
Izquierda moderada	1.00	1.00	1.00	51
Izquierda	1.00	1.00	1.00	44

Figura 20: Reporte de clasificación para DistilBERT

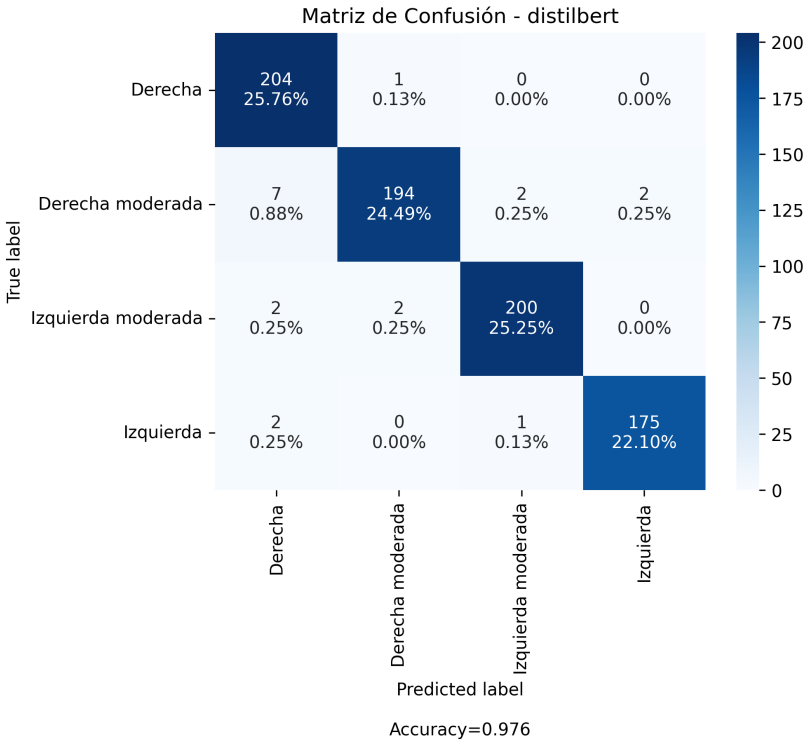

Figura 21: Matriz de confusión para DistilBERT (clúster tamaño 5)

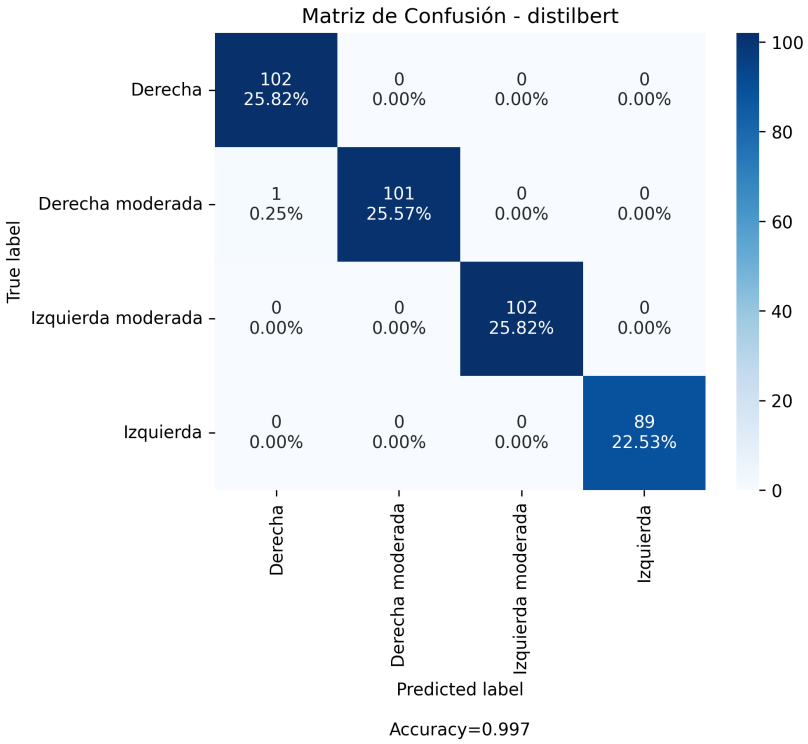

Figura 22: Matriz de confusión para DistilBERT (clúster tamaño 10)

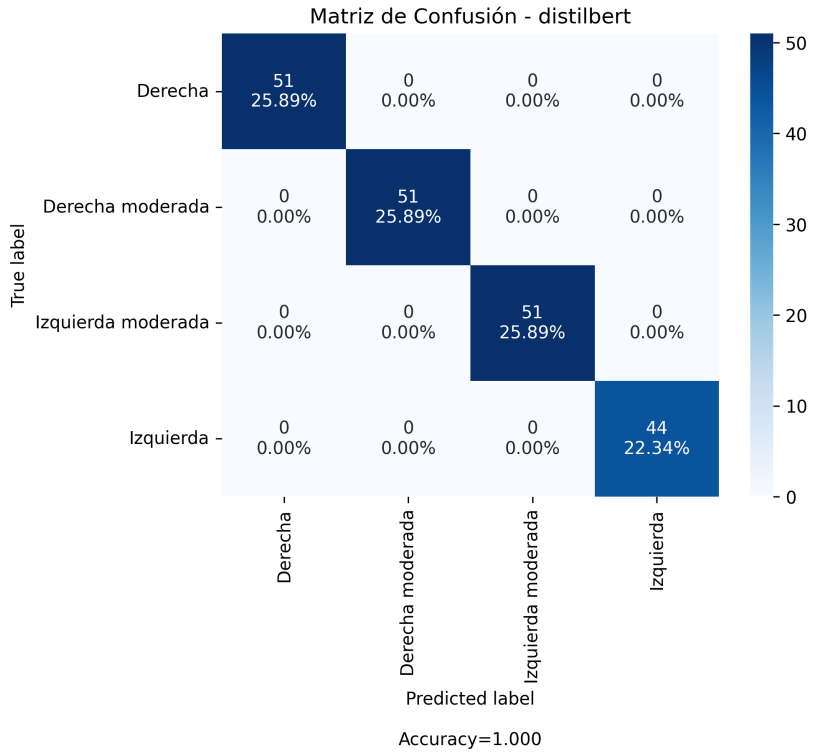

Figura 23: Matriz de confusión para DistilBERT (clúster tamaño 20)

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
	Tamaño de clúster: 5			
Derecha	0.99	1.00	0.99	205
Derecha moderada	0.99	0.99	0.99	205
Izquierda moderada	0.98	0.99	0.98	204
Izquierda	1.00	0.97	0.99	178
	Tamaño de clúster: 10			
Derecha	1.00	1.00	1.00	102
Derecha moderada	1.00	1.00	1.00	102
Izquierda moderada	1.00	1.00	1.00	102
Izquierda	1.00	1.00	1.00	89
	Tamaño de clúster: 20			
Derecha	1.00	1.00	1.00	51
Derecha moderada	1.00	1.00	1.00	51
Izquierda moderada	1.00	1.00	1.00	51
Izquierda	1.00	1.00	1.00	44

Figura 24: Reporte de clasificación para RoBERTa-BNE

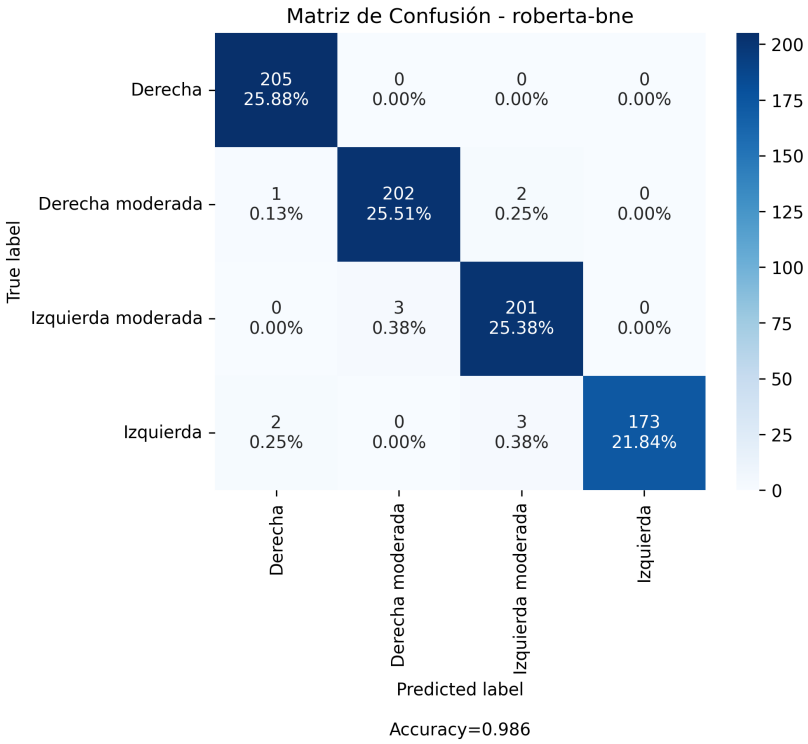

Figura 25: Matriz de confusión para RoBERTa-BNE (clúster tamaño 5)

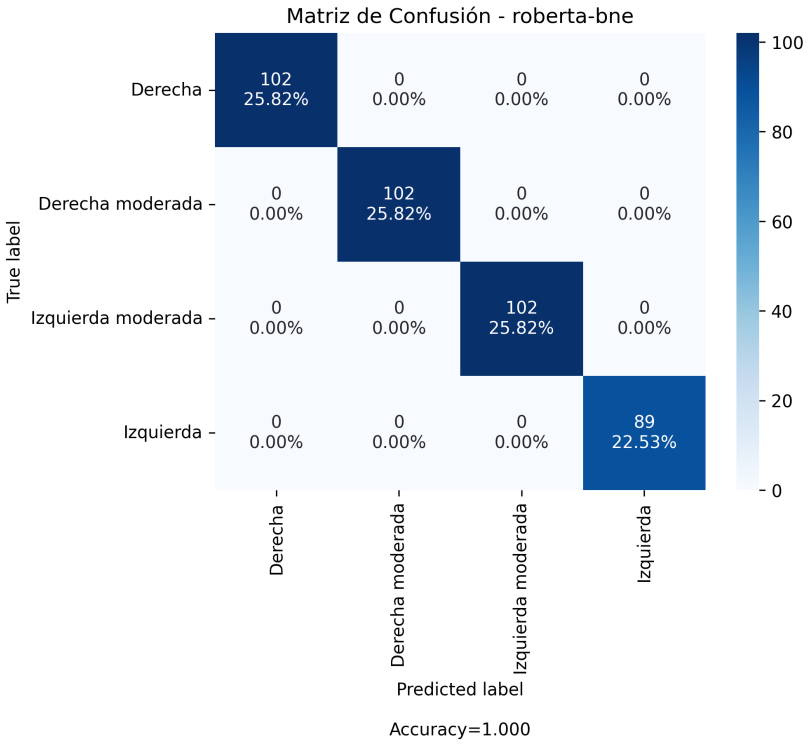

Figura 26: Matriz de confusión para RoBERTa-BNE (clúster tamaño 10)

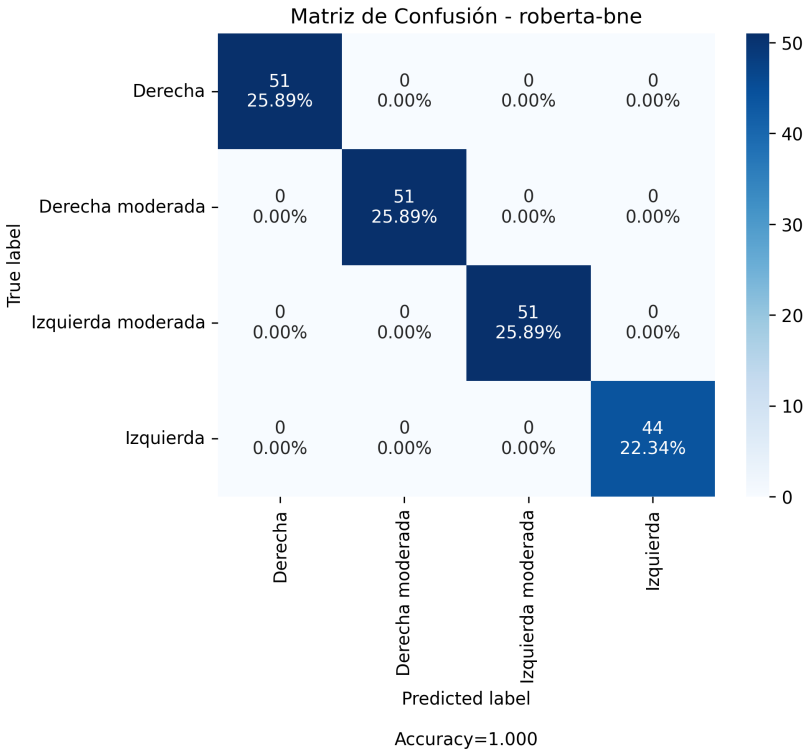

Figura 27: Matriz de confusión para RoBERTa-BNE (clúster tamaño 20)

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
	Tamaño de clúster: 5			
Derecha	0.93	1.00	0.96	205
Derecha moderada	0.94	0.91	0.93	205
Izquierda moderada	0.97	0.95	0.96	204
Izquierda	0.99	0.96	0.97	178
	Tamaño de clúster: 10			
Derecha	0.98	1.00	0.99	102
Derecha moderada	1.00	0.98	0.99	102
Izquierda moderada	1.00	1.00	1.00	102
Izquierda	1.00	1.00	1.00	89
	Tamaño de clúster: 20			
Derecha	1.00	1.00	1.00	51
Derecha moderada	1.00	1.00	1.00	51
Izquierda moderada	1.00	1.00	1.00	51
Izquierda	1.00	1.00	1.00	44

Figura 28: Reporte de clasificación para Google BERT

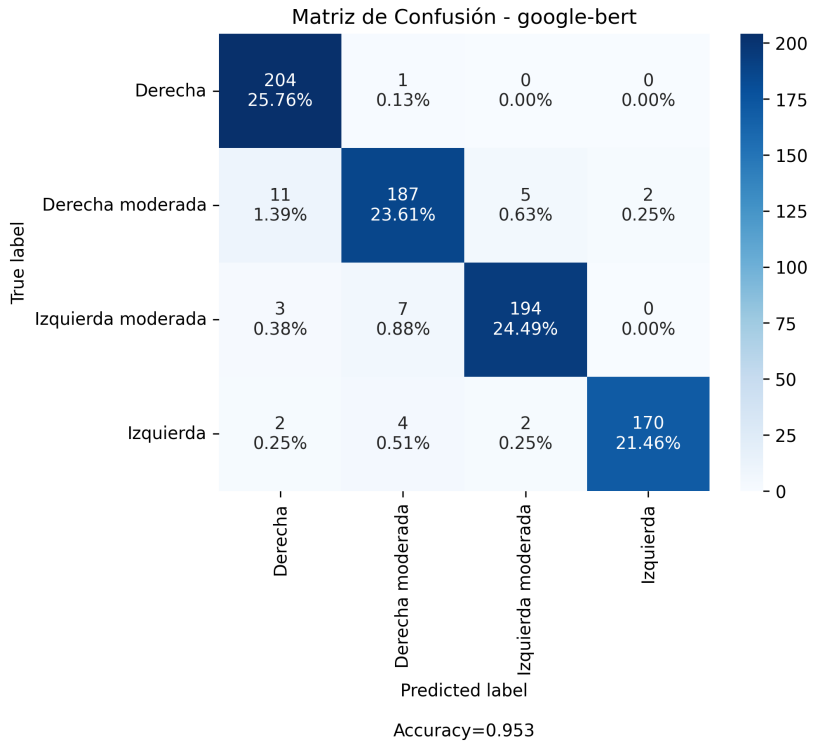

Figura 29: Matriz de confusión para Google BERT (clúster tamaño 5)

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
Derecha	0.05	0.02	0.03	50
Derecha moderada	0.00	0.00	0.00	53
Izquierda moderada	0.00	0.00	0.00	50
Izquierda	0.75	0.35	0.48	51

Figura 30: Reporte de clasificación para BETO con separación estricta por canal

Clase	Precisión	Recall	F1-score	Soporte
Derecha	0.00	0.00	0.00	50
Derecha moderada	0.15	0.40	0.22	53
Izquierda moderada	0.00	0.00	0.00	50
Izquierda	0.23	0.16	0.19	51

Figura 32: Reporte de clasificación para DistilBERT con separación estricta por canal

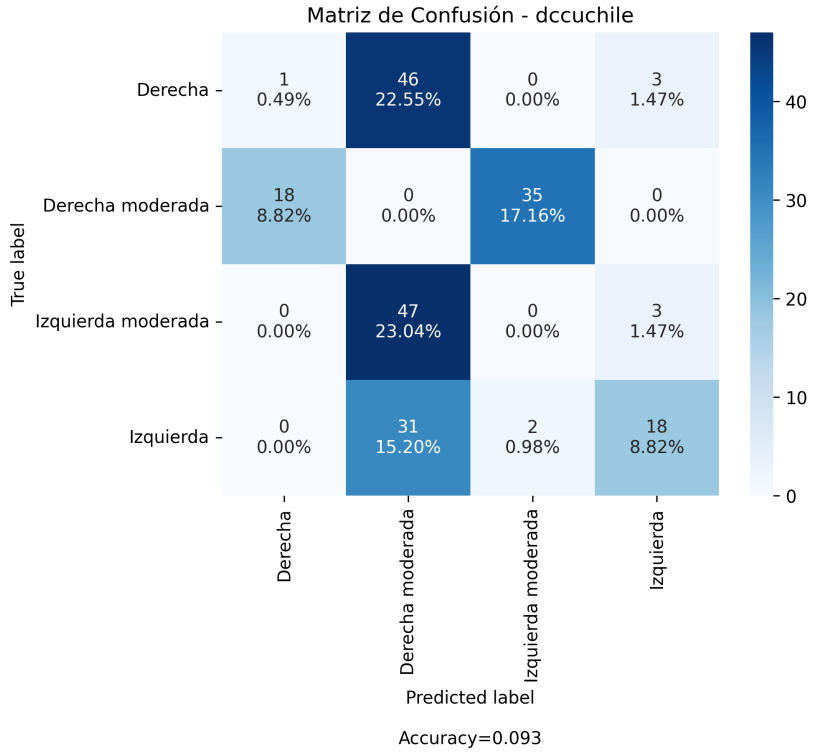

Figura 31: Matriz de confusión para BETO con separación estricta por canal

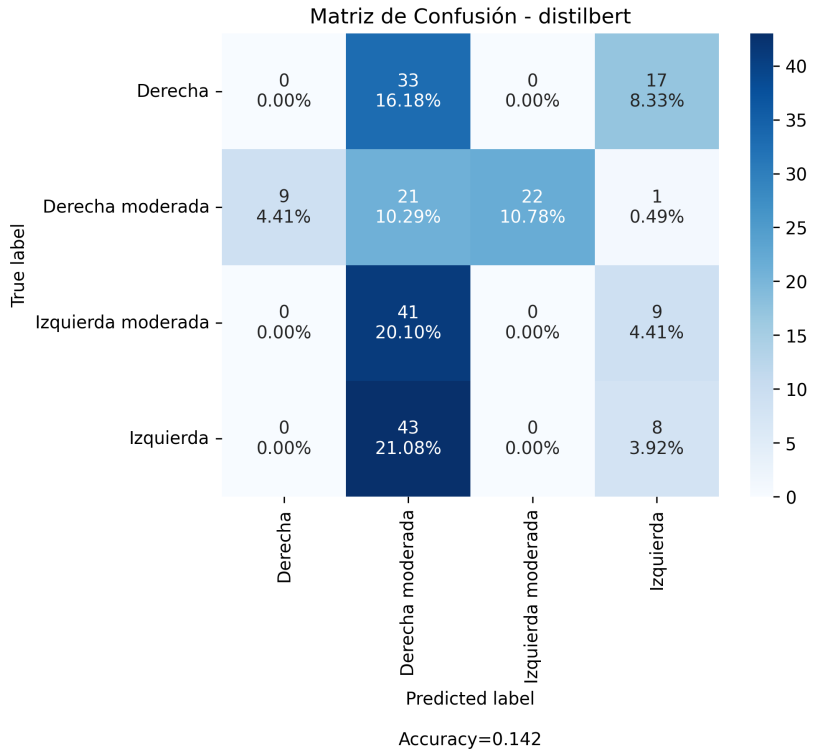

Figura 33: Matriz de confusión para DistilBERT con separación estricta por canal

Clase	Precisión	Recall	F1-score	Soporte
Derecha	0.03	0.02	0.02	50
Derecha moderada	0.04	0.09	0.05	53
Izquierda moderada	0.00	0.00	0.00	50
Izquierda	0.73	0.31	0.44	51

Figura 34: Reporte de clasificación para RoBERTa-BNE con separación estricta por canal

Clase	Precisión	Recall	F1-score	Soporte
Derecha	0.50	0.06	0.11	50
Derecha moderada	0.04	0.09	0.06	53
Izquierda moderada	0.04	0.04	0.04	50
Izquierda	0.16	0.12	0.13	51

Figura 35: Reporte de clasificación para Google BERT con separación estricta por canal

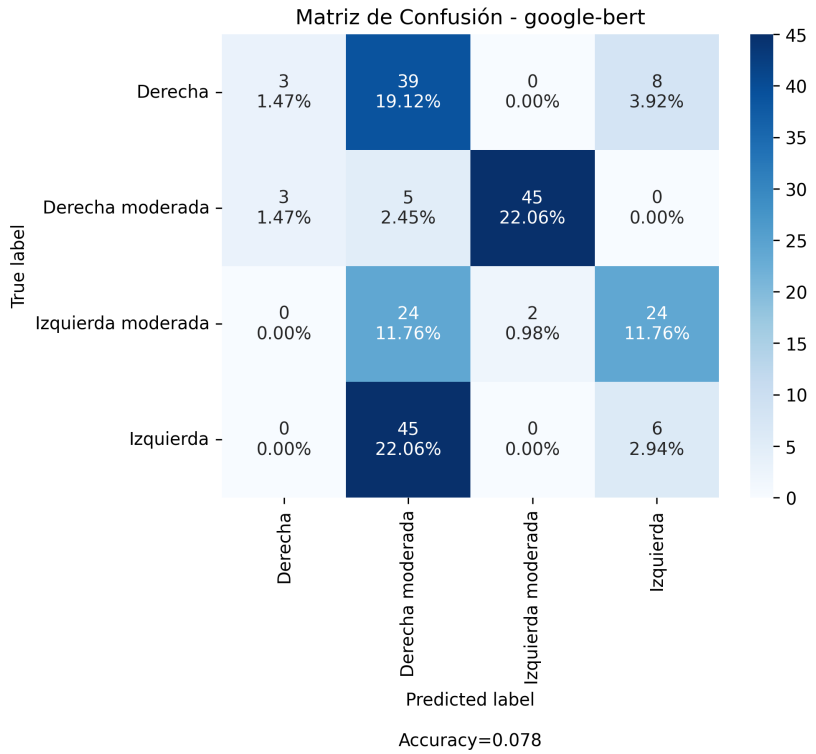

Figura 36: Matriz de confusión para Google BERT con separación estricta por canal