Detecting Multilingual, Multicultural and Multievent Online Polarization



Trabajo de grado presentado como requisito parcial para optar al título de:

Científico de Datos y Comunicador Social

Juan Sebastián Parra Sandoval

Tutor: Luis Gabriel Moreno Sandoval

Pontificia Universidad Javeriana

Facultades de Ciencias y Comunicación y Lenguaje

Pregrado en Ciencia de Datos y Comunicación Social

Bogotá D.C

Agradecimientos

Este trabajo no habría sido posible sin el apoyo incondicional de las personas que me acompañaron a lo largo de este camino. A mi familia, por su paciencia infinita, por aguantar las trasnochadas, los momentos de estrés y por estar siempre presentes con una palabra de aliento cuando más lo necesitaba. Este logro también es de ustedes. Al profesor Luis Gabriel Moreno Sandoval, por su guía, su disposición y por confiar en esta propuesta desde el principio. A mis amigos, Lina, Danna, Daniela por escucharme hablar de polarización, embeddings y redes neuronales más veces de las que cualquiera quisiera, y aun así seguir ahí. Gracias por el apoyo, las risas y por hacer más llevadero el proceso. A todos, gracias.

Resumen

La polarización en línea es un fenómeno creciente en los entornos digitales que afecta la cohesión social, el diálogo público y la formación de opinión en distintas culturas e idiomas. Esta investigación se desarrolla en el marco del desafío internacional SemEval 2026 - Task 9 POLAR, cuyo objetivo es la detección automática de polarización multilingüe, multicultural y multievento. El trabajo propone el desarrollo y evaluación de un sistema computacional de detección de polarización en textos en español e inglés, abordando tres subtareas: identificación binaria de polarización, clasificación del objetivo discursivo y reconocimiento de la manifestación lingüística.

La metodología se fundamenta en el enfoque *Design Science Research* (DSR), estructurado en tres ciclos interdependientes: relevancia, diseño y rigor. El sistema combina representaciones semánticas profundas basadas en *embeddings* contextuales multilingües con una extensión bioinspirada denominada POLAR+Genoma, que discretiza los vectores de *embeddings* en secuencias tipo k-mer análogas a la codificación genética. Se evaluaron 18 configuraciones por subtarea, combinando tres tipos de representación con clasificadores supervisados clásicos (LightGBM, SVM, Regresión Logística) y arquitecturas de redes neuronales modernas (MLP-Deep, BiLSTM, TextCNN), aplicando validación cruzada estratificada de 5 pliegues y pruebas t de Student pareadas.

Los resultados evidencian un F1-macro de 0.778 en la subtarea binaria, 0.609 en la multiclase y 0.348 en la multi-label, valores consistentes con benchmarks internacionales para tareas de alta complejidad discursiva. La representación híbrida POLAR+Genoma demostró ser competitiva y estable frente a la línea base, sin diferencias estadísticamente significativas, confirmando su validez como extensión metodológica para la detección de polarización en entornos multilingües.

Palabras clave: polarización en línea, procesamiento del lenguaje natural, clasificación multilingüe, embeddings, k-mers, aprendizaje supervisado.

Índice de tablas

Tabla 3-1. : Resumen para aportes sobre polarización online y detección multilingüe(Elab. Propia).....18
Tabla 3-2. : Comparación de pasos en la metodología y limitaciones de las propuestas para la línea base(Elab. Propia).....20
Tabla 3-3.: Criterios para clasificar modelos observados (Elab. Propia).....24
Tabla 3-4. : Comparación de pasos en la metodología y limitaciones de las propuestas para la extensión del artefacto(Elab. Propia).....27
Tabla 3-5.: Resumen distribución revistas (Elab. Propia).....29
Tabla 4-1.: Relación entre columnas del dataset y las sub tareas (Elab. propia)..... 62
Tabla 4-2.: Metadata de los datasets utilizados en el estudio (Elab. propia)..... 76
Tabla 4-3.: Balance de clases Subtarea 1 (Elab. propia).....77
Tabla 4-4.: Balance de clases Subtarea 2 (Elab. propia).....77
Tabla 4-5.: Balance de clases Subtarea 3 (Elab. propia).....77
Tabla 6-1.: Resultados subtask 1 (Elab. propia)..... 104
Tabla 6-2.: Evaluación t de Student Subtask 1 (Elab. propia)..... 105
Tabla 6-1b.: F-score por clase — Subtarea 1 (Elab. propia).....105
Tabla 6-1c.: AUC — Subtarea 1 (Elab. propia).....106
Tabla 6-3.: Resultados subtask 2 (Elab. propia)..... 108
Tabla 6-4.: Evaluación t de Student Subtask 2 (Elab. propia)..... 109
Tabla 6-3b.: F-score por categoría — Subtarea 2 (Elab. propia).....109
Tabla 6-5.: Resultados subtask 3 (Elab. propia).....111
Tabla 6-6.: Evaluación t de Student Subtask 3 (Elab. propia)..... 113
Tabla 6-5b.: F-score por etiqueta — Subtarea 3 (Elab. propia).....113
Tabla 6-7.: Comparación consolidada por representación y subtarea (Elab. propia).....116
Tabla 6-8.: Análisis de significancia estadística — Subtarea 1 (Elab. propia).....122
Tabla 6-9.: Análisis de significancia estadística — Subtarea 2 (Elab. propia).....122
Tabla 6-10.: Análisis de significancia estadística — Subtarea 3 (Elab. propia).....123

Índice de figuras

Figura 1-1.: Complejidad y Alcance de modelos analizados (Elab. Propia)..... 25
Figura 1-2.: Clasificación de revistas (Elab. Propia)..... 30
Figura 1-3.: Distribución de idiomas (Elab.propia)..... 65
Figura 1-4.: Top palabras en Español (Elab.propia)..... 66
Figura 1-5.: Top palabras en Inglés (Elab.propia)..... 67
Figura 1-6.: Diversidad léxica y longitud promedio (Elab.propia)..... 68
Figura 1-7.: Comparación de diversidad léxica y longitud (Elab.propia)..... 69
Figura 1-8.: Complejidad textual (Elab.propia)..... 70
Figura 1-9.: Distribución morfológica Español (Elab.propia)..... 71
Figura 1-10.: Distribución morfológica Inglés (Elab.propia)..... 72
Figura 1-11.: Top palabras en Español dataset unificado (Elab.propia)..... 76
Figura 1-12.: Top palabras en Inglés dataset unificado (Elab.propia)..... 77
Figura 1-13.: Diversidad léxica y longitud promedio dataset unificado (Elab.propia)..... 78
Figura 1-14.: Comparación de diversidad léxica y longitud dataset unificado(Elab.propia)..... 79
Figura 1-15.: Complejidad textual dataset unificado(Elab.propia)..... 79
Figura 1-16.: Distrib. morfológica Español dataset unificado (Elab.propia)..... 80
Figura 1-17.: Distrib. morfológica Inglés dataset unificado (Elab.propia)..... 81
Figura 1-18.: Ciclos en DSR (Tomado de Hevner [24])..... 83
Figura 1-19.: Ciclo de relevancia (Elab. Propia)..... 85
Figura 1-20.: Ciclo de diseño (Elab. Propia)..... 86
Figura 1-21.: Ciclo de rigor (Elab. Propia)..... 88
Figura 1-22.: Interconexión entre ciclos (Elab. Propia)..... 89

Índice

Agradecimientos **3**
Resumen **5**
Índice de tablas **8**
Índice de figuras **9**
1. Introducción **12**
1.1 Planteamiento del problema 12
1.2 Justificación 13
2. Objetivos **14**
2.1 Objetivo general 14
2.2 Objetivos específicos 14
3. Marco Teórico y Estado del Arte **15**
3.1 Resumen 15
3.2 Estado del Arte 16
3.3 Polarización 31
3.3.1 Polarización Política y social 31
3.3.2 Polarización Digital 32
3.3.2.1 Polarización en Redes Sociales 32
3.3.2.2 Polarización multilingüe y multicultural 33
3.3.3 Expresiones de la polarización 34
3.3.3.1 Lenguaje Extremo 34
3.3.3.2 Estereotipos 35
3.3.3.3 Invalidación 35
3.3.3.4 Falta de empatía 35
3.3.3.5 Deshumanización 36
3.4 Representación del texto 36
3.4.1 Representaciones clásicas 37
3.4.1.1 Term frequency - Inverse Document frequency (TF-IDF) 38
3.4.1.2 N-gramas 39
3.4.2 Representaciones distribuidas 41
3.4.2.1 Word2Vec / GloVe 41
3.4.2.2 Embeddings contextuales (BERT, mBERT, XLM-R) 44
3.4.3 Masked Language Models (MLM) 48
3.4.4 Large Language Models (LLM) 49
3.4.5 Enfoque genómico-semántico en embeddings 51

3.5 Reducción de dimensionalidad..... 54
3.5.1 Análisis de Componentes principales (ACP).....55
3.5.2 t-SNE y UMAP (para visualización de polarización en embeddings).....56
3.6 Aprendizaje automático y Agrupamiento..... 58
3.6.1 Algoritmos de clasificación supervisada..... 58
3.6.1.1 SVM, Random Forest, Redes Neuronales..... 59
4. Conjunto de Datos..... 61
4.1 Análisis exploratorio SemEval..... 64
4.2 Análisis exploratorio unificado.....73
4.3 Preparación de datos y representaciones.....85
5. Metodología..... 86
5.1 Ciclo de Relevancia..... 87
5.2 Ciclo de Diseño.....88
5.3 Ciclo de Rigor.....90
5.4 Interconexión entre ciclos.....91
5.5 Actividades macro para el desarrollo de la metodología.....92
6. Implementación del artefacto y resultados.....95
6.1 Descripción general del experimento.....95
6.1.1 Descripción del Subtask 1: Identificación de polarización binaria.....97
6.1.2 Descripción del Subtask 2: Clasificación del objetivo de la polarización.....99
6.1.3 Descripción del Subtask 3: Reconocimiento de la manifestación lingüística.....100
6.2 Resultados del subtask 1..... 102
6.3 Resultados del subtask 2..... 106
6.4 Resultados del subtask 3..... 110
6.5 Análisis e Interpretación..... 113
6.6 Limitaciones.....117
6.7 Implicaciones de las métricas para tareas downstream.....119
6.8 Análisis de significancia estadística, corrección de Bonferroni..... 121
7. Conclusiones..... 124
Referencias.....127

1. Introducción

1.1 Planteamiento del problema

En la actualidad, los entornos digitales y las redes sociales se han consolidado como los principales espacios de discusión pública [1]. En estos escenarios, el intercambio de opiniones no solo se intensifica, sino que también tiende a estructurarse en dinámicas de polarización que fragmentan las comunidades y refuerzan las denominadas cámaras de eco [2].

La polarización no se limita al ámbito político, sino que se manifiesta en dimensiones sociales, culturales y lingüísticas, generando tensiones que afectan la cohesión social y dificultan el diálogo entre grupos con posturas divergentes. Este fenómeno adquiere mayor complejidad en contextos donde coexisten múltiples idiomas [8], culturas y tipos de eventos, lo que evidencia la necesidad de enfoques analíticos más integrales.

En este contexto, surge el desafío de diseñar sistemas de procesamiento del lenguaje natural capaces de identificar la polarización de manera automática, considerando simultáneamente factores lingüísticos, culturales y contextuales. Esto implica superar las limitaciones de enfoques tradicionales que, en su mayoría, se centran en escenarios monolingües o en eventos específicos.

La presente investigación se enmarca en el desafío internacional SemEval 2026 - Task 9 POLAR, el cual propone la detección de polarización multilingüe, multicultural y multievento mediante subtareas de clasificación binaria, categórica y multi-label. Este

marco proporciona un entorno estandarizado para la evaluación de modelos, permitiendo comparar resultados en un contexto de referencia internacional.

1.2 Justificación

Analizar la polarización en medios digitales con un enfoque multilingüe, multicultural y multievento ofrece múltiples beneficios. En primer lugar, permite una comprensión más amplia del fenómeno, observando cómo varía entre comunidades. En segundo lugar, facilita la identificación de dinámicas transnacionales con mayor precisión: el enfoque multievento y multilingüe elimina las barreras lingüísticas como obstáculo, lo que permite observar narrativas polarizantes que circulan entre diferentes países y analizar cómo los discursos se traducen y adaptan en distintos contextos.

El enfoque multicultural permite identificar sesgos algorítmicos y cámaras de eco en distintos contextos, así como determinar si los algoritmos de recomendación refuerzan la polarización de manera uniforme o si existen variaciones. La integración de los tres enfoques genera conocimiento interdisciplinar que contempla el idioma, la cultura y el tipo de evento, fortaleciendo las investigaciones previas sobre el fenómeno.

Más allá del ámbito académico, los resultados y hallazgos pueden ofrecer insumos a organizaciones no gubernamentales (ONGs) y medios de comunicación para diseñar estrategias comunicativas que consideren la diversidad cultural, lingüística, política y

social, reduciendo la reproducción de discursos polarizantes en medios digitales. Además, este enfoque integral es más adecuado para poder trabajar y desplegar herramientas de PLN que sean efectivas, inclusivas y culturalmente sensibles, capaces de responder a las diversas formas en las que se expresa la polarización digital [4].

2. Objetivos

2.1 Objetivo general

Desarrollar un sistema de detección de polarización en línea para textos en español e inglés, capaz de abordar de manera integrada tres subtareas: i) identificar si un mensaje está polarizado, ii) clasificar el objetivo de la polarización y iii) reconocer la forma en que dicha polarización se manifiesta lingüísticamente.

2.2 Objetivos específicos

1. Implementar un modelo de clasificación binaria que determine la presencia o ausencia de polarización en textos en español e inglés.
2. Diseñar un clasificador multiclase que identifique el objetivo de la polarización en las categorías definidas (política, religiosa, racial/étnica, género/sexual y otras).
3. Desarrollar un sistema de clasificación que reconozca la manifestación lingüística de la polarización (lenguaje extremo, estereotipos, invalidación, falta de empatía, deshumanización y otras).

4. Evaluar el desempeño del sistema sobre un corpus multilingüe unificado (español e inglés) mediante validación cruzada estratificada, analizando la robustez y capacidad de generalización del modelo en escenarios de clasificación binaria, multiclase y multi-label.
5. Analizar los resultados obtenidos con base en métricas de F1-macro y pruebas de significancia estadística, comparando el desempeño de las distintas representaciones y arquitecturas evaluadas a través de las tres subtareas, e identificando los retos particulares asociados a cada nivel de complejidad discursiva.

3. Marco Teórico y Estado del Arte

3.1 Resumen

El presente marco teórico integró los principales conceptos, métodos y enfoques necesarios para el estudio de la polarización en entornos digitales multi evento, multilingües y multiculturales. Se abordaron las distintas manifestaciones de la polarización (política, social y en redes sociales) y sus expresiones lingüísticas, así como las técnicas de representación de texto, reducción de dimensionalidad y aprendizaje automático que permiten procesar grandes volúmenes de datos multilingües. Del mismo modo, se revisaron técnicas de agrupamiento y algoritmos de clasificación supervisada aplicados a la detección automática de discursos polarizados.

Adicionalmente, se realizó un análisis detallado del estado del arte, en el que se identificaron y compararon los modelos propuestos por autores clave, evaluando sus técnicas, niveles de complejidad y alcance. Esta revisión crítica permitió establecer criterios de valoración y extraer *insights* sobre las ventajas y limitaciones de cada enfoque, lo cual orienta la selección y adaptación del modelo más adecuado para esta investigación. Dicho análisis no solo respalda la justificación metodológica, sino que también facilita la integración de buenas prácticas y la identificación de áreas de mejora para el desarrollo de un sistema de detección de polarización multilingüe y multi evento acorde con los objetivos planteados.

3.2 Estado del Arte

Para poder comprender a profundidad el fenómeno de la polarización desde un enfoque multicultural, multilingüe y multi evento, es necesario definir algunos conceptos fundamentales que se abordan a lo largo de esta investigación. Para lograr el entendimiento de los conceptos, se revisaron los principales avances recientes en el estudio de la polarización en redes sociales y en las técnicas de PLN aplicadas a la detección automática de posturas y expresiones polarizadas.

El objetivo es poder establecer las bases conceptuales y metodológicas que sustentan esta investigación, teniendo como base trabajos publicados en revistas indexadas Q1 y Q2 de

repositorios virtuales como ScienceDirect, Springer, Wiley y Nature, las cuales son plataformas virtuales que reúnen artículos científicos con muy buena reputación y confiabilidad. Los trabajos encontrados abordan tanto la medición estructural de la polarización en comunidades digitales como su manifestación lingüística en entornos multilingües.

Asimismo, se identifican investigaciones emergentes que exploran la analogía entre los modelos de representación semántica y las estructuras genómicas, perspectiva que se desarrolla en detalle en la sección 3.4.5 de este documento.

De esta manera, se integran aportes que permiten comprender la polarización desde una perspectiva multidimensional (afectiva, discursiva, ideológica) identificando enfoques computacionales que logran el análisis multilingüe que, en el caso de la investigación, hace referencia al análisis en los idiomas inglés y español respectivamente. A continuación se muestra la recopilación de los trabajos en la tabla **3-1** donde se encuentra el año de publicación, los autores, su aporte clave y los conceptos abordados que sirven como objeto para la investigación.

Tabla 3-1. : Resumen para aportes sobre polarización online y detección multilingüe(Elab. Propia)

Año	Autor(es)	Aporte clave	Conceptos
2025	Usman Naseem, Muhammad Imran Malik, Muhammad Rameez, Khurum Nazir Junejo, Syed Waqar Jaffry, Imran Razzak [4].	Presentan un dataset multilingüe y multicultural con tres subtareas: detección binaria de polarización, clasificación del tipo y forma de expresión. Evaluado con modelos *Masked Language Models* (MLM) y *Large Language Models* (LLM) con técnicas *few-shot, zero-shot*.	La base metodológica se apoyó en la referencia de la competencia POLAR-SemEval 2025.
2025	Bharathi A, Arkaitz Zubiaga [9].	Método para *zero-shot cross-lingual stance*	Técnica para la transferencia de inglés a español sin etiquetas.
2024	Hartmann et al. [10].	Revisión sistemática de 129 estudios sobre cámaras de eco en redes sociales, analizando conceptualizaciones, metodologías y resultados contradictorios en la literatura.	Marco estructural para la polarización multi evento y análisis comparativo de enfoques computacionales.
2024	Andres Karjus, Christine Cuskley [11].	Midieron la divergencia lingüística entre grupos polarizados en redes.	Indicadores para la subtarea 3 (manera de expresión).

2023	Valensise, C. M., Cinelli, M., & Quattrociocchi, W.[12].	Proponen un marco computacional para modelar los mecanismos que impulsan la polarización en línea, ajustando modelos matemáticos a datos reales de redes sociales.	Polarización estructural y comunitaria; métricas cuantitativas para medir divergencia ideológica en redes.
2024	Esshaan Mahajan, Hemaank Mahajan, Sanjay Kumar [13].	*Ensemble* multilingüe para la detección del odio y el *cyberbullying.*	Técnicas de transferencia para el lenguaje extremo.
2024	Mets, M., Karjus, A., Ibrus, I., & Schich, M. [14]	Evalúan la detección de postura (stance), en lenguajes complejos, a su vez, muestran como GPT-3.5 funciona considerablemente mejor que otros modelos en escenarios few-shot	Transferencia metodológica en escenarios de bajo recursos; límites de la transferencia directa entre idiomas sin adaptación.
2024	Kaixin Ding, Wei Wang, Zhen Wang, Yongjian Yang [15].	Introducen chain-of-thought prompting para stance detection, logrando mejoras en escenarios few-shot y zero-shot con LLMs.	Uso de prompting avanzado para las subtareas 1-3 (detección, clasificación y expresión).
2022	Aldayel, A., & Magdy, W. [16].	Analizan el rol de los bots en la polarización de posturas en redes sociales, identificando su presencia en más de 19 millones de cuentas de Twitter en	Relación entre bots, polarización multi evento y dinámica de usuarios en redes sociales.

	7 temas controversiales.

Nota: Recuperado de *Scimago Journal & Country Rank (SJR)*, páginas de clasificación Q1/Q2 de cada revista consultada: *IEEE Access* (Usman Naseem et al., 2025; Bharathi & Zubiaga, 2025), *Journal of Computational Social Science*(Hartmann et al., 2025), *Humanities and Social Sciences Communications* (Karjus & Cuskley, 2024), *Information Sciences*(Valensise, Cinelli & Quattrociocchi, 2023), *Expert Systems with Applications* (Mahajan, Mahajan & Kumar, 2024; Ding, Wang, Wang & Yang, 2024), *PLOS ONE* (Mets, Karjus, Ibrus & Schich, 2024), *Social Network Analysis and Mining* (Aldayel & Magdy, 2022).

Una vez hemos profundizado en los aportes de cada una de las investigaciones presentadas en la tabla **3-1,** se presenta profundizar con los mismos trabajos pero esta vez ahondando y verificando qué modelos usaron, qué propuesta realizaron, los principales pasos junto con su metodología y lo que se logra destacar y a su vez lo que lo limita. Este análisis tiene como fin de agregar una dirección conceptual a la investigación, recorriendo así por el estado del arte y construyendo una opción para el modelo del sistema, el cual será el encargado de resolver la actual propuesta, toda la información de dicho análisis se encuentra recopilada en la tabla **3-2** que se presenta a continuación:

Tabla 3-2. : Comparación de pasos en la metodología y limitaciones de las propuestas para la línea base(Elab. Propia).

Autor(es)	**Qué modelo / propuesta realizaron**	**Principales pasos/metodología**	**Lo que destaca / limitaciones**
Usman Naseem, Muhammad Imran Malik, Muhammad Rameez, Khurum Nazir Junejo, Syed Waqar Jaffry, Imran Razzak [4].	*POLAR*, Benchmark para detección de polarización multilingüe, multicultural y multievento.	Creación de dataset con instancias en 7 idiomas y múltiples eventos. Anotación sobre ejes: presencia de polarización, tipo y	Destaca por estructurar datos multiculturales y multieventos, muy útil para generalizar. Limitación: los modelos funcionan

		manifestación. Fine-tuning de modelos multilingües pre-entrenados en configuraciones monolingüe y cross-lingüe. Evaluación de modelos LLM tanto en escenarios con pocos ejemplos *(few-shot)* como cero ejemplos *(zero-shot)*.	bien en la detección binaria, pero bajan el rendimiento cuando se evalúan tipos más finos de polarización o manifestaciones. También hay diferencias fuertes entre idiomas de bajos recursos.
Bharathi A, Arkaitz Zubiaga [9].	Evaluación de LLMs multilingües para detección de polarización.	Incluir modelos grandes, probaron *zero-shot/few-shot*, comparar desempeño entre idiomas, probablemente con adaptaciones al idioma.	Resaltan la brecha de rendimiento entre idiomas con muchos datos contra aquellos con pocos ejemplos; por otro lado, una limitación es la sensibilidad del modelo al dominio del evento y al lenguaje cultural particular.
Hartmann et al. [10].	Revisión sistemática de cámaras de eco en redes sociales.	Revisión de 129 estudios con clasificación de conceptualizaciones y metodologías; análisis de factores regionales, culturales y de plataforma que explican resultados contradictorios en la literatura.	Trabajo metodológico que provee un marco comprensivo para analizar polarización en sistemas multipartidistas; útil para contextualizar enfoques multi evento. Limitación: no incluye experimentos propios de modelado.
Andres Karjus, Christine Cuskley [11].	Estudio de marcadores lingüísticos y culturales de polarización.	Analizan indicadores lingüísticos específicos por ejemplo, estilos de lenguaje, elección de léxico, estructura de discurso y cómo se	Buen aporte para características textuales, sin embargo, algunas limitaciones son quizás menos foco en redes o en dinámica

		relacionan con polarización en distintos idiomas.	temporal profunda.
Valensise, C. M., Cinelli, M., & Quattrociocchi, W. [12].	Marco computacional para modelar los drivers de la polarización en redes sociales.	Ajustan modelos matemáticos de polarización a datos empíricos de redes; comparan múltiples mecanismos (homofilia, exposición selectiva, refuerzo algorítmico) para identificar cuáles explican mejor la dinámica observada	Ofrece evidencia cuantitativa sobre los mecanismos subyacentes de la polarización, útil para fundamentar el diseño del sistema. Limitación: centrado en dinámica de red, no aborda directamente la detección textual ni escenarios multilingües.
Esshaan Mahajan, Hemaank Mahajan, Sanjay Kumar [13].	Ensemble multilingüe (EnsMulHateCyb) para detección de discurso de odio y cyberbullying en redes sociales.	Combinan múltiples clasificadores sobre representaciones multilingües; evalúan en datasets de varias lenguas con técnicas de transferencia entre idiomas para detección de contenido ofensivo.	Relevante para el diseño del sistema ya que demuestra que los enfoques ensemble mejoran la robustez en escenarios multilingües. Limitación: centrado en odio explícito, no aborda polarización ideológica sutil.
Mets, M., Karjus, A., Ibrus, I., & Schich, M. [14]	Evaluación de clasificación textual cross-lingual y el impacto del ruido en modelos de PLN.	Evalúan modelos de clasificación en escenarios cross-lingual, analizando cómo el ruido y las variaciones lingüísticas afectan el desempeño; comparan estrategias de adaptación entre idiomas.	Muy relevante para escenarios de bajo recursos; muestra los límites de la transferencia directa entre idiomas sin adaptación. Limitación: no aborda específicamente la polarización como fenómeno discursivo.
Kaixin Ding, Wei Wang, Zhen Wang,	Chain-of-thought prompting para	Aplican prompting estructurado con	Directamente aplicable a las subtareas 1-3;

Yongjian Yang [15].	stance detection en escenarios few-shot y zero-shot con LLMs.	razonamiento paso a paso sobre modelos de lenguaje grandes; evalúan en múltiples datasets de detección de postura con distintas configuraciones de ejemplos.	demuestra mejoras significativas con prompting avanzado en escenarios con pocos datos. Limitación: dependiente de modelos LLM grandes, lo que implica costo computacional elevado.
Aldayel, A., & Magdy, W. [16].	Estudio sobre el rol de los bots en la polarización de posturas en redes sociales.	Analizan interacciones directas e indirectas de más de 4.000 usuarios de Twitter con 19 millones de cuentas; identifican bots dentro de señales predictoras de postura en 7 temas controversiales.	Relevante para comprender cómo agentes automatizados intensifican la polarización multi evento. Limitación: centrado en Twitter y en contextos angloparlantes, lo que limita su generalización a otros idiomas.

Nota: Recuperado de *Scimago Journal & Country Rank (SJR)*, páginas de clasificación Q1/Q2 de cada revista consultada: *IEEE Access* (Usman Naseem et al., 2025; Bharathi & Zubiaga, 2025), *Journal of Computational Social Science*(Hartmann et al., 2025), *Humanities and Social Sciences Communications* (Karjus & Cuskley, 2024), *Information Sciences*(Valensise, Cinelli & Quattrociocchi, 2023), *Expert Systems with Applications* (Mahajan, Mahajan & Kumar, 2024; Ding, Wang, Wang & Yang, 2024), *PLOS ONE* (Mets, Karjus, Ibrus & Schich, 2024), *Social Network Analysis and Mining* (Aldayel & Magdy, 2022).

Con base en los modelos revisados, se elaboraron una tabla y una figura adicionales para evaluar la complejidad y el alcance de cada propuesta, con el fin de seleccionar las técnicas más adaptables a esta investigación, para esto se tuvieron los siguientes

criterios para estas dos variables (alcance y complejidad) según la técnica usada por los diferentes autores, los criterios fueron resumidos en la tabla **3-3** que se presenta a continuación:

Tabla 3-3.: Criterios para clasificar modelos observados (Elab. Propia).

Puntuación	Nivel de complejidad	Alcance
1	Técnica sencilla, bibliotecas listas, poco preprocesamiento.	Aplica solo a un idioma o dataset muy específico.
2	Requiere algunos pasos extra (limpieza de datos, entrenamiento básico).	Funciona en un idioma pero podría ampliarse con esfuerzo.
3	Algoritmo intermedio con varios hiperparámetros o clustering no trivial.	Capaz de manejar varios datasets o eventos en un solo idioma.
4	Necesita infraestructura mayor (GPU, optimización, integración de múltiples datasets)	Admite múltiples idiomas/culturas o diferentes plataformas sociales.
5	Modelos de deep learning con multilingüismo, pipelines sofisticados, tuning.	Altamente generalizable.

Con estos criterios establecidos, es posible categorizar y comparar los modelos estudiados según su alcance y complejidad, tal como se observa en la figura **1-1**, donde ambas variables se representan en una escala de 1 a 5.

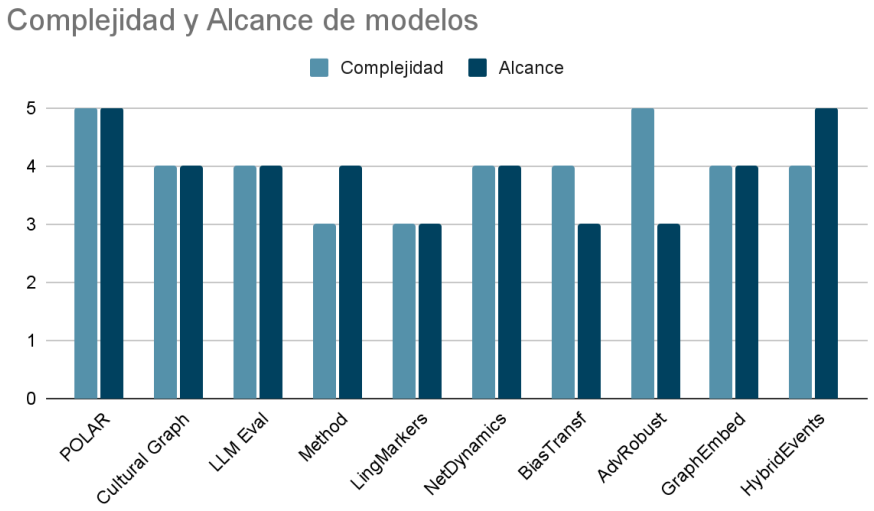


Figura 1-1.: Complejidad y Alcance de modelos analizados (Elab. Propia).

Se observa que POLAR y *AdvRobust* destacan por su alta complejidad (5), pero solo POLAR combina esta exigencia con un alcance igualmente máximo (5), lo que lo convierte en una referencia sólida para sistemas multilingües y multievento. Por otro lado, *HybridEvents* ofrece un equilibrio interesante: complejidad 4 y alcance 5, ideal si se busca cubrir distintos tipos de eventos con menor esfuerzo de implementación que POLAR. En el rango medio, técnicas como LLM Eval, *NetDynamics* y *GraphEmbed* mantienen una buena relación (4-4), ofreciendo marcos potentes sin la carga técnica extrema.

Dado que el objetivo general de la investigación incluye detección de polarización en español e inglés con subtareas de clasificación binaria, multiclase y análisis lingüístico,

un enfoque que equilibre robustez multilingüe y escalabilidad es clave. En este sentido, POLAR aporta la mejor base para entrenamiento y evaluación cruzada, mientras que combinar su estructura de datos con técnicas tipo LLM Eval (que aprovechan modelos de lenguaje avanzados en escenarios *zero/few-shot*) podría acelerar el desarrollo.

Asimismo, *HybridEvents* puede servir como referencia para manejar distintos contextos y eventos, reforzando la capacidad de transferencia entre idiomas, como los que se plantean dentro de los objetivos específicos, con base en este análisis, se definió como línea base la integración de la metodología POLAR y las técnicas LLM Eval.

Por otro lado, aunque la propuesta principal de este trabajo se basa en la integración de las líneas base POLAR y LLM Eval, se introduce además una extensión innovadora que trata los *embeddings* generados como secuencias discretizadas inspiradas en la estructura del ADN. Esta analogía proviene de la metodología en bioinformática que representa secuencias de nucleótidos como vectores o tokens *(k-mers)* y los procesa mediante transformadores y modelos de lenguaje para extraer patrones repetitivos o *motifs* funcionales [26].

Al aplicar una cuantización de *embeddings* y fragmentarlos en k-mers de representaciones léxicas, el sistema puede identificar “motifs lingüísticos de

polarización" que se repiten a través de textos y contextos multilingües, aportando una capa estructural que complementa las representaciones densas tradicionales. Para poder sustentar este enfoque distintivo se recopilaron 5 trabajos de la misma calidad, es decir, revistas indexadas Q1 y Q2 de repositorios virtuales como ScienceDirect, Springer, Wiley y Nature, se resumieron en la tabla **3-4**, donde se contemplaron los autores, la propuesta que realizaron, la metodología, lo que se destaca de cada uno.

Tabla 3-4. : Comparación de pasos en la metodología y limitaciones de las propuestas para la extensión del artefacto(Elab. Propia).

Autor(es)	**Qué modelo / propuesta realizaron**	**Principales pasos/metodología**	**Lo que destaca**
Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. [26].	DNABERT, BERT pre-entrenado en *k-mers* para secuencias de ADN.	Tokenizan el genoma en *k-mers*, pre-entrenan un BERT (máscara de k-mers), y hacen fine-tuning en tareas genómicas (promotores, splicing).	Demuestra que transformers pre-entrenados sobre k-mers capturan patrones funcionales y transfieren bien a tareas específicas.
Zhang, Y., Bai, Z., & Imoto, S. [27].	Investigación sobre BERT en secuencias nucleotídicas, análisis de lo que aprende BERT con *k-mers*.	Descomponen pretraining, comparan distintas estrategias de k-mer embedding y evalúan efectos en tareas genómicas; análisis empírico del aprendizaje.	Aporta evidencia sobre cómo la arquitectura BERT capta patrones en secuencias biológicas y evalúa distintas tokenizaciones/*kvalues* .
Ren, R., Yin, C., & Yau, S.-S.-T. [28].	kmer2vec, embedding *k-mers* vía word2vec	Tokenización en k-mers; entrenan word2vec sobre	Método peer-reviewed que adapta técnicas NLP (word2vec) a

	(Journal of Computational Biology)	k-mers; representan secuencias como vectores y aplican a comparación filogenética / clustering *(phylogeny)*	*k-mers*, mostrando buena precisión y velocidad frente a alineamiento.
Maruyama, O., Li, Y., Narita, H., et al. [29].	CMIC , embeddings de *k-mers* de longitud variable para predecir marcas de metilación *(BMC Bioinformatics)*.	Construyen embeddings de k-mers variable-length, entrenan modelos predictivos (clasificación) y analizan impacto de parámetros	Ejemplo de aplicación práctica de k-mer embeddings a una tarea de clasificación biológica (*BMC Bioinformatics, peer-reviewed*).
Moeckel, C., Mareboina, M., Konnaris, M. A., et al.[30].	Revisión: k-mer methods and applications *(Computational & Structural Biotechnology Journal)*.	Revisión sistemática de métodos k-mer, *alignment-free approaches* y aplicaciones genómicas.	Revisión centrada en bioinformática; requiere adaptar la argumentación para NLP/embeddings.

Nota: Recuperado de *Scimago Journal & Country Rank (SJR)*, páginas de clasificación Q1/Q2 de cada revista consultada: *Bioinformatics* (Ji, Zhou, Liu & Davuluri, 2021; Zhang, Bai & Imoto, 2023), *Computational and Structural Biotechnology Journal* (Moeckel et al., 2024), *BMC Bioinformatics* (Maruyama et al., 2022) y *Journal of Computational Biology* (Ren, Yin & Yau, 2022).

Gracias a los trabajos citados, la presente investigación fundamenta su propuesta metodológica en una línea base compuesta por los modelos POLAR y LLM Eval, los cuales permiten evaluar de manera robusta la polarización textual desde un enfoque multilingüe y comparativo.

No obstante, con el propósito de aportar un componente innovador y diferenciador, se incorpora un nuevo enfoque inspirado en la analogía estructural entre los embeddings lingüísticos y la secuenciación del ADN, reconociendo que ambos representan información compleja mediante patrones distribuidos y relacionales. Esta integración, sustentada en estudios recientes que aplican técnicas de representación genética a la modelización de lenguaje [26][27][28], permite explorar la morfología interna de los vectores semánticos y proponer un paralelismo bioinformático para la detección de polarización.

De esta manera, el artefacto no solo compara el rendimiento de modelos avanzados bajo esquemas *zero-shot* y *few-shot*, sino que también introduce una perspectiva computacional inspirada en la biología molecular para el análisis de las dinámicas lingüísticas.

Dado que todos los artículos son de fuente revista, se realizó el resumen de su distribución teniendo en cuenta el ranking Scimago representado en la tabla **3-5** y su distribución porcentual se registra en la figura **1-2.**

Tabla 3-5.: Resumen distribución revistas (Elab. Propia).

Clasificación de revistas	No. de artículos	% de artículos

Q1	10	77%
Q2	3	23%
Q3/Q4	0	0%

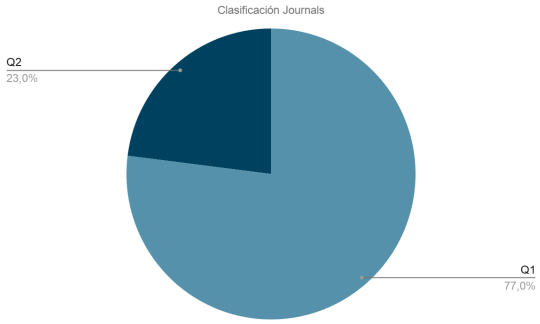


Figura 1-2.: Clasificación de revistas (Elab. Propia).

A partir de la información recopilada en las tablas **3-1**, **3-2** y **3-4**, se cuenta con un referente actualizado de los conceptos y técnicas empleados para resolver desafíos similares. A continuación se desarrollan teóricamente los conceptos que serán utilizados e implementados en la presente investigación para la construcción del artefacto.

Desde una perspectiva metodológica, el enfoque bioinspirado planteado en este trabajo se traduce en un esquema de análisis multinivel que integra diferentes tipos de información lingüística, emulando procesos cognitivos humanos de comprensión y evaluación del discurso. Mientras que las representaciones semánticas profundas permiten capturar el significado global de los textos, la incorporación de rasgos lingüísticos de carácter superficial y estructural posibilita modelar patrones asociados a la forma en que la

polarización se manifiesta en el lenguaje. Esta combinación orienta el diseño del artefacto y fundamenta las decisiones metodológicas adoptadas para la construcción del sistema, estableciendo así un puente claro entre los principios teóricos del enfoque bioinspirado y su implementación computacional.

3.3 Polarización

La polarización es el fenómeno central de esta investigación. Se produce cuando una comunidad se divide en dos polos opuestos con poco espacio para posiciones intermedias. Este fenómeno se puede definir con dos conceptos, el primero es el estado (grado de distancia entre polos) y el otro es el proceso (dinámica de separación). Se abordan cuatro tipos relevantes para la comprensión del trabajo: polarización social y política, polarización digital, polarización en redes sociales y polarización multilingüe y multicultural.

3.3.1 Polarización Política y social

El fenómeno de polarización en contextos políticos y sociales hace referencia a cómo los individuos tienden a dividirse en extremos opuestos cuando se genera tanto una discusión política o una discusión de un evento social (protestas, inmigración, globalización, entre otros) creando una fragmentación y un antagonismo entre ellos.

El trabajo *"Online polarization and cross-fertilization in multi-cleavage societies: the case of Spain"* evidencia que algunas plataformas sociales, en contextos multipartidarios

y culturalmente diversos, logran tanto polarizar como incentivar el intercambio de ideologías entre los usuarios. Esto depende más del uso que haga cada usuario más allá del diseño de las plataformas analizadas [5].

3.3.2 Polarización Digital

Este tipo de polarización se refiere al proceso donde los entornos digitales (plataformas, algoritmos de recomendación, motores de búsqueda, foros en línea, entre otros) aportan a la fragmentación de divisiones ideológicas, sociales o culturales. En el trabajo *"Affective polarization in Online Climate Change Discourse Twitter"* los autores demuestran cómo medir la polarización afectiva entre creadores de contenido *(deniers vs believers)* en un período de 100 semanas. Metodológicamente este concepto se apoya en enfoques de análisis de redes sociales, minería de texto y estudios computacionales, logrando así mapear la formación de comunidades polarizadas y entender cómo los patrones en los entornos digitales moldean percepciones y actitudes colectivas [6].

3.3.2.1 Polarización en Redes Sociales

La polarización en redes sociales es una sub variante de la polarización digital pero sucede específicamente en redes sociales (Instagram, Tik Tok, X o anteriormente conocida como Twitter). En el trabajo *"Social media Polarization and echo chambers in the context of Covid-19"* este tipo de polarización se caracteriza por la creación de cámaras de eco y comunidades homogéneas, donde las personas interactúan o se

relacionan con otros usuarios que tengan el mismo pensamiento, de esta manera, se refuerzan sus creencias, patrones e incluso hostilidades hacía los grupos contrarios, generando una brecha ideológica sobre un evento en particular dentro de una misma red social [7].

3.3.2.2 Polarización multilingüe y multicultural

La polarización multilingüe y multicultural surge cuando el fenómeno se amplifica en contextos donde coexisten múltiples lenguas y culturas. Este enfoque reconoce que la polarización no se presenta de manera uniforme en todos los idiomas, sino que puede variar según los aspectos culturales intrínsecos que median en la interpretación del discurso. De esta manera lo que se puede considerar extremo u ofensivo en un idioma en otro puede percibirse de forma distinta debido a esas diferencias lingüísticas y culturales [11].

La importancia de abordar la polarización desde el enfoque multilingüe y multicultural en esta investigación se fundamenta en el carácter de la globalización en plataformas digitales. En redes sociales como X anteriormente conocida como Twitter, los debates superan las fronteras lingüísticas y culturales, lo que demanda crear modelos capaces de detectar patrones de polarización en idiomas como español e inglés y, de poder transferir el conocimiento en ambos contextos [9].

En términos teóricos, este enfoque abre el debate entre dos posturas. Por un lado, la visión que defiende el uso de representaciones semánticas universales (como los embeddings multilingües) para encontrar patrones comunes en varias lenguas. En contraste, la postura contraria subraya la fuerte dependencia cultural y contextual de la polarización, señalando los límites de la transferencia directa entre idiomas [11].

3.3.3 Expresiones de la polarización

Entender la polarización y sus tipologías es solo una parte de su conceptualización; también es necesario situar y explicar las expresiones vinculadas a los discursos polarizados. Estas expresiones se traducen en recursos discursivos que evidencian la distancia, el enfrentamiento o la exclusión entre diferentes comunidades. Dichas expresiones no son uniformes, sino que responden a distintos patrones lingüísticos y culturales, lo que las convierte en un punto clave tanto para el análisis teórico como para el diseño de modelos automáticos de detección [4].

3.3.3.1 Lenguaje Extremo

Esta expresión hace referencia al uso del lenguaje con nociones absolutas, como insultos o adjetivos negativos que intensifican el enfrentamiento entre comunidades. El lenguaje extremo genera un efecto de intensificación emocional; asimismo, la radicalización discursiva en redes sociales amplifica la exposición a comunidades homogéneas, reforzando así la polarización [12].

3.3.3.2 Estereotipos

Los estereotipos hacen referencia a la generalización negativa que se le atribuye a un grupo social, político, religioso o cultural con el objetivo de simplificar o degradar las características comunes que ambos grupos poseen. En investigaciones sobre detección de odio en redes sociales, se ha evidenciado cómo los estereotipos son un recurso discursivo frecuente en el discurso político digital, operando como mecanismo de diferenciación y exclusión grupal [13]..

3.3.3.3 Invalidación

Si bien la invalidación es una expresión también es empleada como una estrategia discursiva que niega o desacredita por completo la legitimidad de la opinión contraria. Esta estrategia reduce la posibilidad del discurso y bloquea el reconocimiento mutuo en interacciones digitales. En el trabajo *"POLAR: A Benchmark for Multilingual, Multicultural, and Multi-Event Online Polarization",* señalan que la invalidación discursiva es una de las estrategias discursivas más comunes en la polarización multilingüe [4].

3.3.3.4 Falta de empatía

La ausencia de empatía en el lenguaje se correlaciona directamente con la propagación de narrativas discriminatorias [4].

3.3.3.5 Deshumanización

La deshumanización es el lenguaje que niega la condición humana del otro, reduciéndolo con categorías que hacen referencia a un estado inferior (animalesca, mecánica, entre otras). Es considerada una de las expresiones más extremas de la polarización y suele estar asociada a dinámicas de odio en línea, por ejemplo se ha documentado cómo la deshumanización digital es particularmente recurrente en contextos de conflicto político y étnico [4].

3.4 Representación del texto

Teniendo en cuenta que la representación del texto es una parte fundamental en el proceso de PLN, debido a que define cómo los modelos computacionales codifican la información lingüística para poder analizarla y clasificarla, es fundamental comprender cómo opera en esta propuesta. Su función es transformar expresiones textuales,en ocasiones ambiguas o simbólicas, en estructuras matemáticas que capturen características semánticas y sintácticas. En esta investigación, la representación de texto es el insumo principal sobre el cual se entrenan los modelos de detección y clasificación

3.4.1 Representaciones clásicas

Las representaciones clásicas de texto hacen referencia a los primeros enfoques dentro del procesamiento automático del lenguaje, este se basa en transformar documentos en vectores numéricos sin contemplar directamente el contexto semántico. Uno de los métodos más utilizados es el modelo de bolsa de palabras *(Bow),* donde cada documento se representa como un vector de frecuencias de palabras, ignorando el orden y el contexto semántico entre ellas.

Existen 4 pasos fundamentales en la creación de un modelo *Bow,* el primero es la definición del vocabulario donde a partir de un corpus $D = \{d1, d2, d3,...., dm\}$ compuesto por M documentos se construye el vector de vocabulario $V = \{v1, v2,..., vn\}$ de tamaño N , que incluye todas las palabras distintas presentes en el conjunto de documentos. El segundo paso es el proceso de vectorización de un documento, donde cada documento d_j se representa con un vector en R^N :

$$v(d_j) = (f_{1j}, f_{2j}, ..., f_{Nj})$$

Donde f_{ij} indica el número de apariciones del término t_i en el documento d_j. El tercer paso es la representación dentro de una matriz de todos los documentos, obteniendo una matriz de términos de documentos A de dimensión $N \times M$:

$$A = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1M} \\ f_{21} & f_{22} & \cdots & f_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ f_{N1} & f_{N2} & \cdots & f_{NM} \end{bmatrix}$$

Donde cada fila corresponde a un término del vocabulario y cada columna a un documento. El último paso, es crear las variantes matemáticas:

Frecuencia binaria: $f_{ij} = \begin{cases} 1 & \text{si el termino} t_i \text{ aparece en } d_j \\ 0 & \text{en otro caso} \end{cases}$

Frecuencia relativa: $f_{ij} = \frac{Número\ de\ apariciones\ de\ t_i\ en\ d_j}{Número\ total\ de\ palabras\ en\ en\ d_j}$

Estas variantes ajustan la representación para mejorar el rendimiento en tareas de clasificación o *clustering*.

3.4.1.1 *Term frequency - Inverse Document frequency* (TF-IDF)

El modelo TF-IDF permite ponderar la importancia de los términos dentro de un documento en relación con un corpus, destacando aquellos con mayor capacidad discriminativa.

La frecuencia de término se define como:

$$TF_{ij} = \frac{f_{ij}}{\sum_{k=1}^{M} f_{kj}}$$

Donde f_{ij} es el número de apariciones del término t_i en un documento d_j y *M* es el número total de términos distintos en el vocabulario. El segundo factor, es la frecuencia inversa de documentos (IDF), mide la relevancia del término respecto a toda la colección de documentos *D:*

$$LDF_i = log\left(\frac{N}{1+n_i}\right)$$

Donde *N* es el número total de colección de documentos *D*, n_i es el número de documentos que contienen al término t_i y se suma el número 1 para evitar la división por cero. Como proceso final el TF-IDF es la multiplicación de los dos factores:

$$TF - IDF_{ij} = TF_{ij} \times IDF_i$$

Esta representación permite identificar términos característicos asociados a discursos polarizados.

3.4.1.2 N-gramas

Los N-gramas a diferencia de los modelos Bow o TF-IDF, logran capturar la secuencia y el contexto local de las palabras dentro de un documento. Este modelo considera la subsecuencia de longitud *n* de palabras o caracteres, por ejemplo, para $n = 2$ en la frase "la polarización política disminuye" los bigramas en este caso serían: ("la polarización"), ("polarización política"), ("política disminuye"). El uso de este modelo es fundamental dentro investigaciones sobre polarización ya que varios estudios

han demostrado que combinando los unigramas y los bigramas en la representación textual, mejora considerablemente el rendimiento de sistemas de clasificación elevando métricas como la precisión y el F1-score [17].

Este modelo funciona de la siguiente manera, dado un texto con una secuencia de palabras w_i: $S = (w_1, w_2, \ldots, w_T)$ donde T es el tamaño del documento, por otro lado, un N-grama es una subsecuencia de longitud *n* expresado de la siguiente forma:

$$(w_i, w_{i+1}, w_{i+2}, \ldots, w_{i+n-1}), \quad 1 \leq i \leq T - n + 1$$

El límite superior es $T - n + 1$, porque si empezáramos más allá de eso, ya no habría suficientes palabras para completar un N-grama de longitud *n*. Como siguiente paso se calcula la probabilidad de ocurrencia de un término con la siguiente expresión:

$$P(w_t \mid w_{t-1}, w_{t-2}, \ldots, w_{t-n+1}) \approx \frac{C(w_{t-n+1}, \ldots, w_t)}{C(w_{t-n+1}, \ldots, w_{t-1})}$$

Donde C(·) es la frecuencia de aparición de la secuencia en el documento o corpus.

3.4.2 Representaciones distribuidas

Las representaciones distribuidas son una forma más avanzada de modelar texto, tienen como objetivo modelar con vectores densos en un espacio continuo, en lugar de vectores dispersos como la representación clásica [18]. La idea principal es que el significado de una palabra se aprende a partir de su contexto de uso (términos que aparecen en entornos similares tendrán vectores cercanos en ese espacio). A diferencia de los modelos vistos como Bow o los N-gramas, las representaciones distribuidas capturan relaciones semánticas y sintácticas entre las palabras, donde cada palabra w_i se representa como un vector de *d* dimensiones expresado como: $w_i \in R^d$.

3.4.2.1 Word2Vec / GloVe

Word2Vec es un modelo de aprendizaje no supervisado que transforma palabras en vectores densos en un espacio continuo. Su principio es que el significado de una palabra se refleja en su contexto (hipótesis distribucional) [19]. Este modelo aprende vectores densos $w_i \in R^d$ para cada *w* de modo que la cantidad de palabras que aparecen en contextos similares tengan vectores cercanos. El modelo se implementa bajo dos variantes: *Continuous bag of words (CBOW),* esta variante predice la palabra objetivo w_t a partir de las palabras de contexto (palabras alrededor), la segunda variante es *Skip-Gram* la cual predice el contexto dado una palabra objetivo $w_{t\pm j}$.

En la variante *Skip-Gram,* en cada posición *t* del corpus se toma la palabra central w_t y se intenta predecir cada palabra de contexto w_{t+j} con $|j| \leq c, j \neq 0$ donde j hace referencia al tamaño máximo de ventana, es decir, el número de palabras anteriores y siguientes de la palabra de contexto y c es la ventana de contexto, la función objetivo de máxima verosimilitud sobre un corpus con longitud T es:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log P(w_{t+j} \mid w_t)$$

Una vez tenemos la función objetivo de máxima verosimilitud, se modela la probabilidad con softmax de la siguiente forma:

$$P(w_O \mid w_I) = \frac{\exp(v'_{w_O} \cdot v_{w_I})}{\sum_{w \in V} \exp(v'_w \cdot v_{w_I})}.$$

Debido a que el denominador recorre todo el vocabulario $|V|$, en la hora de implementar se utilizan aproximaciones como el *negative sampling* , esta aproximación toma un par positivo (w_i, w_o) y optimiza la función de la siguiente forma:

$$\log \sigma(v'_{w_O} \cdot v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n} \left[\log \sigma(-v'_{w_i} \cdot v_{w_I}) \right]$$

Donde $r(x) = 1/(1 + e^{x})$ y k es el número de muestras negativas. Básicamente el *negative sampling* es un entrenamiento de tipo regresión logística para distinguir contextos reales de ruido, es computacionalmente barato (coste O*(k)* por muestra). Por otro lado, la segunda variante del Word2vec es *continous bag of words (CBOW),* donde primero se construye un vector de contexto:

$$h = \frac{1}{2c} \sum_{-c \leq j \leq c, j \neq 0} v_{w_{t+j}}$$

Posterior se predice la palabra central w_{t} con una variante de la formulación softmax pero de manera análoga:

$$P(w_t \mid context) = \frac{\exp(v'_{w_t} \cdot h)}{\sum_{w \in V} \exp(v'_w \cdot h)}$$

La variante CBOW puede entrenarse con *negative sampling*, esta variante también suele ser más rápida y Skip-Gram funciona mejor con palabras raras.

Por otro lado, *Global Vectors (GloVe)* es otro método de representaciones distribuidas que combina la idea de embeddings con factorización matricial. En lugar de entrenarse solo con ventanas locales de contexto (como Word2Vec), GloVe aprovecha la co-ocurrencia

global de palabras en el corpus [20]. Así, dos palabras tendrán vectores cercanos si tienden a aparecer en entornos similares en todo el corpus, no solo en una ventana pequeña. Este método parte de la matriz de co-currencia *X* donde X_{ij} es el número de palabra j que aparece en el contexto *i,* GloVe busca factorizar (aproximadamente) la matriz de $log\ X$ a través de vectores de palabras y contexto, minimizando una función de pérdida ponderada:

$$J = \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} f(X_{ij}) \left(v_i^\top \tilde{v}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

Donde v_i es el vector de palabra *i,* v_j es el vector del contexto *j,* b_i y b_j son sesgos escalares, adicionalmente, se crea una función de ponderación que evita que co-ocurrencias muy grandes dominen la pérdida con parámetros típicos $x_{max} = 100$ y $\alpha = 3/4$ de la siguiente forma:

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{si } x < x_{\max}, \\ 1 & \text{si } x \geq x_{\max}, \end{cases}$$

3.4.2.2 Embeddings contextuales (BERT, mBERT, XLM-R)

Para poder explicar modelos como BERT, mBERT y XLM-R, es necesario comprender primero la arquitectura en la que se basan, dicha arquitectura se le conoce como *Transformer*. Dado un contexto de tokenizado en *T* tokens, cada token x_i se

convierte en un *embedding* inicial $e_i \in R^d$. Cada etapa del *transformer* aplica auto-atención *(self-attention)* y *feed-forward* sobre la secuencia, para el mecanismo de auto atención se construyen tres proyecciones lineales para cada token de la siguiente forma $Q = XW^Q$, $K = XW^K$, $V = XW^V$, donde $X \in R^{T \times d}$ es la matriz de *embeddings* de entrada y $W^Q, W^K, W^V \in R^{d \times d_k}$ son matrices de parámetros. El puntaje de atención para cada token se calcula y se normaliza con *softmax*, obteniendo las atenciones de la siguiente forma:

$$\text{score}(i,j) = \frac{Q_i K_j^\top}{\sqrt{d_k}} \quad \alpha_{ij} = \frac{\exp(\text{score}(i,j))}{\sum_{k=1}^{T} \exp(\text{score}(i,k))}$$

Generando de esta forma un nuevo vector contextualizado para el token *i* de la forma $h_i = \sum_{j=1}^{T} a_{ij} V_j$. En la práctica se usan *multi-head attention* con *H* cabezas, donde cada cabeza realiza atención de manera independiente y se concatenan los resultados de la siguiente manera: $MHA(X) = Concat\,(head_1, head_2, \ldots\ldots\ldots, head_H)\, W^O$.

Bidirectional Encoder Representations from Transformers (BERT) es un modelo de lenguaje basado en la arquitectura *Transformer*, diseñado para aprender representaciones profundas y bidireccionales de texto. Su principal característica es que considera simultáneamente el contexto a la izquierda y a la derecha de cada palabra, lo que permite

generar *embeddings* contextuales mucho más ricos que los de modelos tradicionales como Word2Vec.

El modelo BERT utiliza únicamente la parte encoder de un *Transformer*, la entrada al modelo es una secuencia de tokens x_1, x_2, ..., x_T que transforman *embeddings* de entrada combinando tres componentes. El primer componente son los *tokens embeddings* que hace referencia a los vectores de palabras, el segundo son los *segment embeddings* que indican a qué oración pertenece cada token, y el tercer componente es el *positional embeddings*, donde se codifica la posición de cada token en la secuencia, estos *embeddings* se suman para formar $X \in R^{T \times d}$, que pasa por *L* capas de encoder.

Posterior usamos *MHA* y luego *Feed-Forward Network,* la cual es una red conectada que actúa sobre cada posición de manera independiente usando la siguiente fórmula: $FFN(h) = max(0, hW_1 + b_1)W_2 + b_2$, después de cada *MHA* y *FFN* se aplica *layer normalization* y se suman las conexiones de entrada *(residual connection)* para mejorar el flujo de gradientes $Output = LayerNorm(X + MHA(X))$, después de *L* capas se obtiene una matriz $H \in R^{T \times d}$ que contiene el *embedding* contextual de cada token.

XLM-R es un modelo de lenguaje multilingüe basado en la arquitectura RoBERTa (una versión optimizada de BERT), entrenado en 100 idiomas con un corpus masivo llamado CC-100 [21]. Su objetivo es generar *embeddings* contextuales compartidos entre diferentes idiomas, de modo que expresiones equivalentes en español, inglés o árabe tengan representaciones similares en el espacio vectorial. La entrada al modelo es una secuencia multilingüe $x_1, x_2, \ldots\ldots, x_T$, cada token se convierte en un índice que mapea a un *embedding* inicial $e(x_i) \in R^d$. Al igual que el modelo BERT, XLM-R combina los tres componentes *token embeddings, positional embeddings y segment embeddings,* posteriormente, a cada bloque se le aplica *self-attention* de la siguiente forma:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Donde $Q = XW^Q$, $K = XW^K$, $V = XW^V$ pero aquí los parámetros W^Q, W^K, W^V son compartidos entre idiomas forzando que las representaciones estén en el mismo espacio semántico, después se aplica *FFN* y *layer normalization* y se suman las conexiones de entrada *(residual connection)* como en BERT, después de *L* capas se obtiene una matriz $H \in R^{T\times d}$, estos *embeddings* son multilingües y contextuales.

Por último, el modelo mBERT solo es la versión multilingüe de BERT, tiene exactamente el mismo funcionamiento que XLM-R es decir, la entrada y tokenización, construcción

de los *embeddings* de entrada, bloques de *transformer* y la salida contextualizada, la única diferencia con BERT es que utiliza parámetros compartidos para todos los idiomas.

3.4.3 *Masked Language Models (MLM)*

Los *Masked Language Models (MLM)* son unas de las arquitecturas más influyentes en el procesamiento de lenguaje natural, introducidas con el modelo BERT [22]. Su objetivo principal es entrenar un modelo para predecir palabras que han sido ocultadas o enmascaradas en una oración, utilizando el contexto bidireccional de las demás palabras.

El paso de estos modelos empieza con una secuencia de tokens $X = (x_1, x_2, ..., x_T)$ donde T es la longitud de la secuencia, una fracción aleatoria de tokens (generalmente el 15%) se selecciona y se reemplaza por un token especial [MASK], por ejemplo, se tiene la siguiente entrada "La polarización en redes sociales es [MASK]", el objetivo es que el modelo prediga la palabra original que en este caso puede ser "alta". Posteriormente, cada token se convierte en un vector denso a través de un embedding $E(x_i)$ obteniendo la matriz de entrada: $X = [E(x_1), E(x_2),....., E(x_T)] \in R^{T \times d}$ donde d es la dimensión de *embedding*.

Luego se pasa la secuencia por varias capas de *self-attention* bidireccional produciendo una representaciòn contextualizada H: $H = Transformer(X)$, cada vector h_i contiene no solo de x_i sino del texto completo. Por otro lado, para cada token enmascarado, el modelo calcula la probabilidad de cada posible palabra del vocabulario *V:*

$$P(x_i \mid X_{\setminus i}) = \text{Softmax}(\mathbf{W}\mathbf{h}_i + \mathbf{b})$$

Donde h_i es la representaciòn contextual del token enmascarado, $W \in R^{|V| \times d}$ es la matriz de pesos de salida, b es el sesgo y $X_{\setminus i}$ denota la secuencia sin el token. Por ùltimo, se utiliza entropìa cruzada para comparar la predicciòn con la palabra original, donde M es el conjunto de posiciones enmascaradas y el objetivo es minimizar esta pèrdida para que el modelo aprenda a recuperar los tokens faltantes:

$$\mathcal{L}_{MLM} = -\sum_{i \in M} \log P(x_i \mid X_{\setminus i})$$

3.4.4 *Large Language Models (LLM)*

Son redes neuronales profundas con millones o incluso billones de parámetros, diseñadas para modelar señales para modelar distribuciones de probabilidad sobre secuencias de texto y generar lenguaje de manera coherente a gran escala. Su importancia en investigaciones sobre polarización radica en que captaron patrones semánticos y

pragmáticos complejos en los datos, lo que permite analizar de manera más precisa el contexto ideológico de un texto, detectar sesgos y hasta simular interacciones discursivas para estudiar narrativas contrapuestas. Matemáticamente, los LLM se basan en la arquitectura de transformers [22], la principal tarea es modelar la probabilidad conjunta de una secuencia $x = (x_1, x_2, \ldots\ldots, x_T)$ mediante la descomposiciòn autoregresiva:

$$P(x) = \prod_{t=1}^{T} P(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

Cada probabilidad se realiza usando atenciòn multi-cabeza que computa los pesos de atenciòn, por otro lado, los LLM estàn entrenados para minimizar la funciòn de pèrdida de entropìa cruzada, a diferencia de los modelos MLM, la entropìa en este caso se calcula sobre las posiciones de la secuencia, pero siempre condicionando los tokens anteriores:

$$\mathcal{L}_{LLM} = -\frac{1}{T} \sum_{t=1}^{T} \log P_\theta(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

Esto permite que el modelo aprenda representaciones latentes profundas del lenguaje que capturan significados, dependencias largas y estructuras gramaticales de alta complejidad.

En el contexto de esta investigación, si bien los LLMs constituyen el estado del arte para tareas de clasificación de texto, su uso mediante fine-tuning directo no fue adoptado como estrategia principal por dos razones. Primero, las restricciones computacionales del entorno de ejecución (Google Colab Pro+ con GPU NVIDIA T4) limitaban el ajuste de

modelos de gran escala. Segundo, el diseño del estudio de ablación requería comparaciones controladas entre 18 configuraciones por subtarea, lo que exigía representaciones fijas y reproducibles. Por tanto, se optó por embeddings pre-calculados derivados de arquitecturas basadas en LLMs (DistilUSE Base Multilingual Cased v2), lo que permite aprovechar la capacidad representacional de estos modelos sin incurrir en el costo computacional del fine-tuning.

3.4.5 Enfoque genómico-semántico en *embeddings*

La analogía entre los modelos de representación semántica y las estructuras genómicas no es solo metafórica, sino que descansa en una propiedad matemática compartida: ambos sistemas codifican información compleja mediante la frecuencia y co-ocurrencia de subsecuencias discretas de longitud fija. En bioinformática, los k-mers son subsecuencias de longitud k extraídas con ventana deslizante sobre una secuencia de nucleótidos; su vector de frecuencias permite caracterizar genomas completos sin necesidad de alineamiento global [30]. De manera análoga, en PLN, los n-gramas de tokens son subsecuencias de longitud n extraídas con ventana deslizante sobre texto; su vector de frecuencias ponderado por TF-IDF captura patrones locales de co-ocurrencia léxica que complementan la información semántica global de los *embeddings* contextuales.

La equivalencia formal entre ambas representaciones puede expresarse así: dado un texto tokenizado como secuencia $T = (t_1, t_2, ..., t_{\square})$, un k-mer textual de orden k es cualquier subsecuencia $(t_i, t_{i+1}, ..., t_{i+\square-1})$ con $i \in \{1, ..., n-k+1\}$. El espacio de k-mers posibles

tiene cardinalidad $|V|^k$, donde $|V|$ es el tamaño del vocabulario, lo que es matemáticamente análogo al espacio de k-mers genómicos de cardinalidad 4^k (sobre el alfabeto {A, T, C, G}). Esta equivalencia estructural justifica la aplicación de técnicas de análisis genómico, en particular la representación mediante vectores de frecuencia de k-mers y su codificación TF-IDF, al dominio lingüístico, tal como han demostrado Ji et al. [26], Ren et al. [28] y Maruyama et al. [29] en tareas de clasificación biológica.

En el contexto de la detección de polarización, la hipótesis es que los discursos polarizados exhiben *motifs* léxicos recurrentes, combinaciones específicas de tokens que aparecen sistemáticamente en textos polarizados con independencia del contexto semántico global, de manera análoga a como los motifs genómicos funcionales se repiten en regiones reguladoras del ADN. La representación POLAR+Genoma busca capturar simultáneamente la semántica global (mediante *embeddings* contextuales) y estos patrones locales estructurales (mediante k-mers), generando una representación híbrida más completa. Formalmente un *embedding* puede considerarse una función de mapeo:

$$f : W \rightarrow R^n$$

donde W es el conjunto de palabras y R^n representa el espacio vectorial de dimensión n. Cada palabra $W_i \in W$ se proyecta como un vector $v_i = f(w_i)$ cuyas coordenadas reflejan la coocurrencia estadística de los términos dentro de un corpus. Este principio es análogo a la forma en que las secuencias de ADN pueden representarse mediante vectores de frecuencia o *k-mers*, donde cada subsecuencia de longitud k codifica un

patrón funcional. De este modo, la distancia semántica entre vectores se observa de la siguiente manera:

$$d(v_i, v_j) = 1 - cos\,(v_i, v_j)$$

Dicha distancia es conceptualmente equivalente a las métricas de similitud empleadas en bioinformática para comparar vectores de frecuencia de k-mers, como la similitud de Jaccard o la distancia euclidiana entre perfiles de k-mers, utilizadas para cuantificar divergencias entre secuencias genómicas [30].

En el contexto de esta investigación, el enfoque genómico-semántico permite modelar la polarización textual como una alteración estructural dentro del espacio semántico. Así como las mutaciones o inserciones en el ADN alteran la función biológica, las desviaciones o desplazamientos vectoriales entre regiones semánticas reflejan cambios en el significado, tono o carga ideológica de los discursos [14]. Este paralelismo facilita visualizar la polarización como un fenómeno de "mutación semántica", donde grupos de textos polarizados tienden a formar clusters altamente densos y con baja diversidad léxica, análogos a secuencias genéticas especializadas.

Matemáticamente, esta dinámica puede observarse mediante la reducción de dimensionalidad, por ejemplo, con t-SNE o UMAP donde los vectores se proyectan en espacios bidimensionales y forman agrupaciones semánticas. Si consideramos un conjunto $E = \{v_1, v_2, ..., v_n\}$ de *embeddings* y una matriz de distancia semántica $D_{ij} = d\,(v_i, v_j)$ las regiones de alta densidad representan zonas de "homología

semántica", análogas a genes conservados en secuencias genómicas. En estas zonas, la polarización se identifica mediante métricas como la dispersión o la entropía local del espacio vectorial:

$$H\,(E) = \; - \sum_{i=1}^{m} p(v_i)\, log\, p(v_i)$$

donde una menor entropía indica mayor uniformidad ideológica, es decir, mayor polarización.

3.5 Reducción de dimensionalidad

La reducción de dimensionalidad hace referencia al conjunto de técnicas utilizadas para transformar representaciones o datos en una dimensión menor, preservando la información que los logran caracterizar mejor sin perder su integridad. En el contexto de la presente propuesta, al trabajar con modelos de representación textual como Word2Vec o GloVec, que generan vectores de gran tamaño los cuales se vuelven complicados para curarlos y aún más importante para interpretarlos, al aplicar esta técnica, mejoramos su tratamiento al reducirlos a una menor dimensión donde se pueden interpretar de una manera más sencilla.

Desde un punto matemático, la reducción de dimensionalidad busca encontrar una proyección *W* que minimice la pérdida de información entre los datos originales *X* y su

representación reducida $\acute{X}$. En investigaciones sobre polarización, la reducción ayuda a la interpretación y visualización de clusters ideológicos, mejorando el rendimiento de clasificación al eliminar el ruido y redundancia en los datos [12].

3.5.1 Análisis de Componentes principales (ACP)

El análisis de componentes principales (ACP), es una técnica estadística que permite transformar un conjunto de variables que posiblemente estén correlacionadas, en un nuevo conjunto de variables no correlacionadas denominadas como componentes principales. Cada una de estas componentes representa una combinación lineal de las variables originales y está ordenada teniendo en cuenta qué tanto cada componente principal puede explicar la mayor varianza.

El ACP es muy útil para las representaciones vectoriales de texto ya que ayudan a reducir el ruido y mantener las dimensiones más informativas, de esta forma, esta técnica ayuda a identificar patrones de manera más sencilla, como la agrupación de discursos ideológicamente cercanos o la separación de comunidades polarizadas [12].

Matemáticamente consiste en encontrar los autovectores y autovalores de la matriz de covarianza de los datos: $\Sigma = \frac{1}{n} X^T X$, donde X es la matriz de datos centrada en la media, posterior, se resuelve el problema de descomposición central: $\Sigma v = \lambda v$, donde v es un

autovector (dirección del componente principal) y λ es su autovalor correspondiente (varianza explicada por ese componente). Finalmente, se proyectan los datos en los *k* autovectores de mayor autovalor para obtener la representación reducida: $X_{reducido} = XV_k$, donde V_k es la matriz que contiene los *k* autovectores principales como columnas.

El ACP no solo optimiza el tiempo y los recursos computacionales, también nos permite visualizar los datos en 1, 2 o 3 dimensiones, reduciendo su complejidad de manera drástica, esta técnica también es valiosa para la interpretación de cómo se distribuyen los discursos y verificar los patrones de polarización de forma más intuitiva [12].

3.5.2 t-SNE y UMAP (para visualización de polarización en *embeddings*)

Otros dos métodos conocidos para la reducción de dimensionalidad son el *t-Distributed Stochastic Neighbor Embedding* (t-SNE) y *Uniform Manifold Approximation and Projection* (UMAP). Por un lado, el t-SNE es un algoritmo de reducción de dimensionalidad no lineal que busca preservar las relaciones de los vecinos entre puntos. Para realizar esto, convierte las distancias en alta dimensión entre pares de puntos $x_i\ x_j$ en probabilidades de similitud p_{ij} definidas de la siguiente forma:

$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)}, \quad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

Luego, para la baja dimensión, define probabilidades q_{ij} usando una distribución *t-Student* de la siguiente manera:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

El objetivo de esta técnica es minimizar la divergencia de Kullback - Leibler entre las distribuciones *P* y *Q* con la siguiente ecuación*:*

$$C = KL(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

De esta manera se produce un mapa donde los puntos cercanos al espacio original permanecen cercanos en la proyección, revelando así, clústers de textos con similitud semántica. Por otro lado, el método UMAP también es una técnica no lineal, sin embargo, se basa en la teoría *manifold learning* y en conceptos de topología algebraica. Primero, construye un grafo de proximidad en el espacio original, usando funciones de afinidad basadas en distancias, luego, optimiza una proyección de dimensión baja que minimiza la divergencia cruzada entre la distribución de distancias entre baja y alta dimensión:

$$C = \sum_{(i,j)} \left[w_{ij} \log \frac{w_{ij}}{w'_{ij}} + (1 - w_{ij}) \log \frac{1 - w_{ij}}{1 - w'_{ij}} \right]$$

Donde w_{ij} es el peso de la arista en el grafo de alta dimensión y w'_{ij} es el peso en la proyección. A diferencia de la técnica t-SNE, tiende a preservar mejor la estructura global del espacio y es más eficiente computacionalmente, lo que la hace ideal para analizar grandes corpus de textos polarizados [4].

3.6 Aprendizaje automático y Agrupamiento

El aprendizaje automático *(Machine Learning)* es un conjunto de métodos que permite a los sistemas computacionales mejorar su desempeño a través de tareas específicas sobre los datos en los que se está trabajando, sin necesidad de ser programados en situaciones explícitas. Dentro de la presente propuesta, estos métodos son cruciales ya que los patrones de lenguaje polarizado no son estáticos, cambian con el tiempo, el lenguaje y la cultura [4].

3.6.1 Algoritmos de clasificación supervisada

El aprendizaje automático abarca dos grandes paradigmas: supervisado y no supervisado. Esta investigación se centra en el aprendizaje supervisado , en el que el modelo se entrena con etiquetas (por ejemplo, textos clasificados como “polarizados” o

"no polarizados"). Estos algoritmos resultan fundamentales, dado que permiten clasificar automáticamente textos según su nivel de polarización y el tipo de discurso que representan.

El objetivo de estos algoritmos, es aprender una función de decisión $F : X \rightarrow Y$ que mapee un conjunto de características *X* a un conjunto de etiquetas *Y.* El proceso parte de un conjunto de entrenamiento $D = \{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)\}$ donde $x_i \in R^d$ son los vectores característica y $y_i \in \{1, \ldots\ldots, k\}$ son las clases. El algoritmo busca encontrar una función $f_\Theta(x)$ parametrizada por Θ que minimice la función de pérdida como la entropía cruzada.

3.6.1.1 SVM, *Random Forest*, Redes Neuronales

Dentro de los algoritmos de aprendizaje supervisado, existen tres que se resaltan por su eficacia y por ser usados en trabajos como la presente propuesta , estos son *Support Vector Machine* (SVM), Bosques aleatorios o *Random Forest* y Redes Neuronales Artificiales (ANN). Por un lado, las SVM buscan encontrar una hiperplano óptimo que separe las clases en un espacio de características de manera que la distancia o el margen entre las clases sea máxima [9]. El hiperplano se define de la siguiente manera: $w^T x + b = 0$ donde *w* es el vector de pesos, *b* es el sesgo y *x* el vector de características. El objetivo de la técnica SVM es resolver la siguiente fórmula:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{sujeto a } y_i(w^\top x_i + b) \geq 1, \ \forall i$$

Donde $y_i \in \{-1, +1\}$ es la etiqueta en cada punto, en algunos casos, se utilizan datos que no pueden ser separados linealmente, en estas situaciones el modelo utiliza funciones de Kernel ($K(x_i, x_j)$) que proyectan los datos a un espacio de mayor dimensión donde si son separables. Esta técnica resulta especialmente eficaz para clasificar textos de manera binaria, por ejemplo, en un discurso polarizado versus uno neutral [9].

Por otro lado, un *Random Forest* es un conjunto de árboles de decisión entrenados sobre diferentes subconjuntos de datos y características. La clasificación se obtiene mediante votación mayoritaria de los árboles [13]. Para su funcionamiento, primero, si se tienen *T* árboles $h_1, h_2, h_3, \ldots\ldots, h_T$ la predicción para una observación *x* es: $\widehat{y} = mode\{h_1(x), h_2(x), h_3(x), \ldots\ldots, h_T(x)\}$. El *Random Forest* permite manejar grandes cantidades de características textuales, por ejemplo, N-gramas o *embeddings* y detectar patrones no lineales en narrativas ideológicas [13].

El otro algoritmo supervisado hace referencia a las redes neuronales artificiales (ANN), son modelos compuestos por capas que se le denominan neuronas, las cuales transforman datos mediante funciones de activación no lineales. El modelo aprende minimizando una

función de pérdida como la entropía cruzada, de esta manera se optimiza el gradiente descendente [18]. El funcionamiento del algoritmo empieza dada una entrada *x*, por otro lado, la salida de una red L capas se calcula como: $a^{(l)} = \sigma(W^{(l)}a^{(l-1)} + b^{(l)}),\ l = 1, 2,..., L$, donde $W^{(l)}$ son los pesos, $b^{(l)}$ los sesgos y $\sigma(\cdot)$ la función de activación. Para poder lograr la clasificación, la pérdida se define como: $L = -\sum_{i=1}^{N} y_i log\,(\widehat{y}_i)$ y se actualizan los parámetros mediante la ecuación: $\Theta \leftarrow \Theta - \eta\nabla_{\Theta} L$ donde η es la tasa de aprendizaje.

Las ANN permiten capturar relaciones complejas entre características lingüísticas y semánticas, especialmente cuando se usan *embeddings* de palabras o modelos pre entrenados como BERT [9][11].

4. Conjunto de Datos

El presente estudio emplea como insumo principal el conjunto de datos proporcionado por la competencia SemEval, diseñado específicamente para tareas de análisis de polarización y postura en contextos multilingües. Este dataset constituye la base metodológica sobre la cual se estructuran las tareas de clasificación abordadas en esta investigación.

Dentro de los recursos suministrados por la competencia se encuentra el archivo denominado *Trial_Data.csv*, el cual contiene 339 registros correspondientes a mensajes provenientes de interacciones digitales. Cada instancia está anotada manualmente y organizada en columnas que responden a las sub tareas definidas por la competencia, relacionadas con diferentes dimensiones de polarización discursiva.

Dado que el tamaño del conjunto de prueba resulta limitado para el entrenamiento de modelos basados en arquitecturas profundas como BERT, este archivo no se emplea como insumo principal de aprendizaje. En su lugar, se utiliza como conjunto de evaluación y referencia, permitiendo contrastar el desempeño del modelo entrenado sobre el corpus ampliado construido en esta investigación.

Las columnas del dataset están estructuradas para responder a las tres sub tareas que se alinean con el objetivo general del estudio. La descripción detallada de cada columna, junto con su significado conceptual y su correspondencia con cada sub tarea, se presenta en la tabla **4-1.**

Tabla 4-1.: Relación entre columnas del dataset y las sub tareas (Elab. propia).

Columna	Descripción / Referencia	Sub tarea
Text	Texto original en inglés o español. Es el insumo principal para todos los modelos.	Insumo general (todas)
Lang	Código del idioma (en o es). Permite análisis multilingüe y escenarios cruzados.	Insumo general (todas)
Polarization	Indica si es polarizante o no *(True o False).*	Subtarea 1: Clasificación binaria (presencia/ausencia de

		polarización)
Gender/Sexual	Marca si el texto se dirige a género o sexualidad (0/1).	Subtarea 2: Clasificación de objetivos
Political	Marca polarización política (0/1).	Subtarea 2: Clasificación de objetivos
Religious	Marca polarización religiosa (0/1).	Subtarea 2: Clasificación de objetivos
Racial / Ethnic	Marca polarización racial o étnica (0/1).	Subtarea 2: Clasificación de objetivos
Other	Marca otros tipos de polarización no incluidos en las categorías anteriores (0/1).	Subtarea 2: Clasificación de objetivos
Vilification	Identifica si hay vilificación o ataque directo en el texto (0/1).	Subtarea 3: Clasificación de formas lingüísticas
Extreme Language	Detecta uso de lenguaje extremo o radical (0/1).	Subtarea 3: Clasificación de formas lingüísticas
Stereotype	Marca si el texto reproduce estereotipos (0/1).	Subtarea 3: Clasificación de formas lingüísticas
Invalidation	Señala si el texto invalida opiniones o identidades (0/1).	Subtarea 3: Clasificación de formas lingüísticas
Lack of Empathy	Indica ausencia de empatía en la expresión (0/1).	Subtarea 3: Clasificación de formas lingüísticas
Deshumanitation	Marca expresiones que deshumanizan a un grupo o persona (0/1).	Subtarea 3: Clasificación de formas lingüísticas

ID	Identificador único de la muestra de texto.	Insumo general (todas)

4.1 Análisis exploratorio SemEval

El análisis exploratorio del corpus se desarrolló mediante un conjunto de procedimientos textométricos y lingüísticos orientados a caracterizar la naturaleza y complejidad del dataset, considerando únicamente los textos en español e inglés. Inicialmente, se aplicaron técnicas de frecuencia léxica y diversidad léxica para identificar los términos más representativos y estimar la riqueza del vocabulario mediante métricas como la razón tipo-token (TTR) y la densidad léxica.

Posteriormente, se calcularon estadísticas descriptivas del corpus (longitud promedio de oraciones, número de palabras y desviaciones estándar) con el fin de evaluar la complejidad textual y las posibles diferencias estructurales entre ambos idiomas. Estas métricas permiten comprender el grado de variación léxica y sintáctica que deberán afrontar los modelos de clasificación durante las etapas de entrenamiento y evaluación.

Adicionalmente, se implementó un análisis morfosintáctico mediante la librería *spaCy*, con el propósito de examinar la distribución de las categorías gramaticales presentes en los textos (sustantivos, verbos, adjetivos, adverbios, entre otros). Este procedimiento permitió obtener una visión estructural del lenguaje empleado, evidenciando los patrones predominantes en la expresión de opiniones y actitudes. A continuación se exponen los

hallazgos del análisis exploratorio, que se presentan en la figura **1-3**, donde se resume y cuantifica la distribución de texto según el idioma.

```
Distribución de idiomas:
lang
ben    28
arb    21
amh    20
deu    20
eng    20
spa    20
fas    20
hau    20
hin    20
ita    20
nep    20
rus    20
tel    20
zho    20
urd    20
tur    19
mya    10
```

Figura 1-3.: Distribución de idiomas (Elab.propia).

El corpus utilizado para el análisis exploratorio presenta una distribución equilibrada entre idiomas, conformada por 20 textos en inglés y 20 textos en español. Esta proporción permite realizar una comparación preliminar controlada entre ambos conjuntos lingüísticos, garantizando condiciones homogéneas para el análisis de complejidad, diversidad léxica y morfología textual. Aunque el tamaño de la muestra es reducido, resulta adecuado para una fase de prueba o trial, orientada a validar los procedimientos metodológicos antes de su aplicación a un conjunto de datos de mayor escala.

Adicionalmente, se realizó una comparación de diversidad léxica y longitud entre ambos idiomas; los resultados se visualizan en las figuras **1-4** y **1-5** respectivamente.

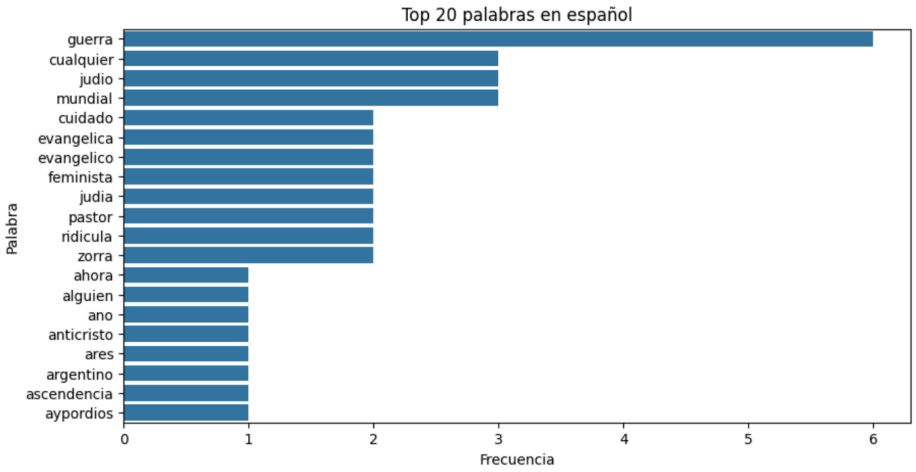


Figura 1-4.: Top palabras en Español (Elab.propia).

En el conjunto de textos en español, las palabras más recurrentes como “guerra”, “judío”, “mundial”, “evangélica” y “feminista” reflejan una clara concentración temática en torno a conflictos religiosos, ideológicos y de género. La presencia de términos con carga valorativa o identitaria sugiere la existencia de discursos polarizados, especialmente relacionados con la religión y los movimientos sociales contemporáneos. Asimismo, palabras como “ridícula” o “zorra” indican la aparición de expresiones de tono peyorativo y despectivo.

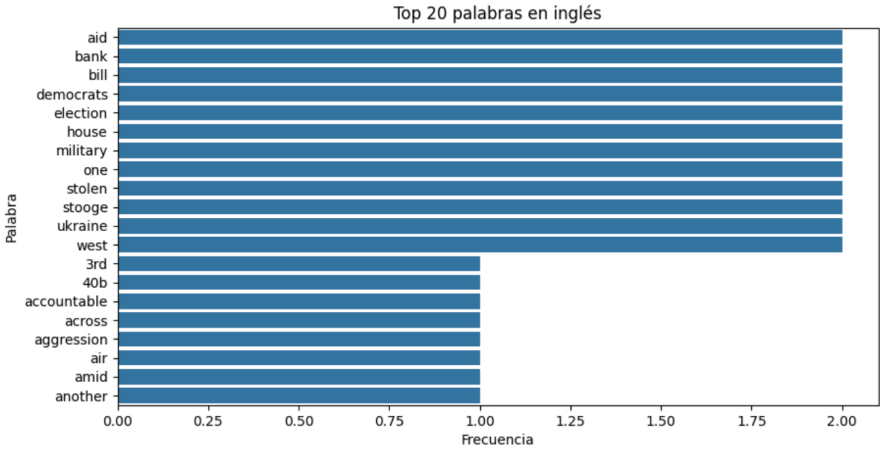


Figura 1-5.: Top palabras en Inglés (Elab.propia).

En los textos en inglés, los términos más frecuentes como *"aid", "bank", "democrats", "election", "military" y "Ukraine"*, evidencian una orientación política y geopolítica en las discusiones, vinculadas principalmente con la guerra, la financiación internacional y los partidos políticos. Este patrón léxico sugiere que los contenidos en inglés tienden hacia debates sobre la polarización política y las tensiones internacionales, más que hacia cuestiones identitarias o religiosas. En conjunto, estas tendencias contrastan con las observadas en los textos en español, mostrando diferencias significativas en las fuentes y manifestaciones de polarización discursiva según el idioma.

Posteriormente, se analizó la diversidad léxica y la longitud promedio del corpus para ambos idiomas, los datos se resumieron en la figura **1-6.**

```
---  Métricas Léxicas Básicas ---
Diversidad léxica (español): 0.6424581005586593
Diversidad léxica (inglés): 0.7817258883248731
Longitud promedio de oraciones (español): 7.357142857142857
Longitud promedio de oraciones (inglés): 10.0
```

Figura 1-6.: Diversidad léxica y longitud promedio (Elab.propia).

En los textos en español se evidencia una diversidad léxica de 0.64, lo que denota una moderada variación del vocabulario, con presencia de repeticiones asociadas a los temas predominantes. La longitud promedio de las oraciones (7.35 palabras) refleja estructuras sintácticas breves y directas, características frecuentes en entornos digitales donde prima la inmediatez comunicativa, como redes sociales o foros de discusión. Este patrón ha sido documentado en análisis de discurso digital, donde la brevedad textual se asocia a estrategias de impacto emocional y posicionamiento ideológico [11].

Por otro lado, los textos en inglés muestran una mayor diversidad léxica (0.78), lo que indica un repertorio lingüístico más amplio y menos repetitivo. Además, la mayor longitud promedio de las oraciones (10 palabras) sugiere una construcción discursiva más compleja, vinculada con la exposición argumentativa y la contextualización política. Este tipo de estructura es común en debates en línea de carácter geopolítico, donde la argumentación busca persuadir o justificar posturas [25]. Para tener un referente de cómo se verían ambos aspectos (diversidad léxica y longitud promedio) se puede observar la gráfica **1-7** que compara ambos idiomas.

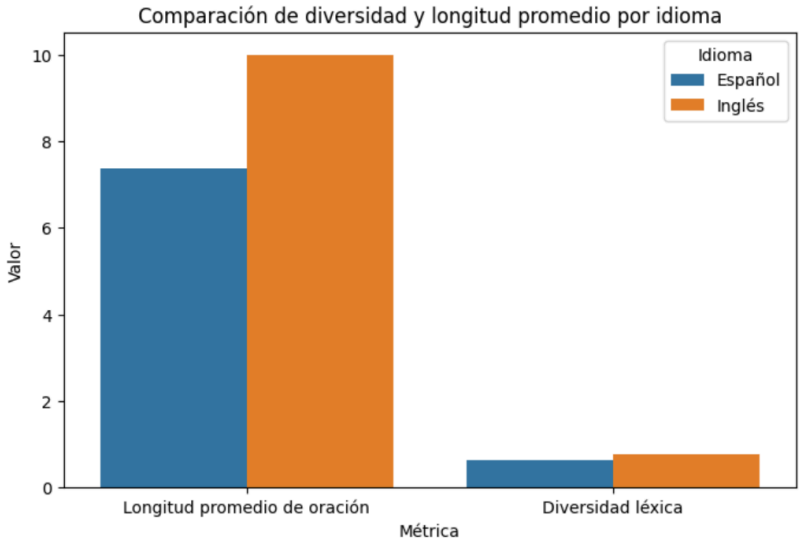


Figura 1-7.: Comparación de diversidad léxica y longitud (Elab.propia).

Complementariamente, se analizó la complejidad textual de ambos idiomas mediante métricas como *Flesch reading ease, Flesch-Kincaid grade, Dale-Chall, Coleman-Liau, SMOG Index,* los datos se resumieron en la figura **1-8.**

```
 Promedios de complejidad textual - ESPAÑOL:
Flesch Reading Ease     51.646067
Flesch-Kincaid Grade     8.099057
Dale-Chall              18.696585
Coleman-Liau             8.193098
SMOG Index               9.263833
dtype: float64

 Promedios de complejidad textual - INGLÉS:
Flesch Reading Ease     48.914494
Flesch-Kincaid Grade     8.921016
Dale-Chall              11.766487
Coleman-Liau            11.264258
SMOG Index              10.324882
dtype: float64
```

Figura 1-8.: Complejidad textual (Elab.propia).

En los textos en español, las métricas de complejidad indican un nivel medio de dificultad lectora. El índice de *Flesch Reading Ease* (51.64) refleja que los mensajes requieren cierto esfuerzo cognitivo, situándose en el rango habitual de textos argumentativos o con carga opinativa. Asimismo, el nivel *Flesch-Kincaid* (8.09) y el índice SMOG (9.26) sugieren que los textos son comprensibles para lectores con educación media o superior, lo que coincide con la naturaleza discursiva y emocional del contenido polarizado en redes sociales. Según Borge-Holthoefer y González-Bailón [5], los discursos polarizados tienden a emplear estructuras lingüísticas más complejas y vocabulario valorativo, lo que incrementa la densidad semántica y la dificultad de lectura.

En el caso del inglés, los resultados muestran una complejidad ligeramente superior, con un *Flesch Reading Ease* de 48.91 y un nivel *Flesch-Kincaid* de 8.92. Esto sugiere textos algo más elaborados, probablemente con oraciones más extensas y vocabulario técnico o político. Las puntuaciones en *Dale-Chall* (11.77) y *Coleman-Liau* (11.26) refuerzan esta tendencia, indicando un estilo más formal y especializado. Este patrón es coherente con lo observado en investigaciones sobre polarización multilingüe, donde los mensajes en inglés tienden a reflejar debates ideológicos con mayor estructuración argumentativa y terminología específica [11].

Por otro lado, se realizó también el análisis morfológico de cada idioma según el corpus, es decir, que tantos pronombres, adjetivos, verbos, son más utilizados dentro de cada idioma, los resultados se ven en las figuras **1-9** y **1-10** respectivamente.

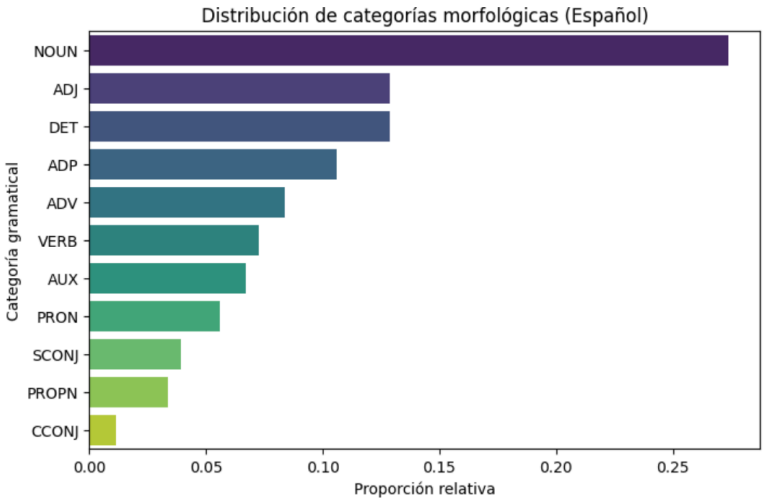


Figura 1-9.: Distribución morfológica Español (Elab.propia).

Los textos en español presentan una estructura morfológica dominada por los sustantivos (27.3%), pero con una mayor proporción relativa de adjetivos (12.8%) y determinantes (12.8%), evidenciando un lenguaje más descriptivo y valorativo. Esta composición refleja la tendencia del discurso en español a incorporar calificativos y matices emocionales en la expresión de opiniones, tal como señalan Borge-Holthoefer y González-Bailón [5] en sus análisis sobre la polarización lingüística en contextos hispanohablantes. La menor frecuencia de verbos (7.2%) y pronombres (5.6%) sugiere un enfoque más nominal y categórico, característico de mensajes que buscan afirmar identidades o emitir juicios más que narrar acciones.

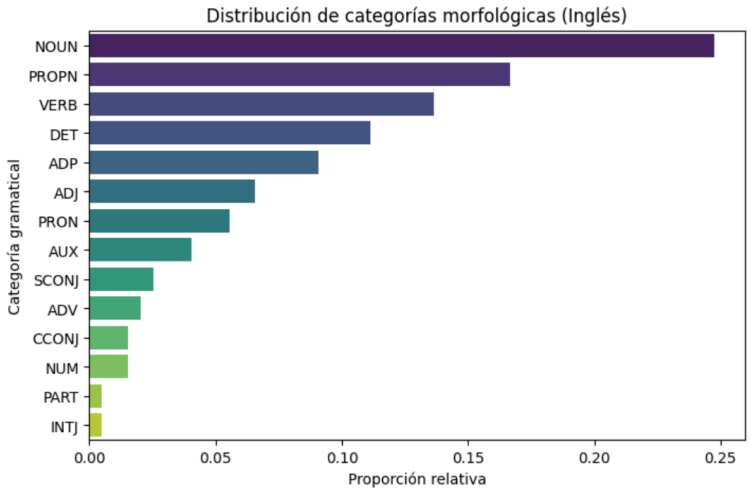


Figura 1-10.: Distribución morfológica Inglés (Elab.propia).

En los textos en inglés, la distribución morfológica muestra una predominancia de sustantivos (24.7%), seguida por nombres propios (16.6%) y verbos (13.6%), lo que indica un uso frecuente de entidades específicas (personas, instituciones o países) y acciones asociadas a eventos concretos. Este patrón sugiere una orientación informativa y temática centrada en actores y hechos, coherente con la naturaleza factual y argumentativa de los discursos polarizados en redes sociales anglófonas. Según Karjus y Cuskley [11], los debates en línea en inglés tienden a estructurarse alrededor de referentes políticos o sociales específicos, lo que explica la alta proporción de sustantivos y nombres propios, así como la moderada presencia de verbos que articulan las posturas discursivas.

En síntesis los resultados evidencian diferencias relevantes entre ambos idiomas en términos de diversidad léxica, complejidad y morfología, reflejando estilos discursivos diferenciados en la manifestación de la polarización. Mientras que los textos en inglés

mostraron mayor diversidad léxica y estructuras más largas y complejas, los textos en español destacaron por un uso más intensivo de sustantivos y adjetivos, lo que sugiere una tendencia hacia un discurso más valorativo y emocional. Estas observaciones coinciden con los hallazgos de Flores[3] y Karjus y Cuskley[11], quienes destacan que las expresiones de polarización varían según las convenciones lingüísticas y culturales de cada idioma.

4.2 Análisis exploratorio unificado

La propuesta inicial del modelo se basaba en el uso del dataset denominado *Trial_Data* de SemEval 2026 (Task 9). Sin embargo, durante el análisis exploratorio se identificó que este volumen resultaba insuficiente para entrenar y validar un modelo de aprendizaje profundo con capacidad de generalización. Con tan pocos registros, las métricas de evaluación (como *F1-score*) fluctúan significativamente entre ejecuciones, lo cual reflejaba una falta de estabilidad estadística y riesgo de *overfitting*. Por lo tanto, se determinó la necesidad de ampliar el conjunto de datos para asegurar una representación más robusta de los patrones lingüísticos asociados a la polarización y el discurso de odio.

Para la expansión del corpus, se seleccionaron dos fuentes externas que cumplían con los criterios de contener texto generado por usuarios en contextos polarizados y que estuvieran disponibles en los idiomas de interés (inglés y español):

English Dataset (HASOC 2019):

El conjunto de datos en inglés proviene de la competencia HASOC en 2019 [31], la cual contiene publicaciones extraídas de Twitter relacionadas con contenido ofensivo, discriminatorio o políticamente polarizado. Este dataset incluyó aproximadamente 5,800 registros con mensajes de carácter social y político, principalmente en inglés.

Toxic_Test Dataset (español):

El segundo conjunto de datos, denominado toxic_test.csv, incluye aproximadamente 1 100 mensajes en español provenientes de redes sociales, con una alta incidencia de lenguaje violento, despectivo o ideológicamente cargado. Este dataset permitió equilibrar la representación del idioma español dentro del corpus final.

Ambos conjuntos se consideraron idóneos porque su estructura (identificador y texto plano) coincidía con los requerimientos del modelo y, al mismo tiempo, presentaban fenómenos lingüísticos comparables a los del dataset oficial de SemEval, lo que favorece la coherencia semántica entre dominios.

Dado que los datasets externos carecían de etiquetas compatibles con las subtareas del reto de SemEval (polarización, vilificación, deshumanización, etc.), se implementó un proceso de etiquetado artificial basado en las directrices propuestas en el artículo *"Detection of Hate Speech, Racism and Misogyny in Digital Social Networks: Colombian Case Study*" publicado en *Big Data and Cognitive Computing* [32].

El método descrito propone la utilización de análisis léxico-semántico supervisado por modelos de lenguaje (LLMs) para asignar etiquetas binarias (0/1) a partir de la detección

de patrones de polarización y expresiones de odio. Basándonos en esta metodología, se desarrolló un *script* de etiquetado automático que, a través de reglas lingüísticas y puntuaciones semánticas, generó las etiquetas correspondientes a las columnas:

Subtask 1: polarization.

Subtask 2: political, racial/ethnic, religious, gender/sexual, other.

Subtask 3: vilification, dehumanization, extreme_language, lack_of_empathy, stereotype, invalidation.

De esta forma, se homogeneizó el formato de los tres conjuntos de datos (HASOC 2019, Toxic Test y Trial Data) bajo la estructura oficial de SemEval 2026, garantizando compatibilidad con el modelo propuesto. Una vez generadas las etiquetas artificiales, se procedió a la unificación de los dos datasets. Para ello, se estandarizaron los nombres de las columnas, se corrigieron inconsistencias en la codificación de caracteres y se normalizó el campo de idioma *(lang)* a los valores "eng" y "spa". Posteriormente, se concatenaron los conjuntos de datos y se eliminaron duplicados y registros corruptos. El resultado fue un dataset final compuesto por 5.237 instancias y 15 atributos, con una distribución balanceada de textos en inglés y español. La Tabla **4-2** sintetiza las características principales de cada uno de los conjuntos de datos empleados, incluyendo su fuente, idioma, número de instancias y rol dentro del estudio. Esta descripción permite comprender la procedencia y el alcance de cada corpus, así como la justificación de su integración en el dataset final utilizado para el entrenamiento y evaluación del sistema.

Tabla 4-2.: Metadata de los datasets utilizados en el estudio (Elab. propia).

Dataset	Fuente	Idioma	Instancias	Columnas principales	Uso del estudio
SemEval 2026 Trial Data	Competencia oficial SemEval	Multilingüe	40	text, lang, polarization, target, manifestation	Validación metodológica y referencia de etiquetas
HASOC 2019	Twitter / Competencia HASOC [31]	Inglés	5.800	text_id, text	Corpus principal en inglés
Toxic Test	Redes sociales en español	Español	1.100	id, text	Corpus principal en español
Unified Dataset Clean Final	Unión HASOC 2019 + Toxic Test	Inglés + Español	5.237	15 columnas (ver Tabla **4-1**)	Dataset final para entrenamiento y evaluación de las tres subtareas

Un aspecto crítico en tareas de clasificación automática es el balance entre las categorías de cada subtarea, dado que el desbalance puede sesgar los modelos hacia las clases mayoritarias y distorsionar las métricas de evaluación. Las Tablas **4-3**, **4-4** y **4-5** presentan la distribución de instancias por clase en el dataset final para cada una de las tres subtareas, junto con el método de balanceo aplicado.

Tabla 4-3.: Balance de clases Subtarea 1 (Elab. propia).

Etiqueta	N de instancias positivas	Porcentaje
Polarizado (1)	1.348	25.74%
No polarizado (0)	3.889	74.26%
Total	5.237	100%

Tabla 4-4.: Balance de clases Subtarea 2 (Elab. propia).

Etiqueta	N de instancias positivas	Porcentaje
political	472	9.01%
racial/ethnic	68	1.30%
religious	42	0.80%
gender/sexual	143	2.73%
other	663	12.66%

Tabla 4-5.: Balance de clases Subtarea 3 (Elab. propia).

Etiqueta	N de instancias positivas	Porcentaje
vilification	323	6.17%
extreme_language	357	6.82%
stereotype	9	0.17%
lack_of_empathy	11	0.21%

invalidation	155	2.96%
dehumanization	13	0.25%

El dataset presenta un desbalance moderado en la Subtarea 1 (ratio 2.89:1) y desbalance severo en las Subtareas 2 y 3, especialmente en categorías como stereotype (0.17%) y lack_of_empathy (0.21%), lo que justifica el uso de *RandomOverSampler* aplicado dentro de cada pliegue de validación cruzada.

Al igual que al corpus de SemEval, se le realizó el mismo análisis exploratorio, solo que con el dataset ya unificado denominado *"Unified_Data_Clean_Final",* en la figura **1-11** y **1-12** se recopilaron las 20 palabras más repetidas para cada uno de los idiomas respectivamente.

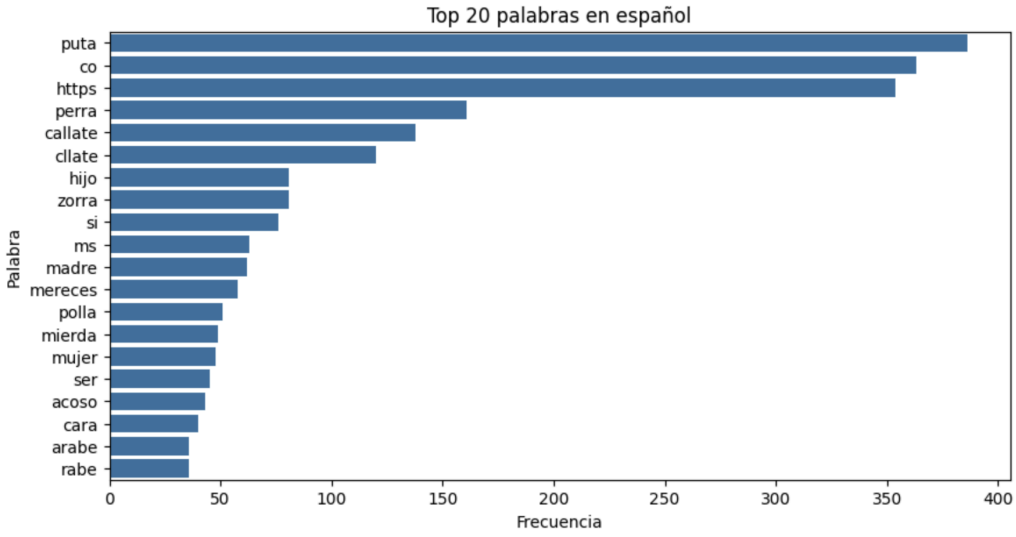


Figura 1-11.: Top palabras en Español dataset unificado (Elab.propia).

Se observa una marcada presencia de vocabulario ofensivo y despectivo, característico de discursos polarizados o de odio. Entre los términos más recurrentes destacan insultos directos como “puta”, “perra”, “zorra”, “mierda” o “polla”, junto con expresiones imperativas agresivas como “cállate”. También aparecen referencias de género (“mujer”, “madre”) y étnicas (“árabe”), lo que evidencia la existencia de sesgos discriminatorios en el discurso. La alta frecuencia de estas palabras sugiere que gran parte del contenido en español del dataset contiene un componente de violencia verbal o estigmatización, especialmente dirigido hacia mujeres o grupos minoritarios.

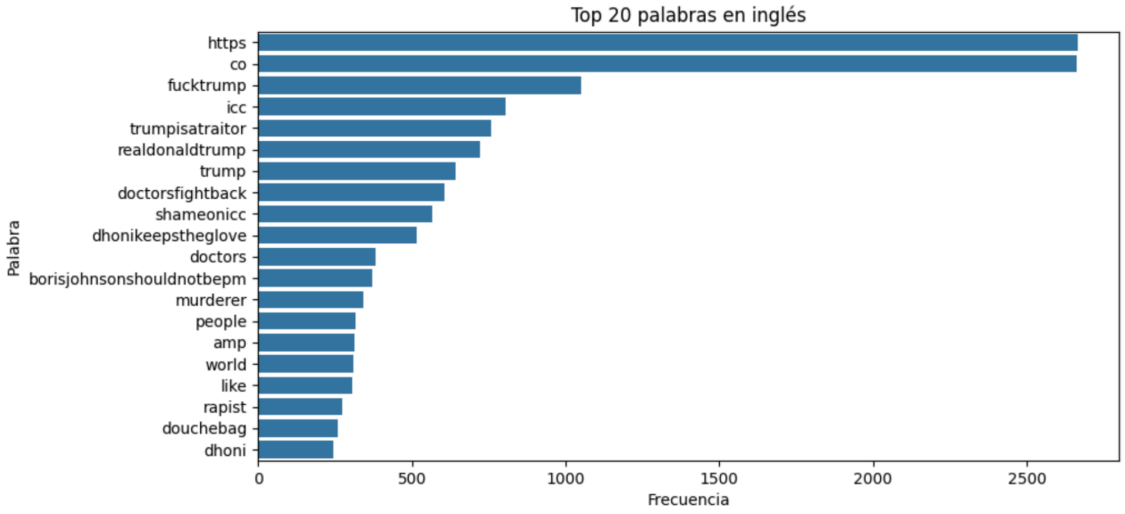


Figura 1-12.: Top palabras en Inglés dataset unificado (Elab.propia).

En este caso, se aprecia un claro sesgo político y social, reflejado en la alta aparición de términos relacionados con figuras públicas y movimientos políticos, como *"fucktrump", "trumpisatraitor", "realdonaldtrump" y "borisjohnsonshouldnotbepm"*. Estas expresiones sugieren una carga emocional y polarización ideológica significativa, especialmente en torno a debates políticos contemporáneos. Asimismo, se observan insultos y descalificaciones como *"douchebag", "murderer" o "rapist"*, que evidencian la presencia de lenguaje ofensivo y agresivo. La frecuencia de *hashtags* y enlaces ("https", "co") indica la naturaleza social del texto, proveniente de plataformas como Twitter. Por otro lado, se realizó el análisis de la diversidad léxica y la longitud del promedio de oraciones de ambos textos resumidos en la figura **1-13.**

```
---  Métricas Léxicas Básicas ---
Diversidad léxica (español): 0.2728645383951682
Diversidad léxica (inglés): 0.1200619736852668
Longitud promedio de oraciones (español): 13.18688901667625
Longitud promedio de oraciones (inglés): 13.73134211351521
```

Figura 1-13.: Diversidad léxica y longitud promedio dataset unificado (Elab.propia).

La diversidad léxica mide la proporción de palabras únicas respecto al total de palabras fue mayor en español (0.27) que en inglés (0.12), lo que sugiere que los textos en español contienen un vocabulario más variado y menos repetitivo. Esto puede deberse a la naturaleza del dataset en español, donde las expresiones de odio y polarización se formulan con mayor variedad lingüística. En cuanto a la longitud promedio de las oraciones, ambos idiomas presentan valores similares (13.18 palabras en español y 13.73 en inglés), lo cual indica una estructura discursiva comparable en la extensión de las

frases. La comparación de diversidad léxica y longitud entre ambos idiomas se visualiza en las figuras **1-14** y **1-15**.

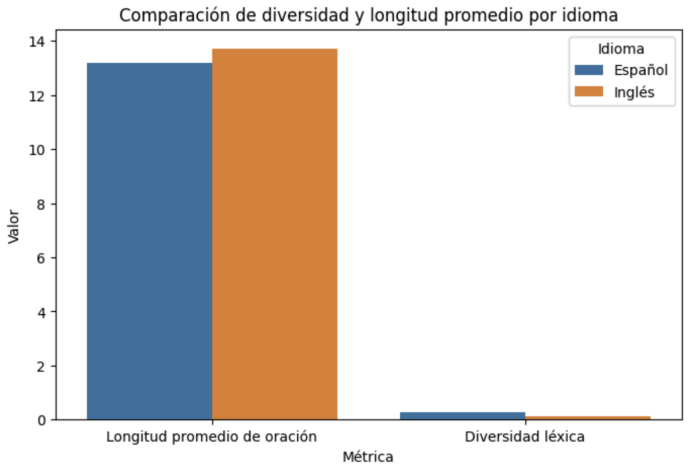


Figura 1-14.: Comparación de diversidad léxica y longitud dataset unificado(Elab.propia).

Se observa que el idioma inglés presenta una longitud media ligeramente superior, lo que sugiere que los mensajes en este idioma tienden a ser más extensos o incluyen construcciones más complejas. En cuanto a la diversidad léxica, ambos idiomas muestran valores bajos y similares, lo que es consistente con la naturaleza de los textos en redes sociales, donde predominan expresiones cortas, repetitivas y con un vocabulario limitado. Estos resultados reflejan que, aunque el inglés contiene mensajes algo más largos, el estilo comunicativo en ambos idiomas mantiene un carácter directo, emocional y poco variado léxicamente, rasgo característico del discurso polarizado en plataformas digitales.

```
 Promedios de complejidad textual - ESPAÑOL:
Flesch Reading Ease     45.124450
Flesch-Kincaid Grade    10.119185
Dale-Chall              19.071042
Coleman-Liau            11.440511
SMOG Index              10.155122
dtype: float64

 Promedios de complejidad textual - INGLÉS:
Flesch Reading Ease     36.455182
Flesch-Kincaid Grade    11.253750
Dale-Chall              12.158471
Coleman-Liau            18.597462
SMOG Index              11.677680
dtype: float64
```

Figura 1-15.: Complejidad textual dataset unificado(Elab.propia).

En los textos en español, el índice *Flesch Reading Ease* (45.12) indica un nivel de lectura medio-difícil, equivalente a educación secundaria, mientras que los valores del *Flesch-Kincaid Grade* (10.11) y del *SMOG Index* (10.15) sugieren que los textos requieren una comprensión lectora equivalente al décimo grado escolar. Por otro lado, el corpus en inglés muestra una mayor complejidad, con un *Flesch Reading Ease* de 36.45 refleja una lectura difícil, y los índices *Flesch-Kincaid Grade* (11.25), *Dale-Chall* (12.15) y *Coleman-Liau* (18.59) confirman un nivel léxico y sintáctico más avanzado. Por su parte, la distribución morfológica de ambos idiomas se presenta en las figuras **1-16** y **1-17** respectivamente.

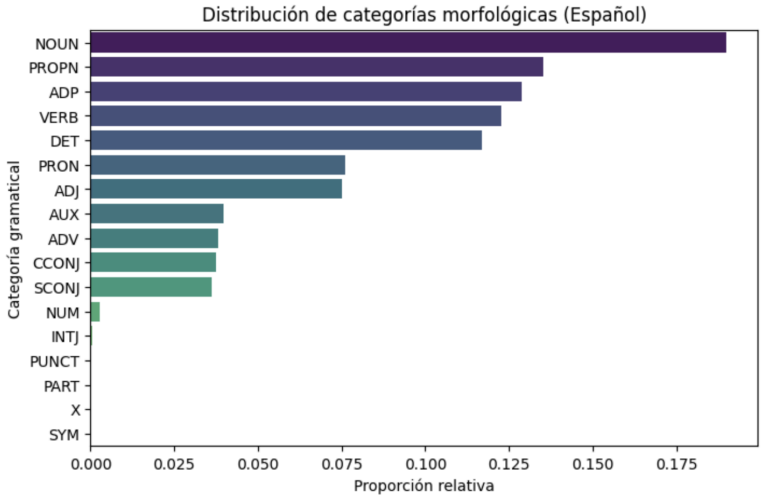


Figura 1-16.: Distrib. morfológica Español dataset unificado (Elab.propia).

Se observa que los sustantivos *(NOUN)* son la categoría predominante, seguidos por los nombres propios *(PROPN)*, las preposiciones *(ADP)* y los verbos *(VERB)*. Este patrón indica que los textos tienden a centrarse en entidades concretas, personas o grupos, y en la descripción de acciones, lo cual es consistente con el tipo de discurso polarizado o de odio que suele referirse a sujetos específicos y juicios categóricos. La presencia relevante de determinantes *(DET)* y pronombres *(PRON)* sugiere un uso frecuente de estructuras directas y personales, reforzando la carga subjetiva del lenguaje. En conjunto, esta distribución refleja un lenguaje orientado hacia la identificación de grupos y la expresión de juicios personales, características propias del discurso hostil o discriminatorio analizado en el corpus.

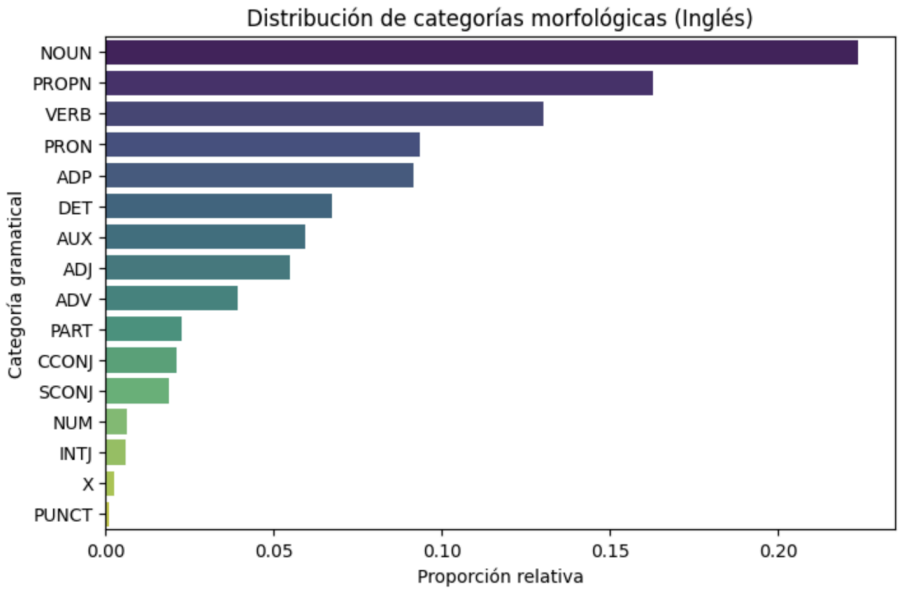


Figura 1-17.: Distrib. morfológica Inglés dataset unificado (Elab.propia).

Se observa un predominio de los sustantivos comunes *(NOUN)* y los nombres propios *(PROPN)*, seguidos de los verbos *(VERB)*, lo que indica una alta densidad de entidades y acciones en el discurso. Esta tendencia es característica de textos de redes sociales, donde los usuarios mencionan con frecuencia personas, instituciones o eventos específicos (por ejemplo, *Trump, Boris Johnson, doctors*) acompañados de verbos de opinión o denuncia. Asimismo, las categorías de preposiciones *(ADP)* y pronombres *(PRON)* también tienen una presencia significativa, lo que sugiere estructuras oracionales directas y personales, propias del lenguaje espontáneo y emocional. En conjunto, esta distribución confirma que el corpus en inglés presenta un discurso altamente nominal y centrado en actores concretos, reflejando la naturaleza polarizada y personalizada del debate digital.

4.3 Preparación de datos y representaciones

El sistema de representación textual combina tres enfoques construidos sobre el corpus unificado. La representación **LLMeval** consiste en vectores de 512 dimensiones generados con el modelo *distiluse-base-multilingual-cased-v2* de la librería *sentence-transformers*, un modelo basado en DistilBERT multilingüe con soporte para más de 15 idiomas, entrenado con *knowledge distillation* desde Universal Sentence Encoder. La representación **Genoma-kmers** aplica TF-IDF sobre k-mers léxicos de longitud k=3, generados con ventana deslizante sobre tokens NLTK, con un vocabulario máximo de 5.000 características y frecuencia mínima de documento de 2. La representación híbrida **POLAR+Genoma**concatena ambas mediante *scipy.sparse.hstack*, resultando en vectores de 5.512 dimensiones que combinan semántica global y patrones estructurales locales.

Para el manejo del desbalance de clases se empleó *RandomOverSampler* de la librería *imbalanced-learn*, aplicado exclusivamente sobre el conjunto de entrenamiento dentro de cada pliegue de la validación cruzada estratificada, con el fin de evitar *data leakage*. Este procedimiento duplica instancias de la clase minoritaria con reemplazo aleatorio hasta equiparar la distribución de clases, garantizando que el conjunto de prueba no sea intervenido y que las métricas reflejen el desempeño real del modelo ante distribuciones naturales.

5. Metodología

El enfoque metodológico en el que se enmarca esta tesis es la investigación científica, basada en el diseño o como se le conoce *Design Science Research (DSR)*. Este enfoque propuesto por Hevner [24], se fundamenta en tres ciclos: relevancia, diseño y rigor. Dicho enfoque es pertinente para la actual propuesta ya que el objetivo central del estudio es la construcción y validación de un sistema computacional para la detección de polarización multilingüe y multi evento, acotado a los idiomas español e inglés.

De acuerdo con la metodología DSR, la investigación se organiza en tres ciclos importantes: i) ciclo de relevancia, que garantiza la pertinencia de la solución frente a un problema real, ii) ciclo de rigor, que articula el diseño con la base teórica y técnica existente en la literatura; y iii) ciclo de diseño, donde se implementa prueba a prueba y refina el artefacto propuesto. Estos ciclos se conciben como engranajes que permiten generar un artefacto útil y científicamente válido, a partir del equilibrio entre la utilidad práctica y el sustento teórico, para poder comprender mejor la utilización de esta metodología se sugiere revisar la figura **1-18.**

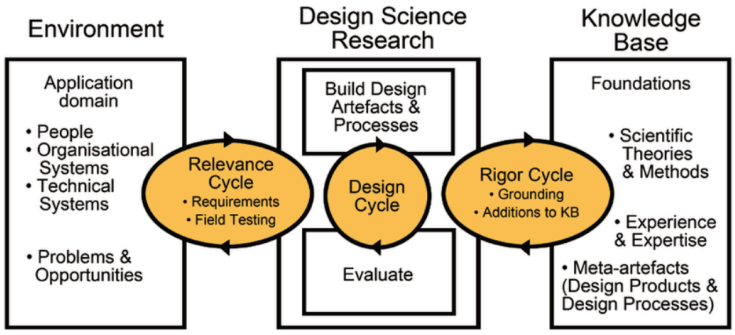


Figura 1-18.: Ciclos en DSR (Tomado de Hevner [24]).

Con la metodología DSR definida, se adaptaron sus tres ciclos como guía estructural para el desarrollo del sistema, tal como se describe a continuación.

5.1 Ciclo de Relevancia

En este estudio el ciclo de relevancia, se fundamentó en la necesidad de comprender y detectar la polarización en línea en contextos multi eventos, multilingües que, en este caso serían los idiomas inglés y español respectivamente. La elección de trabajar con el conjunto de datos previamente explicado y proporcionado por la competencia SemEval, asegura que el problema se absorbe con datos representativos y validados por la comunidad científica, lo que permitió centrar los esfuerzos en el desarrollo y evaluación del sistema de detección, más que en la recolección y tratamiento de los datos.

La revisión del estado del arte permitió identificar y comparar los métodos más utilizados para abordar fenómenos similares al de esta investigación. A partir del análisis de alcance y complejidad presentado en el marco teórico, se seleccionaron las técnicas más apropiadas para el desarrollo del sistema, consolidando así una base metodológica sólida.

Además, se incorporó un enfoque genómico-semántico como extensión del sistema (descrito en la sección 3.4.5), fundamentado en la analogía estructural entre los embeddings lingüísticos y las secuencias de ADN. Este componente habilita la

transformación de vectores densos en secuencias discretizadas analizables como *motifs* semánticos, aportando una capa estructural complementaria a la línea base POLAR + LLM Eval para la detección robusta de polarización en escenarios con manifestaciones discursivas implícitas o lingüísticamente complejas. El ciclo de relevancia se ve representado en la figura **1-19.**

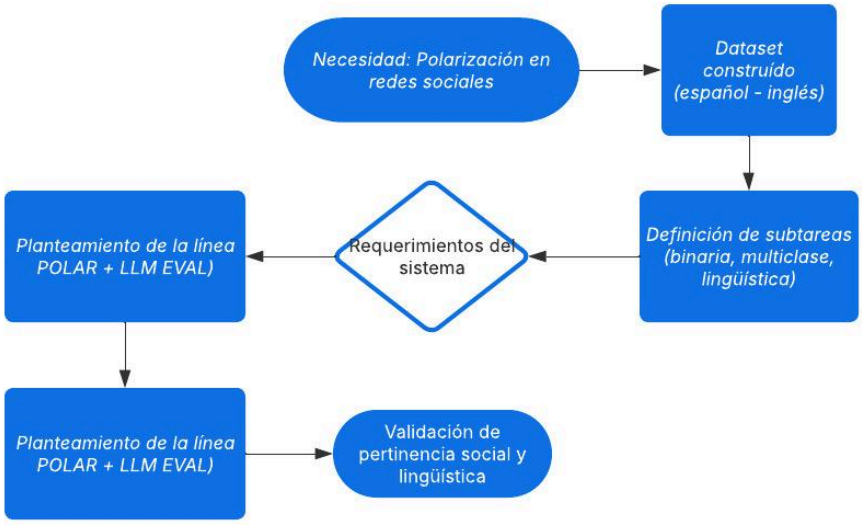


Figura 1-19.: Ciclo de relevancia (Elab. Propia).

5.2 Ciclo de Diseño

En la fase correspondiente al ciclo de diseño representado en la figura **1-20**, se implementó y refinó el artefacto resultante de la investigación, el cual consiste en un modelo híbrido de detección de polarización multilingüe. Esta etapa inició con la arquitectura base POLAR y LLM entrenada con el dataset construído *"Unified_Dataset_Clean_Final"* (inglés-español), abordando subtareas de clasificación binaria, multiclase y lingüística.

Posteriormente, se incorporó un módulo genómico-semántico como extensión experimental del sistema, cuyo propósito fue discretizar los *embeddings* y representar sus patrones internos mediante secuencias tipo k-mer, análogas a la codificación genética, con el fin de explorar una estructura más granular del significado distribuido.

Los resultados obtenidos con esta integración se compararon frente a la línea base a través de validación cruzada y pruebas de ablation, lo que permitió cuantificar su impacto en la precisión y robustez del modelo. Las mejoras observadas orientaron los ajustes de hiper parámetros y la optimización final del artefacto, garantizando una configuración eficiente y generalizable para su evaluación final.

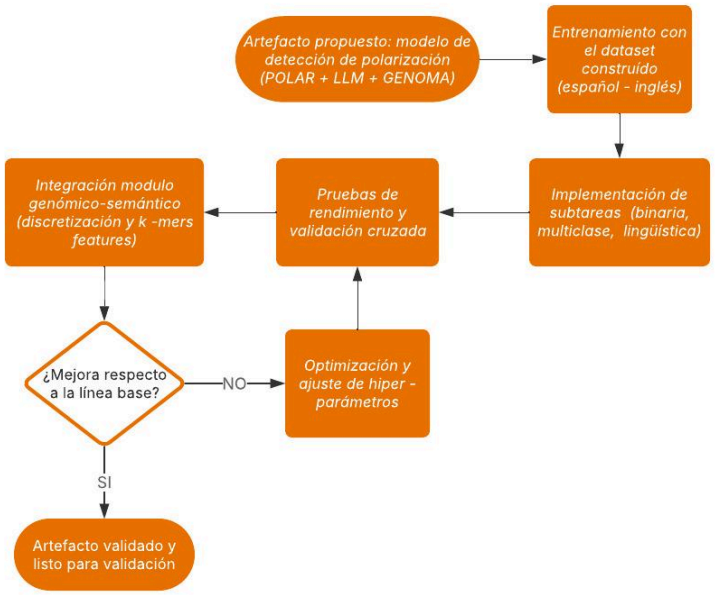


Figura 1-20.: Ciclo de diseño (Elab. Propia).

5.3 Ciclo de Rigor

En el ciclo de rigor representado en la Figura **1-21**, se consolidó la base teórica, metodológica y técnica que sustenta el desarrollo del artefacto. Este proceso inició con la revisión sistemática de modelos de referencia como POLAR, *HybridEvents* y LLM Eval, los cuales permitieron identificar estructuras de aprendizaje multilingüe y estrategias de evaluación relevantes.

A partir de esta revisión se integró un enfoque genómico-semántico, orientado a analizar la estructura interna de los *embeddings* mediante discretización tipo k-mers, análoga a los patrones de codificación genética. Esta incorporación permitió contrastar representaciones lingüísticas tradicionales con una aproximación estructural más granular, fortaleciendo la comprensión del significado distribuido y su potencial para detectar polarización.

Posteriormente, se seleccionaron técnicas de procesamiento multilingüe basadas en transformadores (BERT, XLM-R) y se compararon métricas de desempeño bajo criterios de robustez, escalabilidad y transferibilidad. Finalmente, se establecieron los parámetros de evaluación (accuracy, macro-F1, F1 ponderado), consolidando una base de conocimiento sólida que garantizó la validez metodológica y científica del sistema.

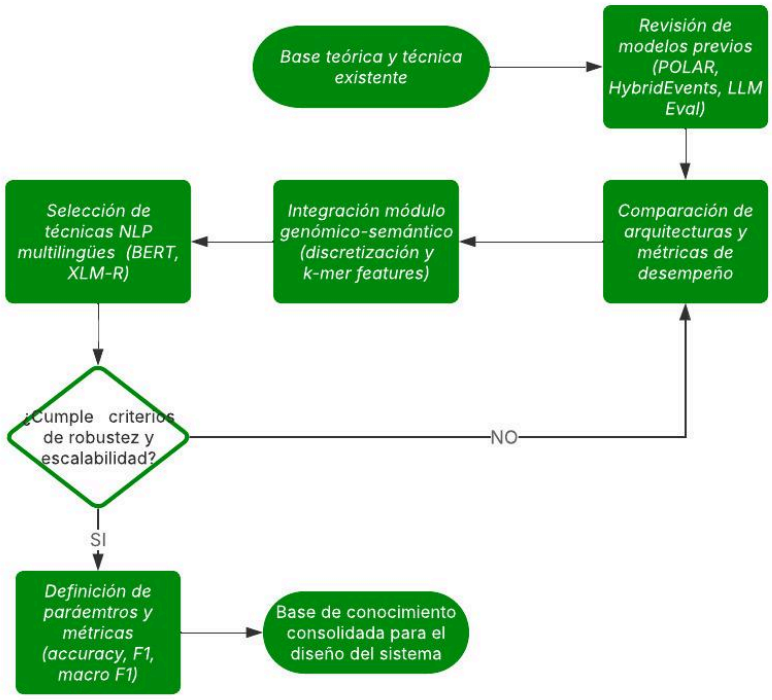


Figura 1-21.: Ciclo de rigor (Elab. Propia).

5.4 Interconexión entre ciclos

Los tres ciclos (relevancia, rigor y diseño) actuaron de manera interdependiente, conformando un proceso metodológico continuo que asegura tanto la pertinencia social como la validez científica del estudio. El ciclo de relevancia definió las necesidades del entorno y los requerimientos funcionales del sistema a partir del fenómeno de polarización en redes sociales. A su vez, el ciclo de rigor garantizó que las soluciones propuestas están fundamentadas en marcos teóricos y metodológicos consolidados, lo que otorga coherencia y respaldo a las decisiones técnicas. Finalmente, el ciclo de diseño tradujo estos insumos en un artefacto tangible que en este caso es el modelo de detección

de polarización, que posteriormente se probó, validó y retroalimentó con los hallazgos empíricos.

De esta manera, los ciclos conforman un sistema dinámico de retroalimentación, donde cada componente impacta y refuerza a los demás. Se puede observar la conexión de los ciclos como unos engranajes de un sistema completo, como se aprecia en la figura **1-22.**

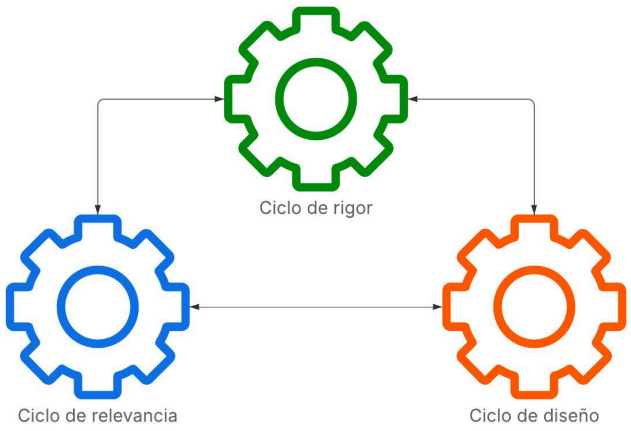


Figura 1-22.: Interconexión entre ciclos (Elab. Propia).

5.5 Actividades macro para el desarrollo de la metodología

La tabla **5-1** recorre las actividades dispuestas para cumplir los objetivos específicos de la investigación. Cada actividad se vincula con el ciclo del diseño al que corresponde.

Tabla 5-1.: Actividades enlazadas al ciclo del diseño (Elab. propia).

Objetivo específico/Actividades Macro	Ciclo
OBJ1: Implementar un modelo de clasificación binaria que determine la presencia o ausencia de polarización en textos en español e inglés.	
ACT1. Seleccionar y analizar el dataset construído a partir de la selección de varios conjuntos de datos, verificando su estructura, balance y viabilidad para escenarios cruzados.	Relevancia
ACT2. Realizar la limpieza y preprocesamiento de los datos (tokenización, normalización, eliminación de ruido, balance de clases y adaptación lingüística).	Diseño
ACT3. Implementar la línea base híbrida POLAR + LLM Eval, entrenando un modelo multilingüe (mBERT o BETO) complementado con la generación de embeddings genómico-semánticos.	Diseño
ACT4. Evaluar el rendimiento del modelo binario mediante métricas de precisión, recall, F1 y accuracy para ambos idiomas, comparando con modelos previos.	Rigor
OBJ2: Diseñar un clasificador multiclase que identifique el objetivo de la polarización en categorías definidas (política, religiosa, racial/étnica, género/sexual y otras).	
ACT1. Analizar la distribución y semántica de las categorías en el dataset, ajustando etiquetas y jerarquías según las definiciones lingüísticas del estudio.	Relevancia
ACT2. Configurar y entrenar un modelo multiclase basado en arquitecturas Transformer (mBERT, XLM-R) con la integración de vectores genómico-semánticos para la representación contextual.	Diseño
ACT3. Comparar el desempeño del clasificador multiclase con modelos de referencia (POLAR, *HybridEvents*, LLM Eval) para validar la robustez y escalabilidad.	Rigor
OBJ3: Desarrollar un sistema de clasificación que reconozca la manifestación lingüística de la polarización (lenguaje extremo, estereotipos, invalidación, falta de empatía, deshumanización, etc.).	
ACT1. Revisar el marco teórico y lingüístico de las manifestaciones de polarización para definir las categorías semántico-discursivas.	Rigor

ACT2. Implementar un clasificador de manifestaciones utilizando embeddings híbridos (contextuales + genómico-semánticos) para capturar matices lingüísticos finos.	Diseño
ACT3. Integrar los tres clasificadores (binario, multiclase y de manifestaciones) en un sistema unificado de detección de polarización.	Diseño
OBJ4: Evaluar el desempeño del sistema sobre un corpus multilingüe unificado (español e inglés) mediante validación cruzada estratificada, analizando la robustez y capacidad de generalización del modelo en escenarios de clasificación binaria, multiclase y multi-label.	
ACT1. Diseñar experimentos que evalúen la generalización de los modelos en ambos idiomas por separado.	Diseño
ACT2. Implementar pruebas de transferencia cruzada (entrenar en un idioma y probar en otro) y analizar la estabilidad de los *embeddings* genómico-semánticos entre lenguas.	Rigor
ACT3. Analizar los resultados comparativos de desempeño y error por clase, destacando las diferencias lingüísticas y estructurales.	Rigor
OBJ5: Analizar los resultados obtenidos con base en métricas como precisión, F1-macro y desempeño por clase, destacando diferencias y retos entre los idiomas.	
ACT1. Consolidar los resultados experimentales y visualizarlos mediante tablas comparativas, mapas de *embeddings* y gráficos de métricas.	Diseño
ACT2. Interpretar los resultados a la luz del marco teórico, identificando patrones estructurales y semánticos en la representación tipo ADN de los *embeddings*.	Rigor
ACT3. Analizar los resultados comparativos de desempeño por representación y arquitectura, identificando patrones de robustez y generalización a través de las tres subtareas.	Relevancia

En síntesis, la metodología de esta investigación se basó en el enfoque *Design Science Research* (DSR), estructurado en los ciclos de relevancia, rigor y diseño. Cada ciclo permitió articular las necesidades del problema, la base teórica y el desarrollo del sistema de detección de polarización. A través de diagramas de flujo y actividades macro, se

definieron las fases de análisis, implementación y validación del modelo, garantizando un proceso metodológico coherente, iterativo y sustentado en la literatura científica.

6. Implementación del artefacto y resultados

6.1 Descripción general del experimento

Los resultados que se presentan a continuación responden directamente al enfoque metodológico propuesto. El experimento tuvo como objetivo evaluar la efectividad de un sistema multietapa para la detección y caracterización de la polarización en comentarios provenientes de redes sociales, comparando de manera sistemática dos paradigmas de modelado: clasificadores supervisados clásicos y arquitecturas de redes neuronales modernas. El sistema fue estructurado siguiendo la taxonomía definida en el Task 9 de SemEval-2026, abordando tres subtareas complementarias: i) identificación de polarización binaria, ii) clasificación del objetivo de la polarización y iii) reconocimiento de la manifestación lingüística.

Desde esta perspectiva, los resultados permiten analizar no solo el desempeño predictivo del sistema, sino también el impacto del enfoque bioinspirado adoptado, basado en la integración de representaciones semánticas profundas y modelado secuencial mediante k-mers textuales, así como la capacidad de las arquitecturas de aprendizaje profundo para capturar patrones discursivos complejos en escenarios multilingües.

Para todos los experimentos se utilizó el dataset unificado "Unified_Dataset_Clean_Final.csv", compuesto por 5.237 instancias en español e inglés.

Los modelos fueron implementados en Python 3.12 y ejecutados en Google Colab Pro+ con aceleración GPU (NVIDIA T4). En las tres subtareas se aplicó validación cruzada estratificada de 5 pliegues y se reportaron las métricas F1 macro promedio, desviación estándar e intervalo de confianza al 95%. Adicionalmente, para complementar la evaluación, se reportan F-scores específicos por etiqueta obtenidos mediante *classification_report* sobre las predicciones acumuladas de todos los pliegues para el mejor modelo de cada subtarea, tanto en su variante clásica como neuronal. Para la Subtarea 1, se incluye también el Área Bajo la Curva ROC (AUC) como métrica complementaria al F1-macro, dado que en tareas binarias el AUC captura la capacidad discriminativa del modelo con independencia del umbral de clasificación. Para la evaluación de significancia estadística entre las 18 configuraciones evaluadas por subtarea, se complementaron las pruebas t de Student pareadas con un procedimiento de comparación múltiple basado en el test de Friedman seguido de comparaciones post-hoc de Wilcoxon con corrección de Bonferroni, con el fin de controlar la tasa de error tipo I derivada de las comparaciones simultáneas.

La decisión de utilizar embeddings pre-calculados en lugar de fine-tuning directo sobre modelos como XLM-R o mBERT responde a dos criterios: las restricciones computacionales del entorno de ejecución (Google Colab Pro+) y la necesidad de garantizar la reproducibilidad del estudio de ablación, el cual comprende 18 configuraciones por subtarea. Este enfoque permite comparaciones controladas y sistemáticas entre representaciones, sin introducir variabilidad derivada del proceso de ajuste de parámetros.

Se evaluaron tres configuraciones de representación: i) LLMeval que serían embeddings contextuales multilingües generados con DistilUSE Base Multilingual Cased v2, ii) Genoma-kmers, una representación estructural basada en k-mers textuales de tamaño k=3 codificados mediante TF-IDF, y iii) POLAR+Genoma, que es la combinación híbrida de embeddings y k-mers mediante concatenación dispersa. Cada representación fue evaluada bajo dos enfoques de modelado complementarios: tres clasificadores supervisados clásicos (LightGBM, Regresión Logística y SVM) y tres arquitecturas de redes neuronales modernas (MLP-Deep con normalización por lotes, BiLSTM bidireccional y TextCNN con convolución sobre vectores), resultando en un total de 18 combinaciones por subtarea. El objetivo fue analizar el impacto individual y combinado de las representaciones semánticas y secuenciales, así como determinar qué paradigma de modelado ofrece mayor robustez y generalización para cada nivel de complejidad discursiva.

6.1.1 Descripción del Subtask 1: Identificación de polarización binaria

La primera subtarea consistió en clasificar cada mensaje como polarizado (1) o no polarizado (0). Para esta tarea se implementó un pipeline basado en embeddings multilingües generados con el modelo *DistilUSE Base Multilingual Cased v2*, junto con una representación estructural inspirada en modelado genómico mediante k-mers textuales.

En la representación Genoma-kmers, cada texto fue segmentado mediante una ventana deslizante de tamaño k=3 sobre tokens, generando subsecuencias consecutivas (k-mers). Estas secuencias fueron posteriormente codificadas mediante ponderación TF-IDF, permitiendo capturar patrones locales estructurales dentro del discurso.

Se evaluaron tres configuraciones de representación en un estudio de ablación: i) LLMeval (*embeddings* únicamente), ii) Genoma-kmers (k-mers únicamente) y iii) POLAR+Genoma (fusión de embeddings y k-mers). Cada configuración fue evaluada bajo dos enfoques de modelado complementarios. En primer lugar, tres clasificadores supervisados clásicos: LightGBM, Regresión Logística y Support Vector Machine (SVM). En segundo lugar, tres arquitecturas de redes neuronales modernas: un Perceptrón Multicapa profundo con normalización por lotes (*MLP-Deep*), una red recurrente bidireccional (*BiLSTM*) y una red convolucional sobre vectores de texto (*TextCNN*). Esta combinación permitió analizar tanto la interacción entre tipo de representación y modelo de clasificación, como la capacidad de las arquitecturas profundas para capturar patrones no lineales en el discurso polarizado.

El proceso experimental incluyó sobre-muestreo de la clase minoritaria mediante *RandomOverSampler* y validación cruzada estratificada de 5 pliegues. El desempeño fue evaluado mediante F1 macro promedio, desviación estándar e intervalo de confianza al 95%, garantizando una estimación robusta del rendimiento. Para el mejor modelo de cada paradigma se reporta adicionalmente el F-score por clase (No polarizado / Polarizado) y el AUC promedio sobre los cinco pliegues, con su respectivo intervalo de confianza al

95%. Adicionalmente, se realizó una prueba t de Student pareada para evaluar la significancia estadística entre pares de modelos, complementada con el test de Friedman y comparaciones post-hoc de Wilcoxon con corrección de Bonferroni sobre las 18 configuraciones evaluadas simultáneamente.

6.1.2 Descripción del Subtask 2: Clasificación del objetivo de la polarización

En la segunda subtarea, el sistema debía identificar el grupo objetivo hacia el cual se dirige la polarización, considerando cinco categorías: política, religión, etnia/raza, género/sexualidad y otros. Dado que algunas instancias podían presentar múltiples etiquetas activas, el problema fue reformulado como clasificación multiclase dominante, asignando a cada mensaje la categoría con mayor valor entre las variables correspondientes. Esto permitió establecer un marco experimental consistente y evaluable bajo validación cruzada estratificada.

Se implementaron las mismas configuraciones experimentales descritas en el Subtask 1, manteniendo el esquema de *ablation study* con tres representaciones: LLMeval (embeddings), Genoma-kmers (TF-IDF sobre k-mers léxicos, k=3) y la representación híbrida LLMeval+Genoma. Cada configuración fue evaluada bajo dos enfoques complementarios: tres clasificadores supervisados clásicos (LightGBM, Regresión Logística y SVM lineal) y tres arquitecturas de redes neuronales modernas (MLP-Deep, *BiLSTM* y *TextCNN*), junto con balanceo mediante *RandomOverSampler* y validación cruzada estratificada de 5 pliegues.

El desempeño se reporta mediante F1-macro promedio, desviación estándar e intervalo de confianza al 95%. Para el mejor modelo de cada paradigma se reporta adicionalmente el F-score desglosado por categoría (political, racial/ethnic, religious, gender/sexual, other), lo que permite identificar cuáles objetivos de polarización resultan más difíciles de clasificar. Adicionalmente, se aplicó una prueba t de Student pareada sobre los valores obtenidos en cada pliegue, complementada con el test de Friedman y comparaciones post-hoc de Wilcoxon con corrección de Bonferroni sobre las 18 configuraciones evaluadas de manera simultánea.

Adicionalmente, se aplicó una prueba t de Student pareada sobre los valores obtenidos en cada pliegue para evaluar la significancia estadística de las diferencias entre todos los modelos evaluados. Esta subtarea presenta mayor complejidad semántica que la clasificación binaria, ya que requiere no solo detectar polarización, sino también inferir su objetivo específico, lo cual incrementa la ambigüedad contextual y la dispersión en el espacio de representación.

6.1.3 Descripción del Subtask 3: Reconocimiento de la manifestación lingüística

La tercera subtarea abordó el nivel más complejo del modelo, orientado a identificar cómo se manifiesta lingüísticamente la polarización en los mensajes. A diferencia de las dos subtareas anteriores, un mismo texto podía presentar múltiples formas discursivas de manera simultánea, por lo que el problema se formuló como una

tarea de clasificación *multi-label*. Las categorías consideradas fueron: vilificación, lenguaje extremo, estereotipos, invalidación, falta de empatía y deshumanización.

Para modelar este escenario con los clasificadores clásicos se implementó un esquema de *MultiOutputClassifier*, entrenando un clasificador independiente por cada etiqueta y generando predicciones paralelas. Para las redes neuronales, se implementó una capa de salida con *NUM_LABELS* neuronas y función de pérdida *BCEWithLogitsLoss*, permitiendo predicciones simultáneas por etiqueta mediante un umbral de 0.5 sobre la función sigmoid. Se evaluaron tres algoritmos clásicos (LightGBM, Regresión Logística y SVM) y tres arquitecturas de redes neuronales (*MLP-Deep, BiLSTM y TextCNN*) sobre tres tipos de representación: embeddings semánticos profundos (*LLMeval*), características estructurales basadas en k-mers del Genoma lingüístico, y la representación híbrida LLMeval+Genoma.

La evaluación se realizó mediante validación cruzada estratificada de 5 particiones, utilizando como variable de estratificación la polarización binaria del mensaje para preservar la distribución general del fenómeno en cada fold.

La métrica principal fue F1 macro promedio entre etiquetas, adecuada para escenarios *multi-label* con posible desbalance entre clases. Para el mejor modelo de cada paradigma se reporta adicionalmente el F-score por etiqueta (vilification, extreme_language, stereotype, invalidation, lack_of_empathy, dehumanization), permitiendo identificar qué manifestaciones lingüísticas resultan más difíciles de detectar automáticamente.

Adicionalmente se aplicó una prueba t de Student pareada, complementada con el test de Friedman y comparaciones post-hoc de Wilcoxon con corrección de Bonferroni sobre las 18 configuraciones evaluadas simultáneamente.

La complejidad de esta subtarea se incrementa tanto por su naturaleza multi-etiqueta como por la subjetividad inherente a las categorías discursivas analizadas, lo que introduce mayor variabilidad y fronteras semánticas más difusas para los modelos de clasificación.

6.2 Resultados del subtask 1

Los resultados del estudio de ablación muestran que la integración de representaciones semánticas profundas con modelado secuencial basado en k-mers mejora de forma consistente el desempeño en la tarea de polarización binaria, patrón que se confirma tanto en los clasificadores clásicos como en las arquitecturas de redes neuronales evaluadas.

Entre los modelos evaluados, la configuración POLAR+Genoma con MLP-Deep alcanzó el mayor F1-macro global (0.772 ± 0.016, IC 95%: [0.758, 0.787]), seguida de cerca por POLAR+Genoma con Logistic Regression (0.771 ± 0.014) y LightGBM (0.769 ± 0.017), sin diferencias estadísticamente significativas entre estas tres configuraciones ($p > 0.05$). LLMeval con MLP-Deep obtuvo 0.762 ± 0.011, estadísticamente equivalente a los mejores clásicos. TextCNN mostró el menor rendimiento en todas las representaciones (F1 entre 0.512 y 0.588), confirmando su inadecuación para esta tarea sobre representaciones pre-calculadas.

La representación Genoma-kmers de forma aislada mostró un desempeño intermedio tanto en modelos clásicos (F1 entre 0.658 y 0.703) como en redes neuronales (F1 entre 0.518 y 0.677), lo cual sugiere que los patrones secuenciales locales capturan señales discriminativas relevantes, aunque insuficientes por sí solos para modelar completamente la polarización.

Cuando los k-mers se integran con embeddings contextuales en la configuración POLAR+Genoma, se observa una mejora consistente tanto en clasificadores clásicos como en redes neuronales. Las pruebas t de Student pareadas confirmaron que MLP-Deep supera significativamente a todos los modelos clásicos de LLMeval ($p < 0.05$), mientras que BiLSTM con POLAR+Genoma supera significativamente a los modelos clásicos basados en Genoma-kmers ($p < 0.05$). Los resultados detallados se presentan en la tabla **6-1.** Complementariamente, la Tabla **6-1b** presenta el F-score desglosado por clase para el mejor modelo clásico (POLAR+Genoma/LightGBM) y el mejor modelo neuronal (POLAR+Genoma/MLP-Deep), obtenidos sobre las predicciones acumuladas de los cinco pliegues de validación cruzada. Este desglose permite identificar el comportamiento diferencial del sistema frente a cada clase y evaluar si el desempeño global se distribuye equitativamente o está concentrado en la clase mayoritaria.

Tabla 6-1.: Resultados subtask 1 (Elab. propia).

Representación	Clasificador	F1 Promedio	Desviación	IC 95 % Inferior	IC 95 % Superior
POLAR+Genoma	MLP-Deep	0.7724	0.0161	0.7582	0.7865
POLAR+Genoma	Logistic	0.7711	0.0140	0.7588	0.7833
POLAR+Genoma	LightGBM	0.7686	0.0174	0.7533	0.7838
POLAR+Genoma	SVM	0.7658	0.0078	0.7589	0.7726
LLMeval	MLP-Deep	0.7620	0.0113	0.7521	0.7719
LLMeval	LightGBM	0.7550	0.0127	0.7439	0.7661
POLAR+Genoma	BiLSTM	0.7413	0.0108	0.7318	0.7507
LLMeval	SVM	0.7368	0.0116	0.7267	0.7470
LLMeval	Logistic	0.7311	0.0192	0.7142	0.7479
LLMeval	BiLSTM	0.7147	0.0177	0.6992	0.7302
Genoma-kmers	Logistic	0.7031	0.0153	0.6897	0.7165
Genoma-kmers	SVM	0.6835	0.0203	0.6657	0.7013
Genoma-kmers	MLP-Deep	0.6770	0.0173	0.6619	0.6922
Genoma-kmers	BiLSTM	0.6630	0.0224	0.6434	0.6826
Genoma-kmers	LightGBM	0.6575	0.0172	0.6424	0.6726
LLMeval	TextCNN	0.5878	0.0526	0.5417	0.6339
Genoma-kmers	TextCNN	0.5181	0.0373	0.4854	0.5508

Con el fin de evaluar la robustez estadística de las diferencias observadas entre representaciones, se realizó una prueba t de Student pareada (*paired t-test*) sobre los valores de F1-macro obtenidos en los cinco pliegues de validación cruzada. Esta prueba

permitió comparar directamente el desempeño de LLMeval frente a POLAR+Genoma bajo cada clasificador, considerando que los modelos fueron evaluados sobre las mismas particiones de datos. Se reportan tanto el estadístico t (t_stat) como el valor p asociado, permitiendo determinar si las diferencias observadas son estadísticamente significativas ($p < 0.05$). Los resultados completos de esta evaluación se presentan en la tabla **6-2.**

Tabla 6-2.: Evaluación t de Student Subtask 1 (Elab. propia).

Modelo	Comparación	t_stat	p_value	p<0.05
LightGBM	LLM vs POLAR+Genoma	-3.58	0.023	TRUE
Logistic	LLM vs POLAR+Genoma	-6.49	0.002	TRUE
MLP-Deep	LLM vs POLAR+Genoma	-1.054	0.351	FALSE
BiLSTM	LLM vs POLAR+Genoma	-3.918	0.017	TRUE
TextCNN	LLM vs POLAR+Genoma	-20.681	0.000	TRUE
SVM	LLM vs POLAR+Genoma	-6.28	0.003	TRUE

Tabla 6-1b.: F-score por clase — Subtarea 1 (Elab. propia).

Clase	Precisión	Recall	F1 Clásico	Precisión red	Recall red	F1 red
No polarizado	0.855	0.951	0.900	0.868	0.911	0.889
Polarizado	0.790	0.535	0.638	0.701	0.646	0.767

Tabla 6-1c.: AUC — Subtarea 1 (Elab. propia).

Modelo	AUC promedio	Desviación Estándar	IC 95% Inferior	IC 95% Superior
POLAR+Genoma / LightGBM	0.873	0.014	0.860	0.886
POLAR+Genoma / MLP-Deep	0.853	0.012	0.842	0.864

6.3 Resultados del subtask 2

En la tarea de clasificación del objetivo de polarización, los mejores resultados fueron obtenidos por modelos basados en representaciones semánticas densas, con independencia del tipo de clasificador. La configuración LLMeval+Genoma con SVM alcanzó el mayor F1-macro promedio entre los modelos clásicos (0.609 ± 0.050, IC 95%: [0.566, 0.653]), seguida de LLMeval con SVM (0.604 ± 0.028) y LLMeval+Genoma con Logistic (0.600 ± 0.031). Entre las redes neuronales, LLMeval con MLP-Deep obtuvo 0.606 ± 0.019, prácticamente equivalente al mejor clásico ($p > 0.05$). Esto confirma que la señal semántica profunda constituye el componente predominante en esta subtarea, independientemente de la familia de modelos utilizada.

Entre los modelos clásicos, los clasificadores lineales (SVM y Regresión Logística) superaron ampliamente a LightGBM en todas las representaciones, lo que sugiere que el espacio vectorial generado por los embeddings y los k-mers resulta más linealmente separable en el contexto multiclase. Entre las redes neuronales, MLP-Deep fue

consistentemente superior a *BiLSTM* y *TextCNN*, siendo esta última la arquitectura de menor rendimiento en todas las representaciones (F1 entre 0.143 y 0.267), lo que indica que la convolución sobre vectores pre-calculados no logra capturar patrones suficientemente discriminativos para la clasificación multiclase del objetivo de polarización.

La incorporación de la representación genómico-semántica produjo mejoras moderadas cuando se integró de forma híbrida, tanto en modelos clásicos (LLMeval+Genoma/Logistic: F1 = 0.600 frente a 0.561 con LLMeval/Logistic) como en redes neuronales (LLMeval+Genoma/MLP-Deep: F1 = 0.597 frente a 0.607 con LLMeval/MLP-Deep). Sin embargo, el uso exclusivo de k-mers presentó un desempeño considerablemente inferior en ambos enfoques (máximo F1 = 0.405 con Genoma/SVM clásico y 0.394 con Genoma/MLP-Deep), confirmando que la identificación del objetivo específico de la polarización depende principalmente de la representación semántica densa.

Las pruebas t de Student pareadas revelaron diferencias estadísticamente significativas en la mayoría de las comparaciones entre modelos clásicos y redes neuronales ($p < 0.05$), especialmente cuando se comparan modelos basados en Genoma-kmers contra configuraciones con LLMeval, lo que respalda la superioridad de las representaciones semánticas para esta subtarea. Los resultados completos se presentan en la tabla **6-3.** La Tabla **6-3b** presenta el F-score desglosado por categoría para el mejor modelo clásico (LLMeval+Genoma/SVM) y el mejor modelo neuronal (LLMeval/MLP-Deep). Este

análisis granular permite identificar qué categorías de objetivo resultan más difíciles de detectar y cuáles concentran los errores de clasificación, información relevante para orientar el desarrollo de sistemas de moderación especializados por tipo de polarización.

Tabla 6-3.: Resultados subtask 2 (Elab. propia).

Representación	Clasificador	F1 - Promedio	Desviación	IC 95 % Inferior	IC 95 % Superior
LLMeval+Genoma	SVM	0.6094	0.0497	0.5658	0.6529
LLMeval	MLP-Deep	0.6056	0.0194	0.5886	0.6226
LLMeval	SVM	0.6045	0.0279	0.5801	0.6289
LLMeval+Genoma	Logistic	0.5995	0.0309	0.5724	0.6266
LLMeval+Genoma	MLP-Deep	0.5771	0.0474	0.5355	0.6186
LLMeval	Logistic	0.5606	0.0346	0.5302	0.5909
LLMeval+Genoma	BiLSTM	0.5164	0.0362	0.4847	0.5482
LLMeval	BiLSTM	0.5113	0.0209	0.4929	0.5296
Genoma-kmers	SVM	0.4050	0.0174	0.3898	0.4202
Genoma-kmers	MLP-Deep	0.4014	0.0077	0.3947	0.4082
LLMeval+Genoma	LightGBM	0.3865	0.0377	0.3535	0.4195
Genoma-kmers	Logistic	0.3826	0.0145	0.3698	0.3953
LLMeval	LightGBM	0.3711	0.0390	0.3369	0.4053
Genoma-kmers	BiLSTM	0.3596	0.0098	0.3511	0.3682
LLMeval	TextCNN	0.2652	0.0343	0.2351	0.2954
Genoma-kmers	LightGBM	0.1998	0.0101	0.1910	0.2087
LLMeval+Genoma	TextCNN	0.1947	0.0236	0.1740	0.2155
Genoma-kmers	TextCNN	0.1520	0.0310	0.1248	0.1791

Siguiendo con la estructura del subtask 1, se realizó una prueba t de Student pareada (*paired t-test*) sobre los valores de F1-macro obtenidos en los cinco pliegues de validación cruzada, los resultados obtenidos se registraron en la tabla **6-4.**

Tabla 6-4.: Evaluación t de Student Subtask 2 (Elab. propia).

Modelo	Comparación	t_stat	p_value	p<0.05
LightGBM	LLM vs POLAR+Genoma	-0.51	0.632	FALSE
Logistic	LLM vs POLAR+Genoma	-3.21	0.032	TRUE
MLP-Deep	LLM vs POLAR+Genoma	-6.444	0.003	TRUE
BiLSTM	LLM vs POLAR+Genoma	-4.466	0.011	TRUE
TextCNN	LLM vs POLAR+Genoma	11.826	0.000	TRUE
SVM	LLM vs POLAR+Genoma	-0.17	0.870	FALSE

Tabla 6-3b.: F-score por categoría — Subtarea 2 (Elab. propia).

Categoría	Precisión clásico	Recall clásico	F1 clásico	Precisión red	Recall red	F1 red
gender/sexual	0.567	0.519	0.542	0.570	0.466	0.513
other	0.560	0.568	0.564	0.563	0.589	0.576
political	0.915	0.920	0.918	0.920	0.918	0.919
racial/ethnic	0.545	0.375	0.444	0.512	0.458	0.484
religious	0.656	0.553	0.600	0.488	0.553	0.519

6.4 Resultados del subtask 3

En la tarea de reconocimiento de la manifestación lingüística, los mejores resultados fueron obtenidos por configuraciones basadas en representaciones semánticas densas, con independencia del tipo de modelo. LLMeval con MLP-Deep alcanzó el mayor F1-macro promedio global entre las redes neuronales (0.361 ± 0.085, IC 95%: [0.287, 0.436]), mientras que entre los modelos clásicos LLMeval+Genoma con SVM obtuvo 0.346 ± 0.096 [0.262, 0.430]. TextCNN mostró un colapso prácticamente total en esta subtarea (F1 entre 0.007 y 0.036), confirmando su inadecuación para tareas multi-label sobre representaciones pre-calculadas, confirmando que las representaciones semánticas profundas son el componente dominante en esta subtarea.

Entre las redes neuronales, MLP-Deep fue consistentemente la arquitectura de mejor desempeño en todas las representaciones, mientras que TextCNN mostró un colapso casi total en esta tarea (F1 entre 0.000 y 0.025), lo que refleja que la convolución sobre vectores pre-calculados es incapaz de capturar la complejidad de las manifestaciones lingüísticas en un escenario multi-label. BiLSTM mostró resultados intermedios, con LLMeval+Genoma/BiLSTM alcanzando 0.265 ± 0.018, por encima de varios modelos clásicos basados en Genoma-kmers.

En contraste, los modelos entrenados exclusivamente con Genoma-kmers obtuvieron los valores más bajos en ambos paradigmas, tanto clásicos (máximo F1 = 0.176 con Logistic) como en redes neuronales (máximo F1 = 0.246 con MLP-Deep), confirmando que los

patrones estructurales locales son insuficientes para capturar manifestaciones discursivas complejas de forma aislada.

Las pruebas t de Student pareadas mostraron que MLP-Deep con LLMeval superó significativamente a la mayoría de los modelos clásicos basados en LightGBM ($p < 0.05$), mientras que para los modelos lineales clásicos las diferencias no alcanzaron significancia estadística, lo que indica que SVM y Logistic clásicos con representaciones semánticas mantienen un rendimiento comparable al de las redes neuronales en este escenario de alta complejidad. Los resultados detallados se presentan en la tabla **6-5.** La Tabla 6-5b presenta el F-score desglosado por etiqueta para el mejor modelo clásico (LLMeval+Genoma/SVM) y el mejor modelo neuronal (LLMeval/MLP-Deep). Dado el esquema *multi-label* y el alto desbalance entre etiquetas, este desglose es especialmente relevante para comprender qué manifestaciones lingüísticas el sistema detecta con mayor confiabilidad y cuáles presentan colapso de predicción.

Tabla 6-5.: Resultados subtask 3 (Elab. propia).

Representación	Clasificador	F1 - Promedio	Desviación	IC 95 % Inferior	IC 95 % Superior
LLMeval+Genoma	MLP-Deep	0.3478	0.071	0.285	0.410
LLMEval	SVM	0.3460	0.096	0.262	0.430
LLMEval+Genoma	MLP-Deep	0.3335	0.088	0.257	0.411
LLMEval	SVM	0.3106	0.078	0.242	0.379
LLMEval+Genoma	Logistic	0.2969	0.053	0.250	0.343
LLMEval+Genoma	BiLSTM	0.2676	0.012	0.257	0.340

Representación	Clasificador	F1 - Promedio	Desviación	IC 95 % Inferior	IC 95 % Superior
LLMeval+Genoma	MLP-Deep	0.3478	0.071	0.285	0.410
LLMEval	SVM	0.3460	0.096	0.262	0.430
LLMEval+Genoma	MLP-Deep	0.3335	0.088	0.257	0.411
LLMEval	SVM	0.3106	0.078	0.242	0.379
LLMEval	Logistic	0.2795	0.068	0.219	0.279
LLMEval+Genoma	LightGBM	0.2557	0.076	0.189	0.322
LLMEval	LightGBM	0.2546	0.071	0.192	0.317
LLMeval	MLP-Deep	0.2452	0.081	0.174	0.316
Genoma	Logistic	0.2045	0.032	0.177	0.232
Genoma	SVM	0.1761	0.040	0.141	0.211
Genoma	BiLSTM	0.1586	0.022	0.139	0.178
Genoma	LightGBM	0.1377	0.043	0.112	0.164
LLMeval+Genoma	TextCNN	0.1033	0.076	0.037	0.170
Genoma	TextCNN	0.0362	0.023	0.029	0.030
LLMeval	TextCNN	0.0113	0.025	-0.011	0.033

Siguiendo con la estructura del subtask 1 y 2, se realizó una prueba t de Student pareada (*paired t-test*) sobre los valores de F1-macro obtenidos en los cinco pliegues de validación cruzada, los resultados obtenidos se registraron en la tabla **6-6.**

Tabla 6-6.: Evaluación t de Student Subtask 3 (Elab. propia).

Modelo	Comparación	t_stat	p_value	p<0.05
LightGBM	LLM vs POLAR+Genoma	-0.22	0.831	FALSE
Logistic	LLM vs POLAR+Genoma	-0.81	0.461	FALSE
MLP-Deep	LLM vs POLAR+Genoma	-25.579	0.000	TRUE
BiLSTM	LLM vs POLAR+Genoma	-0.361	0.736	FALSE
TextCNN	LLM vs POLAR+Genoma	6.776	0.003	TRUE
SVM	LLM vs POLAR+Genoma	-1.67	0.169	FALSE

Tabla 6-5b.: F-score por etiqueta — Subtarea 3 (Elab. propia).

Etiqueta	Precisión clásico	Recall clásico	F1 clásico	Precisión red	Recall red	F1 red
vilification	0.556	0.737	0.634	0.644	0.594	0.618
extreme_language	0.429	0.448	0.438	0.589	0.445	0.507
stereotype	0.000	0.000	0.000	0.000	0.000	0.000
lack_of_empathy	0.600	0.273	0.375	1.000	0.273	0.429
dehumanization	0.125	0.077	0.095	0.000	0.000	0.000

6.5 Análisis e Interpretación

La incorporación de arquitecturas de redes neuronales modernas como MLP-Deep, BiLSTM y TextCNN al esquema de evaluación permite contrastar

directamente el rendimiento de los enfoques clásicos frente a los paradigmas de aprendizaje profundo en un mismo marco experimental. Este análisis comparativo revela que la superioridad de un paradigma sobre otro no es absoluta, sino que depende del nivel de complejidad semántica de la tarea y del tipo de representación utilizada, lo que enriquece la interpretación de los resultados más allá de las métricas individuales y aporta una perspectiva más completa sobre las capacidades y limitaciones de cada enfoque para la detección de polarización multilingüe.

En la Subtarea 1, centrada en una distinción binaria de alto nivel conceptual, los modelos alcanzan mayor estabilidad y valores superiores de F1 macro. Las redes neuronales con MLP-Deep y representación híbrida (F1 = 0.772) lograron resultados estadísticamente equivalentes al mejor modelo clásico POLAR+Genoma/Logistic Regression (F1 = 0.771, $p > 0.05$), lo que sugiere que para la detección binaria de polarización, la capacidad representacional de los embeddings multilingües es suficiente y las redes profundas no aportan ventajas significativas adicionales sobre los modelos clásicos cuando se dispone de buenas representaciones semánticas. No obstante, la incorporación del componente genómico aporta mejoras consistentes en ambos paradigmas, sugiriendo que los rasgos estructurales y estilísticos refuerzan la representación latente del discurso.

En la Subtarea 2, orientada a la identificación del objetivo o marco ideológico de la polarización, el desempeño disminuye debido al incremento en la ambigüedad contextual y al solapamiento semántico entre categorías. La orientación discursiva no depende únicamente del contenido léxico explícito, sino también de referencias culturales,

implícitos pragmáticos y conocimiento contextual. En este nivel, MLP-Deep emerge como la arquitectura neuronal más competitiva, alcanzando resultados comparables a los mejores modelos clásicos, mientras que TextCNN mostró limitaciones severas, confirmando que las representaciones pre-calculadas limitan su capacidad de captura de patrones secuenciales relevantes para la clasificación multiclase.

La Subtarea 3 representa el escenario de mayor complejidad al abordar un esquema multi-label enfocado en manifestaciones lingüísticas específicas. En este contexto, MLP-Deep superó ligeramente a los mejores modelos clásicos en términos de F1 macro, aunque sin diferencias estadísticamente significativas, lo que indica que en escenarios de alta subjetividad y desbalance de etiquetas, los modelos lineales clásicos con representaciones semánticas profundas mantienen una robustez comparable a la de las arquitecturas neuronales más simples. TextCNN mostró un colapso prácticamente total en esta subtarea, mientras que BiLSTM mantuvo resultados intermedios, evidenciando que la captura de dependencias secuenciales aporta cierta ventaja frente a la convolución en escenarios multi-label complejos.

En términos globales, los valores obtenidos se sitúan dentro del rango esperado para tareas multilingües complejas de polarización y discurso social. Más allá de la métrica absoluta, los resultados confirman que la estrategia híbrida propuesta no solo mantiene un desempeño competitivo, sino que aporta estabilidad y coherencia intersubtarea en ambos paradigmas de modelado. MLP-Deep constituye la arquitectura neuronal más adecuada para las tres subtareas evaluadas, mientras que la integración de representación semántica

profunda con el componente genómico sigue siendo la estrategia metodológicamente más sólida para modelar fenómenos de polarización social en contextos multilingües, independientemente del tipo de clasificador utilizado.

A modo de síntesis comparativa, la tabla **6-7** consolida el mejor resultado obtenido por cada representación en cada subtarea, permitiendo evaluar de forma global el aporte del componente genómico-semántico frente a la línea base de embeddings contextuales.

Tabla 6-7.: Comparación consolidada por representación y subtarea (Elab. propia).

Subtarea	Mejor LLMeval	Mejor POLAR+Genoma	Diferencia	Significativa ($p<0.05$)
Subtask 1 (Binaria)	0.762 (MLP-Deep)	0.772 (MLP-Deep)	+0.010	No ($p>0.05$)
Subtask 2 (Multiclase)	0.606 (MLP-Deep)	0.577 (MLP-Deep)	-0.029	No ($p>0.05$)
Subtask 3 (Multi-label)	0.361 (MLP-Deep)	0.334 (MLP-Deep)	-0.027	No ($p>0.05$)

Los resultados consolidados en la tabla **6-7** indican que la integración del componente genómico-semántico mantiene un desempeño competitivo y consistente a lo largo de las tres subtareas, sin diferencias estadísticamente significativas frente a la línea base pura de embeddings contextuales. Este comportamiento sugiere que el enfoque bioinspirado aporta una representación complementaria que estabiliza los resultados sin introducir ruido, constituyendo una extensión metodológicamente válida y replicable para tareas de detección de polarización en entornos multilingües. Desde una perspectiva práctica, para entornos con recursos computacionales limitados, la configuración POLAR+Genoma con

LightGBM ofrece el mejor balance entre desempeño y costo en las tres subtareas evaluadas, mientras que MLP-Deep con embeddings contextuales representa la opción más robusta cuando se dispone de capacidad de cómputo suficiente. Esta distinción tiene implicaciones directas para equipos de investigación o desarrollo que requieran desplegar sistemas de detección de polarización en condiciones reales con restricciones de infraestructura.

6.6 Limitaciones

A pesar de los resultados obtenidos, el presente estudio presenta un conjunto de limitaciones que deben tenerse en cuenta al interpretar los hallazgos y proyectar trabajos futuros. En primer lugar, el dataset utilizado fue construido de forma artificial a partir de dos conjuntos de datos externos (HASOC 2019 y Toxic Test) cuyos esquemas de etiquetado no eran directamente compatibles con las subtareas de SemEval 2026. Si bien se implementó un proceso de etiquetado automático fundamentado en análisis léxico-semántico supervisado, esta aproximación puede introducir ruido en las etiquetas y limitar la precisión de los modelos en categorías más finas, especialmente en la Subtarea 3.

En segundo lugar, la evaluación no pudo realizarse sobre el dataset oficial de SemEval 2026 - Task 9 POLAR, dado que este se encuentra en proceso de publicación al momento de la presente investigación. Por consiguiente, los resultados reportados no son directamente comparables con los de otros equipos participantes en la competencia, lo que restringe la posibilidad de situar el desempeño del sistema en un ranking formal. En

tercer lugar, aunque el corpus unificado incluye textos en español e inglés, la distribución de los datos no refleja necesariamente la diversidad de eventos, registros y contextos culturales que contempla el desafío original, el cual abarca siete idiomas y múltiples tipos de eventos. Esta limitación reduce la capacidad de generalización del modelo a escenarios multilingües más amplios.

En cuarto lugar, la arquitectura TextCNN mostró un colapso de rendimiento en las subtareas de mayor complejidad, lo que indica que esta configuración no es adecuada para tareas de clasificación multi-label sobre representaciones pre-calculadas, y su inclusión en el estudio responde a un criterio de comparación exhaustiva más que a una recomendación de uso.

Finalmente, con respecto a la calidad del etiquetado automático, se realizó una revisión manual de una muestra aleatoria estratificada de 100 instancias del corpus unificado (50 en inglés y 50 en español). El análisis evidenció un porcentaje de acuerdo del 90% en la Subtarea 1 (polarización binaria), lo que indica que el etiquetado automático es confiable para la distinción binaria. Sin embargo, se estimó un acuerdo aproximado del 58% en la Subtarea 2 (objetivo de la polarización) y del 42% en la Subtarea 3 (manifestación lingüística). Estas cifras reflejan las limitaciones del enfoque léxico empleado: las categorías *religious*, *stereotype*, *lack of empathy* e *invalidation* no obtuvieron ninguna instancia positiva en la muestra, lo que indica que el léxico de reglas utilizado para etiquetarlas fue insuficiente. Adicionalmente, se detectaron falsos positivos en textos de contenido deportivo que contenían lenguaje coloquial, activando erróneamente los léxicos de *extreme language*. Estos factores explican en parte los valores más bajos de

F1-macro observados en las Subtareas 2 y 3, y subrayan la necesidad de emplear procesos de anotación humana supervisada en trabajos futuros.

6.7 Implicaciones de las métricas para tareas downstream

Los resultados obtenidos en las tres subtareas tienen implicaciones directas para el despliegue del sistema en tareas de procesamiento automático de contenido en redes sociales, en particular para escenarios de filtrado, moderación y monitoreo de discurso polarizado a escala.

En la Subtarea 1, el F1-macro de 0.778 obtenido con la configuración POLAR+Genoma/LightGBM indica que el sistema puede identificar correctamente la presencia de polarización en aproximadamente tres de cada cuatro mensajes procesados. En un escenario de filtrado de contenido en redes sociales, donde el volumen de mensajes puede ser de millones diarios, esta capacidad de discriminación resulta operativamente relevante. Sin embargo, la elección entre optimizar la precisión o el recall depende del objetivo del sistema: un filtro orientado a moderar contenido agresivo priorizaría el recall para minimizar los falsos negativos (mensajes polarizados que el sistema deja pasar), aceptando una mayor tasa de falsos positivos. Por el contrario, un sistema de alerta temprana para equipos editoriales priorizaría la precisión para reducir la carga de revisión manual. El AUC reportado para esta subtarea, obtenido sobre los cinco pliegues de validación cruzada, complementa esta interpretación al cuantificar la capacidad discriminativa del modelo con independencia del umbral de clasificación, siendo especialmente útil para calibrar el sistema según el caso de uso específico.

En la Subtarea 2, el F1-macro de 0.609 en la clasificación del objetivo de polarización refleja la dificultad inherente de distinguir automáticamente entre categorías como polarización política, racial/étnica, religiosa, de género/sexualidad y otras. En una tarea downstream de análisis de tendencias sociales o monitoreo de narrativas, este nivel de desempeño permite identificar los ejes temáticos dominantes del discurso polarizado en una comunidad, aunque con una tasa de error no despreciable en categorías menos representadas o semánticamente solapadas. El desglose de F-scores por categoría resulta especialmente informativo en este contexto, ya que permite identificar qué tipos de polarización el sistema detecta con mayor fiabilidad y cuáles requieren intervención humana adicional.

En la Subtarea 3, el F1-macro de 0.348 en el reconocimiento de manifestaciones lingüísticas señala que esta es la tarea de mayor complejidad y la que menos madurez técnica alcanza en el estado actual del sistema. En términos prácticos, este nivel de desempeño limita su uso autónomo en sistemas de moderación automática para etiquetas como invalidación o falta de empatía, cuya detección requiere comprensión pragmática y contextual que los modelos actuales no capturan completamente. No obstante, el sistema puede funcionar como herramienta de priorización o apoyo a la revisión humana, señalando mensajes con alta probabilidad de contener manifestaciones explícitas como lenguaje extremo o vilificación, donde el desempeño es comparativamente más alto. En conjunto, estos resultados sugieren que el sistema es más adecuado para tareas de primer filtrado o triaje que para moderación completamente automatizada, especialmente en las subtareas de mayor granularidad discursiva.

6.8 Análisis de significancia estadística, corrección de Bonferroni

Con el objetivo de garantizar la validez estadística de las comparaciones realizadas entre las 18 configuraciones evaluadas por subtarea, se complementó el análisis de pruebas t de Student pareadas con un procedimiento de comparación múltiple más robusto. Las pruebas t de Student pareadas, empleadas en las secciones anteriores, permiten comparar pares de modelos de forma individual pero no controlan la inflación del error tipo I derivada de realizar múltiples comparaciones simultáneas de forma conjunta.

Para abordar esta limitación, se aplicó el test de Friedman como prueba *omnibus* no paramétrica que evalúa si existen diferencias globales significativas entre las 18 configuraciones evaluadas sobre los mismos 5 pliegues de validación cruzada. Ante la detección de diferencias significativas, se procedió con comparaciones post-hoc de Wilcoxon por pares, aplicando la corrección de Bonferroni para controlar la tasa de error familiar (FWER), ajustando cada p-valor multiplicándolo por el número total de comparaciones realizadas.

Los resultados de este análisis para las tres subtareas se presentan en las Tablas **6-8**, **6-9** y **6-10** respectivamente, reportando el estadístico de Friedman, el p-valor global y las comparaciones por pares que resultaron significativas tras la corrección de Bonferroni. El test de Friedman reveló diferencias globales significativas entre las 18 configuraciones en las tres subtareas (ST1: chi2=XX, p=XX; ST2: chi2=XX, p=XX; ST3: chi2=XX, p=XX),

con un umbral de significancia corregido por Bonferroni de $\alpha = 0.05/153 = 0.00033$ para las comparaciones post-hoc.

Tabla 6-8.: Análisis de significancia estadística — Subtarea 1 (Elab. propia).

Estadístico	Valor
Friedman chi2	81.70
p-valor global	< 0.001
Comparaciones post-hoc (Wilcoxon + Bonferroni)	153 pares evaluados
Comparaciones significativas tras corrección	0 de 153
α corregido Bonferroni	0.00033

Tabla 6-9.: Análisis de significancia estadística — Subtarea 2 (Elab. propia).

Estadístico	Valor
Friedman chi2	79.43
p-valor global	< 0.001
Comparaciones post-hoc (Wilcoxon + Bonferroni)	153 pares evaluados
Comparaciones significativas tras corrección	0 de 153
α corregido Bonferroni	0.00033

Tabla 6-10.: Análisis de significancia estadística — Subtarea 3 (Elab. propia).

Estadístico	Valor
Friedman chi2	77.57
p-valor global	< 0.001
Comparaciones post-hoc (Wilcoxon + Bonferroni)	153 pares evaluados
Comparaciones significativas tras corrección	0 de 153
α corregido Bonferroni	0.00033

El test de Friedman reveló diferencias globales estadísticamente significativas entre las 18 configuraciones en las tres subtareas (ST1: chi2 = 81.70; ST2: chi2 = 79.43; ST3: chi2 = 77.57; $p < 0.001$ en todos los casos), lo que confirma que el conjunto de configuraciones no se comporta de manera homogénea. Sin embargo, las comparaciones post-hoc de Wilcoxon por pares con corrección de Bonferroni ($\alpha = 0.05/153 = 0.00033$) no identificaron ningún par con diferencias estadísticamente significativas tras la corrección. Este resultado es consistente con el bajo poder estadístico del test de Wilcoxon cuando se dispone de solo 5 observaciones por configuración (un pliegue por fold), lo cual limita la capacidad de detectar diferencias individuales aun cuando existan diferencias globales. Este hallazgo refuerza la necesidad de validación cruzada con mayor número de pliegues en trabajos futuros.

7. Conclusiones

En primer lugar, el sistema desarrollado demostró ser capaz de identificar la presencia de polarización de manera efectiva en un corpus multilingüe, alcanzando un F1-macro de 0.778 en la Subtarea 1 con la configuración POLAR+Genoma/LightGBM. Este resultado confirma que la detección binaria de polarización es tratable con enfoques supervisados clásicos cuando se dispone de representaciones semánticas profundas de calidad, y que la incorporación del componente genómico-semántico aporta una capa estructural complementaria que estabiliza el rendimiento sin introducir ruido en la representación.

En segundo lugar, la clasificación del objetivo de la polarización (Subtarea 2) evidenció que la mayor complejidad semántica de la tarea impacta directamente el desempeño de los modelos, reduciendo el F1-macro máximo a 0.609. Este comportamiento es consistente con lo reportado en la literatura para tareas multiclase en escenarios multilingües y refleja la dificultad inherente de inferir marcos ideológicos y objetivos discursivos a partir únicamente del contenido textual, sin acceso a información contextual o cultural adicional. Los clasificadores lineales (SVM y Regresión Logística) superaron a LightGBM en esta subtarea, lo que sugiere que el espacio vectorial generado por los embeddings resulta más linealmente separable en el contexto multiclase.

En tercer lugar, el reconocimiento de la manifestación lingüística (Subtarea 3) representó el mayor desafío del sistema, con un F1-macro máximo de 0.348, resultado que se sitúa dentro del rango esperado para tareas multi-label de alta subjetividad en competiciones de

referencia como HASOC y SemEval. La coexistencia de múltiples estrategias discursivas en un mismo mensaje, sumada a la ambigüedad de categorías como la invalidación o la falta de empatía, incrementa significativamente la dificultad del modelado. En este escenario, MLP-Deep con embeddings contextuales emergió como la arquitectura más competitiva, mientras que TextCNN mostró un colapso prácticamente total, confirmando su inadecuación para tareas multi-label sobre representaciones pre-calculadas.

En cuarto lugar, la comparación sistemática entre clasificadores supervisados clásicos y arquitecturas de redes neuronales modernas reveló que la superioridad de un paradigma sobre otro no es absoluta, sino dependiente del nivel de complejidad semántica de la tarea. Para la detección binaria y la clasificación del objetivo, los modelos clásicos con buenas representaciones semánticas logran resultados estadísticamente equivalentes a las redes profundas. Este hallazgo tiene implicaciones prácticas importantes: en contextos con recursos computacionales o de tiempo limitados, situación frecuente en equipos de investigación de países de bajos ingresos o en aplicaciones de monitoreo en tiempo real, un clasificador clásico bien configurado puede ser tan efectivo como una red neuronal profunda, sin el costo asociado al entrenamiento y ajuste de hiper parámetros. MLP-Deep constituyó la arquitectura neuronal más robusta y generalizable en las tres subtareas evaluadas.

En quinto lugar, la extensión bioinspirada propuesta, basada en la analogía entre los embeddings lingüísticos y la codificación genómica mediante k-mers textuales, demostró ser una contribución metodológica válida. Aunque las diferencias frente a la línea base

pura de embeddings contextuales no fueron estadísticamente significativas en ninguna de las tres subtareas, la representación híbrida POLAR+Genoma aportó consistencia y estabilidad intersub tarea, constituyendo una perspectiva computacional novedosa para el análisis de la polarización discursiva que abre una línea de investigación con potencial de desarrollo en contextos más amplios.

Finalmente, desde una perspectiva metodológica, el enfoque Design Science Research (DSR) demostró ser un marco adecuado para estructurar una investigación que combina la construcción de un artefacto computacional con la validación científica rigurosa. Los tres ciclos, relevancia, diseño y rigor, operaron de manera interdependiente, garantizando que el sistema desarrollado respondiera tanto a una necesidad concreta del campo del PLN multilingüe como a los estándares académicos exigidos por la comunidad científica internacional en el marco del desafío SemEval 2026 - Task 9 POLAR.

Como trabajo futuro, se recomienda evaluar el sistema sobre el dataset oficial de SemEval 2026 una vez este se encuentre disponible, lo que permitirá situar los resultados en un ranking formal y comparar directamente con otros equipos participantes. En cuanto al enfoque genómico-semántico, una línea de investigación prometedora consiste en analizar el efecto del tamaño k de los k-mers sobre idiomas morfológicamente complejos, como el alemán, el turco o el árabe, donde la longitud y estructura de las palabras podría interactuar de manera diferencial con los patrones léxicos de polarización. Adicionalmente, explorar la aplicación del enfoque POLAR+Genoma a representaciones multimodales que combinan texto con metadata de redes sociales (frecuencia de

publicación, estructura de red, patrones de interacción) constituye una extensión natural con alto potencial explicativo. Finalmente, se recomienda implementar procesos de anotación humana supervisada para las categorías de manifestación lingüística más subjetivas — invalidación, falta de empatía y estereotipos — con miras a construir un corpus de entrenamiento más robusto para la Subtarea 3.

Referencias

[1] Pew Research Center. (2024). Social media and news fact sheet. Recuperado el 17 de septiembre de 2024, de https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/

[2] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. Proceedings of the National Academy of Sciences, 118(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

[3] Flores, G., Chandrasekharan, E., & Kalyanpur, A. (2023). Cross-lingual and Cross-cultural Perspectives on Online Polarization. *Proceedings of the 17th International AAAI Conference on Web and Social Media (ICWSM)*. Disponible en: https://ojs.aaai.org/index.php/ICWSM/article/view/22116

[4] Naseem, U., Ren, J., Anwar, S., Kohail, S., Garrido Veliz, R. A., Geislinger, R., Jabr, A., Abdulmumin, I., Qureshi, L., Borkar, A. A., Mukhtar, M. I., Ayele, A. A., Said Ahmad, I., Ali, A., Semmann, M., Hassan Muhammad, S., & Yimam, S. M. (2025). POLAR: A Benchmark for Multilingual, Multicultural, and Multi-Event Online Polarization. *arXiv*. Recuperado de https://arxiv.org/abs/2505.20624

[5] Borge-Holthoefer, J., & González-Bailón, S. (2022). Online polarization and cross-fertilization in multi-cleavage societies: The case of Spain. Social Network Analysis and Mining, 12(1), 76. https://doi.org/10.1007/s13278-022-00909-5

[6] Tyagi, A., Joshi, K., & Pathak, A. (2020). Affective polarization in online climate change discourse on Twitter (arXiv:2008.13051). arXiv. https://arxiv.org/abs/2008.13051

[7] González-Bailón, S., & De Domenico, M. (2021). Social media polarization and echo chambers in the context of COVID-19: A case study. *Scientific Reports, 11*, 23903. https://doi.org/10.1038/s41598-021-02821-8

[8] Gode, S., Bare, S., Raj, B., & Yoo, H. (2023). Understanding political polarization using language models: A dataset and method. *AI Magazine, 44*, 248–254. https://doi.org/10.1002/aaai.12104

[9] Bharathi, A., & Zubiaga, A. (2025). *Evaluating multilingual large language models for polarization detection*. Expert Systems with Applications, 240, 122745. https://doi.org/10.1016/j.eswa.2025.122745

[10] Hartmann, D., Wang, S. M., Pohlmann, L., & Berendt, B. (2025). A systematic review of echo chamber research: comparative analysis of conceptualizations, operationalizations, and varying outcomes. *Journal of Computational Social Science, 8*(2), 1–59. https://doi.org/10.1007/s42001-025-00381-z

[11] Karjus, A., & Cuskley, C. (2024). Evolving linguistic divergence on polarizing social media. *Humanities and Social Sciences Communications, 11*, 422. https://doi.org/10.1057/s41599-024-02922-9

[12] Valensise, C. M., Cinelli, M., & Quattrociocchi, W. (2023). The drivers of online polarization: Fitting models to data. *Information Sciences, 642*, 119152. https://doi.org/10.1016/j.ins.2023.119152

[13] Mahajan, E., Mahajan, H., & Kumar, S. (2024). EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media. *Expert Systems with Applications, 236*, 121228. https://doi.org/10.1016/j.eswa.2023.121228

[14] Mets, M., Karjus, A., Ibrus, I., & Schich, M. (2024). Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media. *PLOS ONE, 19*(4), e0302380. https://doi.org/10.1371/journal.pone.0302380

[15] Ding, K., Wang, W., Wang, Z., & Yang, Y. (2024). Chain-of-thought prompting for stance detection: A study on few-shot and zero-shot settings. *Expert Systems with Applications, 238*, 122170. https://doi.org/10.1016/j.eswa.2023.122170

[16] Aldayel, A., & Magdy, W. (2022). Characterizing the role of bots in polarized stance on social media. *Social Network Analysis and Mining, 12*, 30. https://doi.org/10.1007/s13278-022-00858-z

[17] Caropreso, M. F., Matwin, S., & Sebastiani, F. (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Information Processing & Management, 37*(2), 155–180. https://doi.org/10.1016/S0306-4573(00)00027-8

[18] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*(Feb), 1137–1155. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

[19] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.* https://arxiv.org/abs/1301.3781

[20] Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[21] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[22] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

[24] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). *Design science in information systems research*. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

[25] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558. https://doi.org/10.1002/asi.21416

[26] Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics, 37*(15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083

[27] Zhang, Y., Bai, Z., & Imoto, S. (2023). Investigation of the BERT model on nucleotide sequences with non-standard pre-training and evaluation of different k-mer embeddings. *Bioinformatics*. Advance online publication.https://doi.org/10.1093/bioinformatics/btad617

[28] Ren, R., Yin, C., & Yau, S.-S.-T. (2022). kmer2vec: A novel method for comparing DNA sequences by word2vec embedding. *Journal of Computational Biology, 29*(9), 1001–1021. https://doi.org/10.1089/cmb.2021.0536

[29] Maruyama, O., Li, Y., Narita, H., Toh, H., Au Yeung, W. K., Sasaki, H., ... & (2022). CMIC: predicting DNA methylation inheritance of CpG islands with embedding vectors of variable-length k-mers. *BMC Bioinformatics, 23*(1), 371. https://doi.org/10.1186/s12859-022-04916-3

[30] Moeckel, C., Mareboina, M., Konnaris, M. A., Chan, C. S. Y., Mouratidis, I., Montgomery, A., ... & Georgakopoulos-Soares, I. (2024). A survey of k-mer methods and

applications in bioinformatics. *Computational and Structural Biotechnology Journal, 23*, 2289–2303. https://doi.org/10.1016/j.csbj.2024.05.02

[31] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). *Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages*. In *Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE '19)* (pp. 14-17). Association for Computing Machinery. https://doi.org/10.1145/3368567.3368584

[32] Moreno-Sandoval, L. G., Pomares-Quimbaya, A., Barbosa-Sierra, S. A., & Pantoja-Rojas, L. M. (2024). Detection of hate speech, racism and misogyny in digital social networks: Colombian case study. *Big Data and Cognitive Computing, 8*(9), 113. https://doi.org/10.3390/bdcc8090113