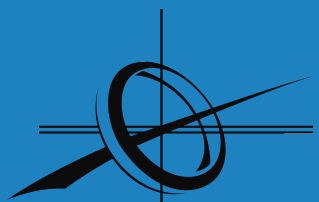
POLITÉCNICA



Universidad Politécnica de Madrid

ETSI SISTEMAS INFORMÁTICOS

Análisis del perfilado de autor y del tratamiento automatizado de la información en contextos digitales globales

Proyecto Fin de Grado

Grado en Ingeniería del Software

Autor:
Lázaro Granados Martín

Tutores:
Sergio D'Antonio Maceiras

Julio 2025

Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos

Análisis del perfilado de autor y del tratamiento automatizado de la información en contextos digitales globales

Proyecto Fin de Grado

Grado en Ingeniería del Software

Curso académico 2024-2025

Autor:
Lázaro Granados Martín

Tutores:
Sergio D'Antonio Maceiras

A todas las personas que han sabido quererme, aceptarme y me han acompañado durante mis peores etapas

Resumen

El presente Trabajo de Fin de Grado tiene por objeto el estudio del *author profiling* como disciplina dentro del procesamiento del lenguaje natural, con especial atención a su evolución reciente en el marco de las competencias PAN (*Plagiarism Analysis, Authorship and Social Software Misuse*), celebradas entre los años 2020 y 2024. El objetivo principal consiste en analizar de forma sistemática los enfoques metodológicos aplicados en dichas ediciones, identificando patrones tecnológicos, estrategias recurrentes y su rendimiento relativo en tareas de clasificación de autor, verificación de autoría, detección de cambios de estilo y análisis moral.

La investigación se apoya en 40 trabajos seleccionados en función de su rendimiento técnico y relevancia documental. A partir de un proceso de revisión cualitativa y análisis comparativo, se realiza una clasificación de modelos en cuatro grandes categorías: enfoques tradicionales, redes neuronales clásicas, arquitecturas basadas en *transformers* y sistemas híbridos. El estudio revela la progresiva consolidación de modelos preentrenados de gran escala (LLMs) y la aplicación de técnicas de *few-shot learning* en entornos con recursos limitados.

El trabajo concluye con una reflexión crítica sobre los desafíos éticos asociados al perfilado automatizado de autoría y propone una línea futura de investigación centrada en el análisis moral del discurso digital.

Abstract

Este Trabajo de Fin de Grado analiza de forma sistemática el estado del arte en *author profiling*, disciplina del procesamiento del lenguaje natural que busca inferir características del autor a partir de sus textos. El estudio se centra en las tareas propuestas por PAN entre 2020 y 2024, en el marco de la conferencia CLEF. A partir de 40 contribuciones destacadas, se identifican los principales enfoques técnicos utilizados en la disciplina: modelos tradicionales, redes neuronales clásicas, arquitecturas basadas en *transformers* y soluciones híbridas. Se evalúa el rendimiento de cada enfoque según la naturaleza de la tarea (clasificación, verificación de autoría, cambio de estilo, análisis moral), destacando la progresiva consolidación de los grandes modelos de lenguaje y el uso creciente de *few-shot learning* en escenarios con datos escasos. El trabajo también aborda las implicaciones éticas del perfilado automatizado, como la privacidad, los sesgos algorítmicos y la transparencia. Finalmente, se propone una línea futura de investigación orientada al análisis del impacto moral en el discurso digital, integrando aspectos éticos y sociales en el diseño de sistemas inteligentes. Este trabajo combina análisis técnico con una reflexión crítica sobre el papel de la IA en entornos comunicativos contemporáneos.

Índice

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

La coyuntura política y comunicativa actual se encuentra profundamente marcada por una crisis de legitimidad informativa y una creciente instrumentalización del lenguaje en los entornos digitales. Los medios de comunicación y las plataformas sociales actúan simultáneamente como altavoces de discursos polarizados y como agentes de reproducción de sesgos, configurando un ecosistema donde la viralidad y la emocionalidad prevalecen sobre la veracidad y el pensamiento crítico. Esta deriva tiene consecuencias morales y sociales de gran calado, que van desde la normalización del discurso de odio hasta la erosión del consenso democrático. En este contexto, la ingeniería del software adquiere una dimensión estratégica: no solo como disciplina técnica, sino como herramienta de intervención en la esfera pública. Desarrollar sistemas capaces de analizar, perfilar y auditar los discursos que circulan en redes no es una opción secundaria, sino una respuesta urgente a los desafíos de una sociedad que consume información de forma acelerada, fragmentaria y a menudo acrítica. Este trabajo parte de esa urgencia y propone, desde el ámbito del procesamiento del lenguaje natural, una contribución fundamentada al estudio y mejora del perfilado de autor como vía de comprensión y, eventualmente, de regulación ética del discurso digital.

1.1. Contexto

El presente trabajo surge en un contexto marcado por profundas transformaciones en la manera en que se produce, difunde y consume información en entornos digitales. La proliferación de contenidos generados por usuarios, el auge de las redes sociales y el papel creciente de los modelos de inteligencia artificial (IA) en la creación de texto han multiplicado tanto las posibilidades expresivas como los riesgos asociados a la manipulación y distorsión de la información.

Desde una perspectiva político-social, nos encontramos en un periodo especialmente sensible, donde el discurso público se polariza con facilidad, proliferan fenómenos como la desinformación o las *fake news*, y emergen nuevos actores comunicativos con capacidad de influencia masiva. Este escenario ha motivado una preocupación creciente por parte de organismos públicos, medios de comunicación y ciudadanía, no solo respecto al contenido en sí, sino también sobre las intenciones y perfiles de quienes lo emiten.

En paralelo, disciplinas como la psicología computacional y el análisis de comportamiento textual han cobrado relevancia, al ofrecer herramientas para entender cómo ciertos estilos lingüísticos, narrativas emocionales o patrones discursivos pueden estar vinculados a características del emisor, como su identidad ideológica, nivel de toxicidad o capacidad persuasiva. En este sentido, el *author profiling* adquiere un valor estratégico para detectar patrones de radicalización, perfilar emisores de desinformación o identificar riesgos comunicativos.

Justamente por ello, las tareas impulsadas en el marco de **PAN**[1][1], organizadas anualmente en **CLEF** [2][2], ofrecen un campo de experimentación riguroso para evaluar soluciones técnicas que combinan lingüística computacional, aprendizaje automático y responsabilidad ética. Este trabajo se plantea, por tanto, como una contribución al estudio de las arquitecturas, métodos y dilemas que configuran el *author profiling* actual, evaluando su eficacia frente a los desafíos reales que plantea el ecosistema comunicativo contemporáneo.

1.2. Objetivos

El objetivo de este Proyecto de Fin de Grado es ofrecer una revisión de los modelos de *Machine Learning* en materia de análisis de textos. Usando como referencia los estudios y talleres que se publican anualmente en la convención **CLEF**, para ello, tomaremos las publicaciones de los equipos y las revisiones de las tareas publicadas en PAN de los últimos años, basadas en *fake news*, *author profiling*, identificación de la autoría de textos y el pensamiento crítico.

Se establecen los siguientes objetivos analíticos en torno al corpus de trabajos presentados en las competencias PAN entre 2020 y 2024:

1. Identificar patrones tecnológicos predominantes: Analizar la evolución en el uso de modelos y tecnologías en las tareas de *author profiling*, prestando especial atención a:

- La adopción progresiva de redes neuronales profundas.
- La aparición de modelos basados en transformers y LLM[3].
- El uso de arquitecturas híbridas y técnicas tradicionales frente a soluciones basadas en aprendizaje profundo.

2. Evaluar el rendimiento relativo de las soluciones: Estudiar las puntuaciones obtenidas por los equipos en función de:

- El tipo de tecnología empleada.
- La naturaleza de la tarea o subtarea (clasificación, verificación, multilingüismo, etc.).
- La evolución de un mismo enfoque técnico en diferentes ediciones.

3. Detectar estrategias recurrentes y su mejora a lo largo del tiempo: Identificar si existen equipos o grupos de investigación que han mantenido líneas continuas de trabajo, y cómo han optimizado o reconfigurado sus soluciones a lo largo de los años.

[1]Siglas en inglés de *Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*

[2]Siglas en inglés de *Convention and Labs of Evaluation Forum*

[3]Siglas en inglés de *Large Language Model*

4. Comparar tareas similares entre distintas ediciones: Establecer comparaciones cualitativas y cuantitativas entre tareas de enunciado análogo, valorando la dificultad inherente, los enfoques utilizados y la evolución en la calidad de los resultados.

5. Proponer una tipología de soluciones técnicas aplicadas al *author profiling* sobre un escenario: Analizar si los distintos enfoques técnicos permitirían el desarrollo de aplicaciones para determinar el impacto moral tanto en redes sociales como en mensajes por parte de instituciones.

Con el resultado de este estudio se ofrece una línea futura para responder a uno de los mayores problemas sociales y políticos que están surgiendo en las principales redes sociales: los sistemas de verificación de contenido y la regulación de su consumo.

1.3. Estructura del documento

El presente documento se organiza en cinco capítulos principales y ofrece un análisis sistemático del uso de modelos de *Machine Learning* aplicados al *author profiling* en las competiciones PAN entre los años 2020 y 2024.

El **Capítulo 1**, dedicado a la introducción, contextualiza el auge de la inteligencia artificial en la generación de contenidos, en un entorno mediático marcado por la polarización política, la desinformación y el uso estratégico del lenguaje en redes sociales. En este marco, el *author profiling* se propone como una herramienta analítica clave para identificar patrones discursivos, inferir características de los emisores y detectar posibles focos de radicalización. Se presentan los objetivos principales del estudio, que incluyen la identificación de patrones tecnológicos, la evaluación del rendimiento por tipo de tarea, la sistematización de estrategias metodológicas, la comparación entre ediciones sucesivas de PAN y la formulación de una tipología de soluciones técnicas aplicadas a escenarios sensibles desde el punto de vista moral.

En el **Capítulo 2**, correspondiente al estado del arte, se traza una revisión crítica del fenómeno desde múltiples ángulos. Se analizan las dinámicas sociotécnicas que favorecen la viralización de desinformación en contextos digitales y se abordan aspectos psicológicos asociados al consumo y producción de contenido, como la moralidad percibida, el sesgo de confirmación y los mecanismos de licenciamiento moral. También se examina la evolución del Procesamiento del Lenguaje Natural (PLN), desde sus raíces en la computación simbólica hasta la consolidación de arquitecturas profundas como BERT[4], GPT[5] o LLaMA [6]. Se describen distintos enfoques técnicos y se discuten sus implicaciones éticas. Asimismo, se profundiza en el impacto de las *fake news*, el papel de los algoritmos de recomendación y la necesidad de herramientas automatizadas para su detección. Finalmente, se introduce el *author profiling* como disciplina, destacando su aplicabilidad en contextos sociales y sus dilemas éticos relacionados con la privacidad y los sesgos algorítmicos.

[4]Siglas en inglés de *Bidirectional Encoder Representations from Transformers*
[5]Siglas en inglés de *Generative Pre-trained Transformer*
[6]Siglas en inglés de *Large Language Model Meta AI*

El **Capítulo 3** desarrolla el proceso metodológico seguido en el análisis. Se parte de una recopilación de 40 trabajos presentados en PAN entre 2020 y 2024, seleccionando aquellos que obtuvieron las mejores puntuaciones en tareas de atribución de autoría, detección de plagio y perfilado. Los sistemas se clasifican en categorías según su arquitectura (tradicionales, redes neuronales clásicas, *transformers*/LLMs y modelos híbridos), y se aplica una revisión sistemática cualitativa y cuantitativa que permite registrar las variables clave de cada propuesta.

El **Capítulo 4** presenta los resultados obtenidos. Se confirma la consolidación de los modelos basados en transformadores a partir de 2021 como las soluciones más eficaces en tareas de clasificación, aunque también se observa un rendimiento notable de métodos híbridos y clásicos, especialmente en ediciones anteriores. Se describen las tendencias tecnológicas año a año y se analiza el rendimiento diferencial según el tipo de tarea (clasificación de autor, verificación de autoría o detección de cambios de estilo), subrayando la complejidad inherente a las tareas de verificación. Además, se documentan soluciones representativas por tipo de arquitectura, como el modelo mT0-XL con ORPO [7] para desintoxicación de texto, el sistema híbrido ADHOMINEM o el enfoque tradicional (n-gramas y SVM [8]). La sección concluye con la evaluación del cumplimiento de los objetivos analíticos propuestos, confirmando la solidez metodológica del trabajo.

El **Capítulo 5** recoge las conclusiones del estudio y propone líneas de trabajo futuras. Se reflexiona sobre el impacto social y medioambiental de las tecnologías analizadas, con especial atención a su rol en la generación y control de contenidos en plataformas digitales. Entre las líneas futuras se propone la creación de nuevas tareas PAN orientadas al análisis del impacto moral del discurso digital y del consumo político en redes sociales, formulando una propuesta denominada AIMEDD (Análisis del Impacto Moral y Ético en el Discurso Digital). Esta propuesta contempla el uso de *datasets* anotados y modelos avanzados para identificar juicios morales, medir la intensidad ética del lenguaje, detectar manipulación o sesgos algorítmicos, y caracterizar moralmente al autor del contenido.

[7] Siglas en inglés de *Odds Ratio Preference Optimization*
[8] Siglas en inglés de *Support Vector Machine*

Capítulo 2

Estado del arte

La actualidad político-social está dominada por un discurso generalmente populista; las redes sociales se inundan de comentarios que van más allá de la opinión; en las noticias oímos hablar de bulos, mentiras, discursos de odio... todo ello generando un clima de confrontación social que inmediatamente se refleja en las redes sociales [3, 4, 5]. Esta nueva realidad se ve intensificada y amplificada por el papel que desempeña la tecnología. El uso de *bots*, la inteligencia artificial y el análisis masivo de datos participan activamente en este espacio colectivo, hasta el punto de convertirse en la herramienta más poderosa y determinante para transformar —o incluso distorsionar— nuevas realidades. Estas realidades no son otra cosa que percepciones individuales, personalizadas y sesgadas en función del consumo de información [6, 7, 8].

Para comprender este problema en su conjunto, es útil considerar tres grandes grupos que lo configuran. El primero hace referencia al **espacio en el que se desarrolla la confrontación: Internet**. Aunque se trata de un entorno virtual, su uso puede verse condicionado por factores geográficos, políticos o sociales, lo que provoca que el acceso a la información y su difusión varíen según el lugar. Esta desigualdad incide directamente en la forma en que los contenidos se consumen y se comparten en la red.

El segundo grupo está formado por las **personas que participan en estas dinámicas comunicativas**. En este caso, resultan relevantes los factores psicológicos y motivacionales que influyen en cómo los usuarios interactúan, qué tipo de mensajes emiten y cómo responden a los demás. Estas interacciones tienen consecuencias que no se limitan al ámbito digital, sino que también afectan a la vida social fuera de la pantalla.

Por último, el tercer grupo se relaciona con **la tecnología y las disciplinas que la hacen posible**, en particular aquéllas que permiten analizar grandes volúmenes de datos y automatizar el tratamiento del lenguaje. Estas herramientas no solo facilitan la interacción, sino que también influyen en la construcción de ideas, opiniones y percepciones dentro del entorno digital.

Partiendo de esta estructura, las secciones siguientes del trabajo se centran en examinar Internet como entorno condicionado, los factores psicológicos implicados en la interpretación moral del discurso y las tecnologías de procesamiento del lenguaje natural que permiten su análisis a gran escala.

2.1. Internet y su conducta social

Internet es un lugar prácticamente sin límites, en una constante expansión y lucha entre el anonimato y la identificación de los individuos. Todos los gobiernos intentan regular su uso, ya sea mediante la censura en los gobiernos más autoritarios o leyes flexibles en instituciones más avanzadas en derechos, como la europea [9]. Muchos gobiernos, en este sentido, ven Internet como una amenaza y, lejos de abordar el problema mediante una regulación progresista de

los derechos, directamente optan por prohibir o limitar su acceso. En 2011 [10] se analizó cuál era el nivel de agresión a nivel de derechos por diferentes países en *Freedom on the Net: A Global Assessment of Internet and Digital Media*, tal y como se recoge en la **Figura 2.1**. Resultando para algunos países una puntuación muy alta en términos de violaciones y privación de libertad, sobre todo una coacción dirigida a activistas que luchan por la libertad de prensa o de opinión [11].

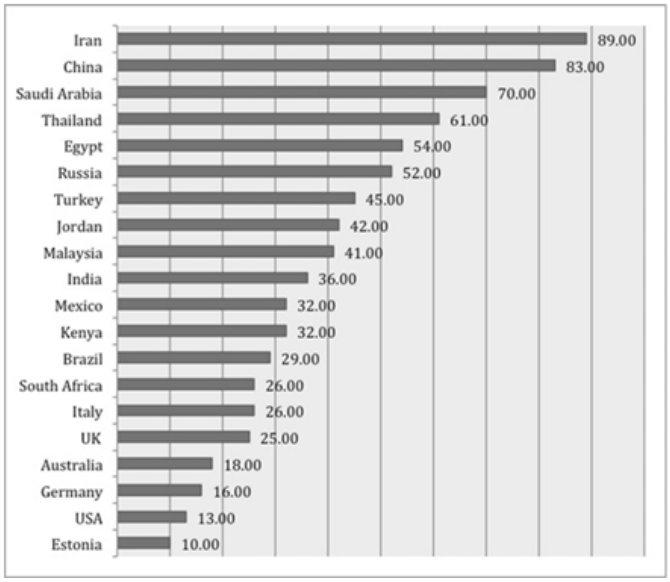


Figura 2.1: Puntuación de libertad gubernamental en Internet (2011): 0 mejor - 100 peor puntuación [10]

No es de extrañar que esto haya tenido un impacto tanto a nivel social como gubernamental; los gobiernos más autoritarios, como el de Irán, ponen grandes esfuerzos en censurar y eliminar del dominio público cualquier información [12]. Mientras que otros, como el de Egipto, limitan su uso dependiendo de la religión a la que se pertenezca.

En su edición más reciente, *Freedom on the Net 2024* [13], se constata un deterioro sostenido en los niveles de libertad en internet a escala global, caracterizado por la expansión del control gubernamental sobre el discurso digital, el incremento de la vigilancia y la proliferación de desinformación promovida por los propios estados. Esta nueva revisión, que abarca 72 países y representa al 87 % de los usuarios globales de internet, permite contrastar la tendencia descendente observada a lo largo de más de una década con los diagnósticos iniciales ya recogidos en estudios previos. Así, el análisis comparativo entre ambas ediciones ofrece una perspectiva crítica sobre la erosión de derechos digitales fundamentales y el impacto de los marcos regulatorios y tecnológicos actuales.

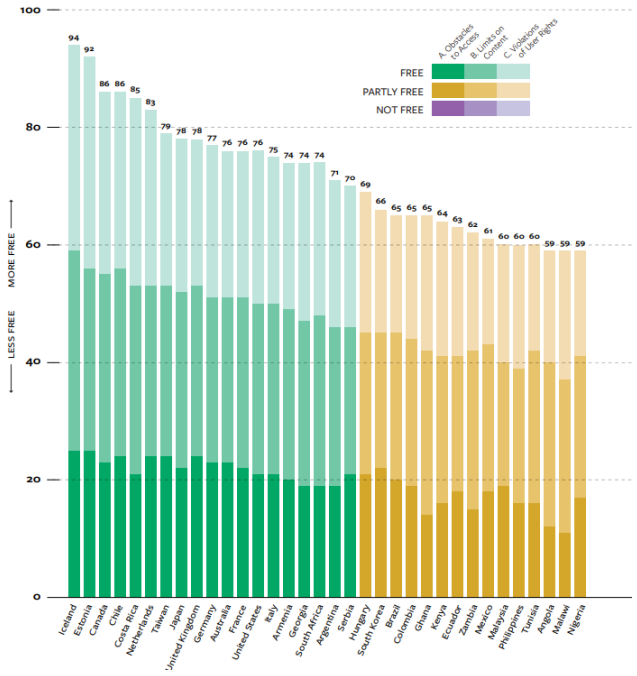


Figura 2.2: Puntuación de libertad gubernamental en Internet (2024) [13]

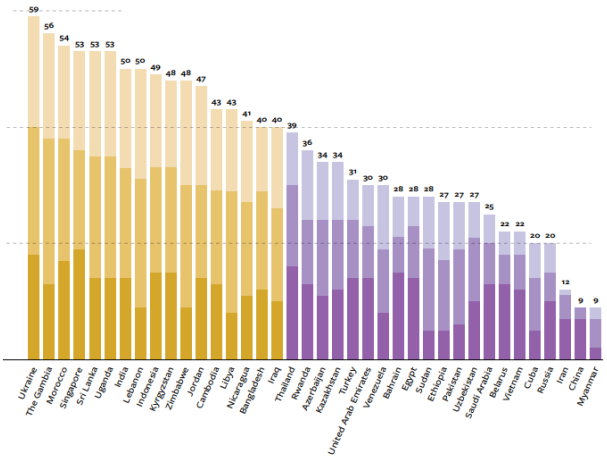


Figura 2.3: Puntuación de libertad gubernamental en Internet (2024) [13]

La información disponible en Internet puede ser interpretada y difundida libremente por cualquier persona con acceso a un dispositivo conectado. Esta característica convierte a la red en un instrumento de doble filo: si bien facilita el acceso rápido y masivo al conocimiento, también favorece la circulación descontrolada de contenidos. En particular, existe una marcada tendencia a compartir información sensacionalista o inexacta, fenómeno que se evidenció de forma es-

pecialmente aguda durante la crisis sanitaria provocada por la COVID-19. En última instancia, esta dinámica nos hace vulnerables a la manipulación informativa, al influir en nuestras percepciones y juicios sin que, en muchos casos, seamos plenamente conscientes de ello [14].

Internet se ha convertido en un medio de comunicación masivo y viral. A diferencia de otros medios, hemos construido perfiles sociales y ha cambiado la forma de relacionarnos. Eso nos ha obligado a tener que participar en la dinámica legal y civil para poder regularlo [15].

Las personas suelen elegir de qué manera consumen o se relacionan y, normalmente, se obedece a lo que más se acerque a nuestros pensamientos o a nuestra ética [16]. Esto crea y favorece un estrecho círculo entre las personas que desean y comunican el mismo interés, sobre lo que llamamos una red social.

Internet es en un medio que influye en nuestro día a día, no solo en términos sociales, sino que también ha supuesto cambios en la manera en que aprendemos y nos desarrollamos. En 2019 apareció el término "*online brain*" en un artículo periodístico en el que se hizo énfasis en cómo afecta Internet a nuestras características cognitivas y nuestra necesidad de estar "en línea" [17]. Este artículo nos presenta aún en una etapa prematura como para evaluar el impacto a largo alcance de estos cambios. Pero existen evidencias y cambios sobre cómo aprendemos, memorizamos o nos relacionamos hoy en día. Presentándose una alta co-dependencia con las circunstancias en las que se desarrolla una persona a lo largo de su vida. Como los problemas de atención derivados de un aprendizaje multitarea visto en adolescentes [18].

Nos aprovechamos de Internet para justificar mentiras, incentivar el odio o la ignorancia mediante una actitud deshonesta a través del anonimato. Muchas veces vemos cómo se traduce el descontento o el enfado en opiniones políticas extremistas [19]. Este uso desmedido y violento de opiniones (principalmente en redes sociales) se conoce como discurso de odio, y suele atender a prejuicios y cuestiones emocionales [20]. Resulta cada vez más complicado desmentir la información: es más voraz, fácil y eficaz desinformar, atacar y atentar contra la verdad [21], la sociedad carece de una valoración crítica previa, sobre qué escucha, qué consume, cómo afecta a su vida y si responde a sus necesidades [22].

En definitiva, la forma de comunicarnos está sujeta a cómo decidimos comunicarnos de forma moral. Esto depende en gran medida de sus circunstancias y la "intensidad moral" [23, 16].

2.2. Psicología moral

El *Annual Review of Psychology* ha recogido a lo largo de varias décadas artículos que recogen diversos aspectos conductuales sobre la moral y muestra tanto nuestro comportamiento moral como nuestra cognición moral se han visto afectados significativamente. En 2016, Elizabeth Mullen y Benoît Monin [24] analizaron la *"consistency"*, dicho del mismo comportamiento perpetuado a lo largo del tiempo por parte de una persona, frente al *"licensing"*, el comportamiento opuesto de forma liberada. Esta revisión de la literatura ([25, 26, 27, 28, 29, 30, 31, 32, 33]) amplifica estos conceptos y sugiere qué circunstancias se dan en los juicios morales, tanto los de hechos positivos como negativos:

- *Consistency* -

- Cuando se piensa de manera abstracta
- Cuando la persona se centra en su compromiso
- Cuando existen inferencias sobre sus valores a partir de su comportamiento inicial

- *Licensing* -

- Cuando se piensa de forma concreta (objetos físicos, experiencias inmediatas...)
- Cuando la persona en cuestión se centra en el progreso realizado
- Cuando existe una discordancia entre el valor moral que se pone a prueba y sus valores
- Cuando se enfrentan conductas ambiguas
- Cuando esa persona se encuentra agotada o exhausta.

Estos elementos nos ayudarán a contextualizar que valores cognitivos justifican o se explican en un determinado juicio moral. Un juicio moral cae en diferentes definiciones y formas dependiendo de variables distintas [34]. Para definir un juicio moral, podemos agrupar todas las variables que participan y resumirlo al balance entre hechos o comportamientos negativos frente a otros positivos en favor de un resultado que no genere desaprobación, o por lo menos la menor posible [35, 36, 37, 38, 39, 40]. En este mismo estudio se ofrece una clasificación de cuatro clases de juicio moral; dependiendo del grado de complejidad que tiene procesar la información en cada uno de ellos. Estas cuatro categorías corresponden con las mostradas en la **Figura 2.4**, siendo la base de ellas la categoría que más información requiere a la hora de establecer un juicio moral y, por lo tanto, la que se encuentra arriba la de menor información.

Figura 2.4: Pirámide de clasificación de juicios morales en base de la cantidad de información que procesa. Fuente: Bertram F. Malle, 2021.*Annual Review of Psychology, Moral Judgments* [34]

Estas dos características nos ponen en contexto de cómo una persona valora un hecho y cómo va a ser su deliberación. Para medir la intensidad del lenguaje en la respuesta, necesitamos medir la intensidad moral [41]. Este modelo recoge las etapas que determinan la manera en que el ser humano traslada su comportamiento moral en base a la intensidad moral. Se dividen en cuatro fases o secciones, cada una alimentando a la siguiente:

1. Reconocimiento del problema moral: el ambiente socio-cultural, económico y organizacional son las principales fuentes que influenciarán a esta fase.
2. Realizar el juicio moral en base a lo identificado por el individuo, donde se ve involucrado el desarrollo de la moral cognitiva.
3. Establecer la intención moral en base al juicio previo.
4. Llegar o participar en un comportamiento moral haciendo uso de la intención y el resultado de su juicio.

En estudios posteriores, como el realizado en 2024 [42], se puso a prueba el modelo teórico de Thomas Jones [41] para analizar si la intensidad moral era un modulador del comportamiento moral en dilemas éticos tecnológicos. Obteniéndose como resultado evidencia favorable sobre cómo se puede extrapolar este comportamiento teórico a un contexto social y tecnológico.

Es decir, somos capaces de identificar qué elementos en una narrativa nos incitan a responder o interactuar. Por ello, a la hora de evaluar escritos o contextos, podremos anticipar qué tipo de respuesta nos incentiva.

2.3. Procesamiento del Lenguaje Natural

El concepto de Inteligencia Artificial (IA) surgió en la década de 1940 gracias a la teoría de la computación de Alan Turing y al test homónimo que publicó en 1949 [43]. Considerado el padre de la computación, Turing definió el marco teórico para evaluar la capacidad de una máquina de reproducir el pensamiento humano con la intención de replicar su lógica. Este nuevo concepto no sólo se aplicó en el campo de la filosofía; Turing sentó las bases teóricas para entender qué era un lenguaje de programación y la arquitectura en la que se sustentaba.

Durante las décadas de 1950 y 1960, surgieron los primeros intentos de procesar el lenguaje humano, de forma que una máquina pudiera replicarlo. En 1966 ve la luz ELIZA, desarrollado por Joseph Weizenbaum, considerado el primer procesador de lenguaje natural [44]. ELIZA intentaba mantener diálogos básicos mediante la técnica de *pattern matching* a través de reglas lingüísticas simples.

Durante las décadas de 1970 y 1980, el enfoque en PLN estuvo dominado por métodos basados en reglas gramaticales y estructuras sintácticas formales, se llevaron a cabo diversos esfuerzos para encontrar aplicaciones para esta incipiente IA, destacando el desarrollo de *chatbots* [45]. El objetivo era crear una IA capaz de mantener una conversación con una persona sin que esta advirtiera que se trataba de una máquina.

En los años 1990, el auge de los datos y la mejora en la capacidad de cómputo impulsaron un cambio hacia enfoques estadísticos. Modelos como los n-gramas comenzaron a utilizarse para estimar la probabilidad de secuencias de palabras [46], permitiendo avances en tareas como traducción automática, corrección

ortográfica y clasificación de textos, por ejemplo, la detección de números manuscritos o el diagnóstico de cáncer de mama. En estos años, surgieron las redes neuronales recurrentes (RNN, *Recurrent Neural Network*) [47], que introdujeron una forma de modelar dependencias temporales y contextuales, mejorando significativamente el rendimiento en tareas secuenciales como el análisis de sentimientos o la generación de texto.

No fue hasta 2012 cuando el concepto de *Deep Learning* se consolidó. Este enfoque, basado en las RNN, permitió el procesamiento masivo de datos gracias a los avances tecnológicos en las GPU (*Graphics Processing Unit*) y la computación en la nube. La clasificación de estos modelos está reflejada en la **Figura 2.5**. Estos modelos nos ayudan a la hora de detectar *fake news*, verificar la autoría de textos, averiguar las características que definen al autor (*author profiling*), entre otros.

A partir de 2017, con la publicación del artículo *Attention is All You Need* [48], desarrollado por ocho investigadores de Google, se produjo un avance significativo en la calidad de las inteligencias artificiales. Este trabajo introdujo una nueva arquitectura denominada *transformer* [49], la cual permitió mejorar la contextualización en la selección de *tokens*, representaciones numéricas del texto más relevantes para el tema principal. Uno de los modelos más extendidos basados en esta arquitectura es el modelo BERT (*Bidirectional Encoder Representations from Transformers*), presentado en 2018 [50], siendo el primer modelo bidireccional preentrenado con solo texto plano. BERT se entrenó en tareas de enmascaramiento de palabras (*masked language modeling*) y predicción de la siguiente oración, lo que lo hace especialmente eficaz en tareas de comprensión y clasificación de texto.

La evolución del *Deep Learning* ha llevado desde el concepto de una máquina capaz de evolucionar en distintos estados hasta redes neuronales que pueden generar o discriminar textos, imágenes y audios. Uno de los modelos más utilizados para la generación de contenido es el modelo GPT, desarrollado por OpenAI, cuya primera versión fue lanzada en 2018. GPT emplea un entrenamiento unidireccional (de izquierda a derecha), optimizado para tareas de generación de texto. Su tercera versión, GPT-3, destacó por su capacidad de producir textos coherentes, responder preguntas, traducir, escribir código y mucho más, con escasos ejemplos (few-shot learning) [51]. Estos modelos marcan un punto de inflexión en el PLN, ya que no requieren ajustes específicos por tarea y generalizan de forma sorprendente a múltiples dominios, gracias al entrenamiento en grandes cantidades de datos textuales.

2.3.1. Diversificación de Modelos y Técnicas en IA

La evolución de la IA ha estado marcada no solo por avances tecnológicos, sino también por una creciente diversificación en los tipos de modelos utilizados, que ha permitido abordar problemas cada vez más complejos y específicos, en contextos que van desde el análisis estadístico tradicional hasta arquitecturas neuronales avanzadas. A continuación, se describen los principales enfoques y su relevancia en distintas tareas.

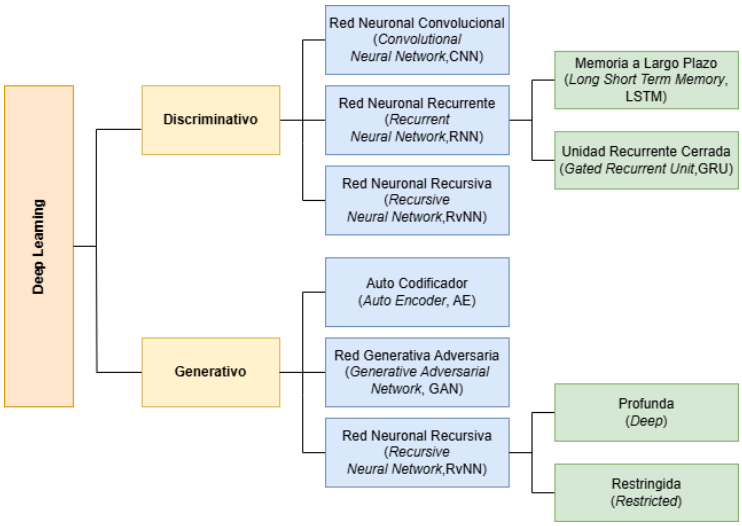


Figura 2.5: Clasificación de modelos de *Deep Learning* Fuente: *Combating multimodal fake news on social media: methods, datasets, and future perspective* [52].

2.3.1.1. Modelos Basados en Árboles

Uno de los enfoques más utilizados en problemas de clasificación y regresión son los modelos basados en árboles de decisión. Entre ellos, el algoritmo *Random Forest* [53] destaca por su robustez frente al sobreajuste y su capacidad para manejar grandes volúmenes de datos con variables tanto categóricas como continuas. *Random Forest* genera múltiples árboles utilizando subconjuntos aleatorios de los datos y características, y luego agrega sus predicciones, lo cual mejora significativamente la precisión del modelo. Este enfoque es particularmente útil en aplicaciones como detección de fraudes [54], clasificaciones de texto [55] y predicción de riesgos médicos [56].

2.3.1.2. Modelos No Supervisados - *Clustering*

Los modelos de aprendizaje no supervisado, en particular los algoritmos de *clustering*, permiten encontrar estructuras ocultas en los datos sin necesidad de etiquetas. Un ejemplo es el *B0-maximal clustering* [57], diseñado para tareas como la verificación de autoría de texto. Este enfoque combina características estilométricas con técnicas de agrupamiento para separar textos de distintos autores. Por su parte, la "*Logical Combinatorial Pattern Recognition*" [58] ofrece una alternativa basada en lógica y teoría combinatoria, útil para clasificar patrones estructurados en contextos donde los métodos estadísticos tradicionales no son suficientes.

2.3.1.3. Modelos Probabilísticos y Estadísticos

Antes de la llegada del aprendizaje profundo, los modelos probabilísticos y estadísticos dominaron muchas tareas de procesamiento de lenguaje. Los modelos de N-gramas [46] se utilizan para predecir la probabilidad de aparición de una palabra dado su contexto anterior, y aún son relevantes en sistemas de

autocompletado, corrección ortográfica y análisis de estilo. Otro modelo clásico es la regresión logística [59], utilizada principalmente en tareas de clasificación binaria. A pesar de su simplicidad, es efectiva cuando se interpretan relaciones lineales entre variables y continúa utilizándose como modelo base o comparativo en contextos de salud, finanzas y ciencias sociales.

2.3.1.4. Redes Neuronales

- Redes neuronales convolucionales (CNN) [60]: originalmente diseñadas para el reconocimiento de imágenes, las CNN también han demostrado utilidad en tareas de texto, como la clasificación de documentos o el análisis de sentimientos, mediante el tratamiento del texto como secuencias estructuradas.

- Redes neuronales recurrentes y su variante *Long Short-Term Memory* (LSTM) [61]: diseñadas para manejar datos secuenciales, las LSTM resuelven el problema del desvanecimiento del gradiente y permiten aprender dependencias a largo plazo, esenciales para el procesamiento de lenguaje natural, traducción automática o generación de texto.

- *Graph Neural Networks* (GNN) [62]: estas redes procesan datos estructurados en forma de grafos (como redes sociales, relaciones entre palabras, o estructuras moleculares) y permiten representar de forma eficiente las conexiones y dependencias entre entidades.

- Redes neuronales siamesas: empleadas en tareas de comparación, como verificación de autoría o autenticación biométrica. Estas redes aprenden una función de similitud entre pares de entradas. Algunas variantes avanzadas incluyen la *Graph-Based Siamese Network* [63], que combina grafos y arquitectura siamesa para capturar relaciones semánticas complejas, y la *Hierarchical Recurrent Siamese Network* (HRSNN) [64], que incorpora estructuras jerárquicas y temporales, ideales para comparar textos largos o documentos técnicos.

- *Deep Bayes Factor Scoring* [65]: un modelo que combina inferencia bayesiana con redes profundas, particularmente útil en contextos donde se necesita evaluar la verosimilitud de una hipótesis, como en la verificación de autoría o análisis forense de textos.

- Redes Neuronales de Base Radial (RBF)[66]: Estas redes emplean funciones de base radial como funciones de activación en su capa oculta, lo que les permite modelar regiones locales del espacio de características. Han sido exploradas en tareas de perfilado de autor por su capacidad para clasificar textos cortos mediante representaciones densas.

Esta diversidad de modelos no solo refleja la madurez técnica del campo, sino también su capacidad de adaptación a múltiples contextos y dominios. Desde modelos interpretables y rápidos como los árboles de decisión, hasta arquitecturas neuronales altamente expresivas, la IA actual cuenta con un repertorio de herramientas para abordar una amplia gama de problemas, desde los más estructurados hasta los más semánticos y contextuales.

El auge en el uso de estas tecnologías ha generado continuos dilemas ético-tecnológicos, con respuestas vagas por parte de las instituciones, pese al esfuerzo de los expertos en señalar y proponer soluciones [67]. La IA se ha convertido no solo en una herramienta para el análisis de datos. Resulta ser también una nueva fuente de generación de textos, tanto es así que ya no bastan las herramientas para evitar el plagio en la educación pública, sino que se están haciendo grandes esfuerzos en reducir los trabajos generados con IA [68, 69]. Uno de los principales problemas que se le atañe a la IA y al *deep learning* es lo que se conoce como "*Black Box dilema*", actualmente los modelos y los datos con los que se entrenan dichas IA permanecen ocultos y no existe un control rígido por parte de sus creadores sobre el contenido que usan[70]. Esto nos impide evaluar la calidad del aprendizaje o posibles vulneraciones intelectuales en materia de derechos[71].

2.4. Difusión de la información

Sabemos que nuestras circunstancias definen en un primer grado nuestra línea de pensamiento, nuestra moral y estas circunstancias serán la base de nuestro juicio. Las emociones acaban siendo quienes regulan el contenido moral que vemos y consumimos en redes sociales. Tanto es así que la presencia de palabras moral-emocionales aumenta un 20 % la transmisión de mensajes en la red social: ***Twitter*** (actual "X") [72]. Explicándose así el efecto de contagio moral dentro de grupos, aumentando un 95 % la interacción dentro de grupos altamente influenciados por estas palabras o este lenguaje moral. Lo que da lugar a que encontremos más interacciones entre estos mensajes; de hecho, ha quedado constatado que un uso del lenguaje moral-emocional acarrea mayores discursos de odio (*hate speech*) como respuesta a este lenguaje. Sobre todo dirigido a personas cuyos perfiles se enmarcan dentro del activismo; al contrario que para otros perfiles que se desarrollan en el ámbito informativo o político [73].

2.4.1. *Fake News*

Las *fake news* son una expresión derivada del término posverdad. En el año 2016, el *Oxford English Dictionary* eligió el término post-truth como palabra del año. Esta palabra describe aquellas "situaciones que derivan en la formación de la opinión pública en apelaciones emocionales y creencias personales antes que en hechos objetivos" [74]. En cambio, el Diccionario de la Real Academia Española define posverdad desde 2017 como "distorsión deliberada de una realidad, que manipula creencias y emociones con el fin de influir en la opinión pública y en actitudes sociales" [75]. La aparición de estos términos obedece a una evolución de la instrumentalización por parte de la élite política, sumado a una "sumisión voluntaria" por parte de la ciudadanía, principalmente la americana y la europea, ya que persiguen desentenderse de la verdad [76].

No es casualidad que esta terminología aparezca en 2016, este mismo año se dieron las campañas de las elecciones estadounidenses que enfrentaban a Donald Trump contra Hillary Clinton. Numerosos estudios han demostrado cómo se han visto afectadas estas elecciones por las *fake news* ([77] [78], [79]). Para el 14 % de la población estadounidense, las redes sociales han sido la principal fuente de información de estas elecciones, creándose así perfectos caldos de cultivo para la proliferación de estas noticias. Esto ha marcado un antes y después, no solo

en la política internacional, sino que es una manifestación directa de cómo el mundo globalizado socializa a través de Internet y el impacto que tienen estas noticias sobre nuestras vidas. [80]

En 2021 [81], se ofreció un estudio del modelo por el cual estas noticias se distribuían por la red social Twitter. Bajo el modelo de dos estados, las *fakes news* se presentan en primera instancia como noticias comunes y ordinarias. En un segundo estado, una vez que los usuarios detectan elementos falsos y no creíbles, esta misma noticia evoluciona a otra noticia nueva. De esta manera, la noticia original queda parcialmente o completamente renovada, por lo que, una vez que queda en entredicho la noticia original, antes de que acabe su verificación, existe otra que compite por desinformar del mismo modo.

La información que se presenta en estas noticias debe pasar un filtro de veracidad para cada individuo, esto depende en gran medida de tres grupos principales que determinan nuestra predisposición a tomar una noticia como real [82]:

1. Las características del mensaje que estimulen nuestras propias creencias y su presentación.
2. Los factores individuales que determinan nuestra interpretación: cognición de la persona, alfabetización, tanto informativa como de las noticias, y nuestra predisposición inicial.
3. Las intervenciones de medios o personas, así como los avisos o los *nudges* (estrategias utilizadas para modificar el comportamiento de las personas, sin que estas las adviertan) que influyan en la precisión de la información y, por tanto, a la hora de juzgar la credibilidad.

¿Por qué proliferan las *fake news* y sus interacciones, resultando en contenido viral? Por una parte, el diseño de los algoritmos que ofrecen contenido a los usuarios [83] hace que este contenido se consuma por perfiles afines. Entendemos que los algoritmos están diseñados y se refinan para mantenernos en la plataforma de consumo. Por otra parte, el impacto emocional del lenguaje incrementa la interacción de este contenido. En 2019 [84] un estudio concluye con cinco categorías que influyen a la hora de compartir estas noticias de forma activa o no, como podemos ver en la **Tabla** 2.1

Influencia positiva	Influencia negativa
Confianza a la hora de compartir información en línea	Comparación social
Compartir información personal	Relación: Confianza en línea – autenticar noticias antes de compartirlas
Miedo a perderse algo (*Fear of Missing Out*)	
Fatiga de las redes sociales	

Tabla 2.1: Influencia a la hora de compartir *fake news* Fuente: [84]

También las normas sociales impulsan en gran medida la intensidad moral con la que los individuos actúan, liberando y sopesando dilemas de privacidad y

seguridad, como fue el caso estudiado de la *German Corona-Warning-APP* [23]. O el aumento de los discursos conspiracionistas en plataformas como *Youtube* o webs específicamente creadas para la difusión de teorías conspiranoicas [85], que destacan del resto si tienen un carácter científico o pseudocientífico.

Como se ha señalado, dos acontecimientos recientes han marcado un punto de inflexión en la propagación masiva de noticias falsas, bulos y teorías conspirativas: las campañas electorales de Estados Unidos en 2016 y, posteriormente, la pandemia de COVID-19 en 2020. Estos eventos evidenciaron la necesidad urgente de desarrollar herramientas eficaces para la detección temprana y la verificación automatizada de información.

En respuesta, se han producido avances significativos gracias al desarrollo de nuevas arquitecturas de procesamiento y al aumento de la capacidad computacional. Un estudio publicado en 2022 ofrece una panorámica actualizada del estado del arte en la lucha contra las *fake news*, recopilando diversas estrategias y sistemas en función de las técnicas de análisis que aplican [52].

En el plano técnico, la mayoría de estos enfoques comienzan con una etapa de preprocesamiento en la que los textos se convierten en vectores, se normalizan y se eliminan elementos no significativos (como preposiciones o signos de puntuación), con el objetivo de optimizar la velocidad y precisión de los modelos aplicados. Esta fase es clave para garantizar que los sistemas de detección puedan operar de forma eficiente ante el creciente volumen de información digital.

En este contexto, el *author profiling* adquiere un papel complementario y estratégico. La identificación de características del autor (como su edad, género, estilo de escritura o incluso patrones ideológicos) puede contribuir significativamente al análisis del origen y la intencionalidad de los contenidos falsos. Al perfilar al emisor, no solo se mejora la trazabilidad de la información, sino que también se posibilita una clasificación más precisa de los mensajes en función de su posible credibilidad o nivel de manipulación. Así, el perfilado de autor se presenta como una herramienta clave en los sistemas integrados de verificación y control de la desinformación.

2.4.2. *Author Profiling*

El *author profiling* es una subdisciplina del procesamiento de lenguaje natural y la estilometría que se define como "análisis de textos con el objetivo de identificar las características del autor" [86]. Estas características permiten clasificar al autor de un texto según su estilo de escritura [87]. Debido al crecimiento del uso de las redes sociales y la generación de contenido multimedia, esta disciplina ha visto un despunte relevante por su uso a la hora de analizar grandes cantidades de datos y la toma de decisiones en base a ellos.

2.4.2.1. Enfoques metodológicos

Los primeros estudios se basaban en la extracción de características estilométricas en textos. La frecuencia del uso de determinadas palabras, la longitud de las oraciones, así como patrones gramaticales o signos de puntuación, son elementos estadísticos que permiten a los algoritmos clásicos como las máquinas de vectores de soporte, la regresión logística o árboles de decisión clasificar al autor del texto. Estos algoritmos suelen tener una limitación en textos breves.

Es una limitación natural ya que la estadística precisa de una gran cantidad del espacio muestral para obtener un mejor resultado.

Con la llegada del *Deep Learning* y su consolidación, los modelos más actuales en vez de analizar todo el conjunto del texto, extraen representaciones semánticas significativas que permiten identificar el contexto. De esta manera se crean proyecciones de palabras en espacios vectoriales continuos, donde el significado de cada palabra queda relacionado y contextualizado con el propio texto. Esta segregación más atómica y relevante del texto sirve de entrada a redes neuronales para obtener resultados muy precisos, incluso en textos breves que son por norma, los que podemos leer en las redes sociales [88]. En este mismo estudio, veremos que la tendencia actual en los talleres PAN es utilizar lenguajes preentrenados (en mayor medida BERT), ya que la arquitectura en la que se sustenta permite trasladar el contexto de forma más precisa y con ello obtener mejores resultados a la hora de clasificar al autor.

2.4.2.2. Aplicaciones

El *author profiling* se ha aplicado extensamente en el análisis de contenido generado por usuarios en redes sociales. Por ejemplo, se ha utilizado para identificar perfiles demográficos de usuarios en Twitter, Reddit y otras plataformas, lo que permite personalizar contenido, detectar comportamientos anómalos y comprender mejor a las audiencias [89].

Según el artículo *Multilingual author profiling on Facebook* [90], la investigación en este campo ha evolucionado hacia tres enfoques principales: análisis basado en el estilo de escritura del autor, en las características inferidas del contenido del texto y en la temática abordada.

Esta división es la que utiliza normalmente PAN a la hora de enunciar sus tareas anuales. PAN es uno de los principales talleres dedicados a la estilometría y el análisis forense de textos. Desde 2007, organiza tareas compartidas en las que equipos de investigadores analizan distintos conjuntos de datos según temáticas específicas [91]. Por ejemplo, en 2013 se enfocaron en la identificación de edad y género a través de redes sociales, especialmente Twitter [92]. Más adelante, en 2016, ampliaron el estudio a otras plataformas, entre las que se incluían blogs [93]. En 2020, debido al impacto de la propagación de *fake news* durante la pandemia de COVID-19, se propuso analizar qué tipos de usuarios tenían mayor predisposición a difundir desinformación [94]. Para 2022, los esfuerzos se centraron en la verificación de autoría, la detección de ironía mediante el análisis de etiquetas, la difusión de estereotipos y el estudio de cambios de estilo [95]. En 2024, se continuó con la línea de 2023 en el análisis del estilo de escritura de múltiples autores y la eliminación de contenido tóxico mediante su reformulación. Además, se introdujo una nueva tarea orientada a diferenciar el pensamiento crítico de las narrativas conspirativas e identificar sus principales factores subyacentes. También se planteó uno de los mayores desafíos en la disciplina: distinguir entre textos generados por inteligencia artificial y aquellos de autoría humana [91].

El *author profiling* constituye sólo una de las múltiples líneas de investigación abordadas en PAN, junto con tareas relacionadas como la detección de cambios de estilo, la atribución de textos generados por inteligencia artificial o la identificación de discursos de odio. Aunque los enunciados específicos varían cada año, la mayoría mantienen como eje común la caracterización del autor a

través de su escritura. La aplicabilidad de estas técnicas en ámbitos tan diversos como la seguridad, la justicia, la educación o el marketing evidencia su creciente relevancia. No obstante, el potencial de estas herramientas conlleva también importantes implicaciones éticas, que deben ser consideradas con rigor antes de su adopción en entornos reales.

2.4.2.3. Dilemas éticos

El análisis de textos con el objetivo de inferir características del autor, junto con la recopilación de datos necesaria para entrenar estos sistemas, puede suponer una amenaza para derechos fundamentales como la privacidad y el consentimiento informado [96, 97]. Esta situación genera un entorno de desconfianza en torno al uso de la inteligencia artificial, especialmente cuando se percibe una falta de transparencia por parte de las entidades que desarrollan estas tecnologías, así como un uso poco regulado o excesivo de las mismas [98].

Por otro lado, los modelos utilizados en el perfilado de autor pueden perpetuar sesgos algorítmicos si se entrenan con datos no representativos o se diseñan sin tener en cuenta principios de equidad [99]. Esto puede derivar en resultados discriminatorios que afectan de forma desigual a determinados grupos sociales [100]. En muchos casos, dichos sesgos no son producto de una intención explícita, sino la consecuencia de decisiones técnicas mal fundamentadas o de conjuntos de datos históricamente sesgados.

En este contexto, el *author profiling* requiere una aplicación ética y responsable, que integre principios de privacidad, transparencia y no discriminación. Solo bajo estos pilares puede garantizarse un uso justo, equitativo y socialmente aceptable de esta tecnología.

Capítulo 3

Desarrollo del proyecto

Este apartado expone el proceso llevado a cabo para analizar el uso de distintas arquitecturas y metodologías aplicadas a tareas de *author profiling* en las competencias PAN celebradas entre 2020 y 2024. El análisis se basa en un corpus de trabajos seleccionados por su desempeño destacado y accesibilidad técnica, lo que ha permitido estudiar en detalle la evolución de los modelos y estrategias empleadas por los equipos participantes.

A lo largo del desarrollo, se describen los criterios de selección del corpus, el procedimiento metodológico seguido para la clasificación de tecnologías, así como las principales tipologías de modelos identificadas. Se presta especial atención al papel emergente de arquitecturas profundas como las redes neuronales recurrentes, convolucionales y basadas en *transformers*, con el objetivo de trazar un mapa claro del estado técnico actual del perfilado de autor en el ámbito competitivo. Los resultados específicos derivados de este análisis se presentan en un apartado posterior.

3.1. Objeto de estudio

Para este estudio se ha seleccionado como objeto de análisis el conjunto de trabajos presentados en las competencias PAN en el marco de la conferencia CLEF durante el período 2020–2024. Estas competencias se centran en el diseño, evaluación y comparación de sistemas automáticos aplicados a tareas específicas como la atribución de autoría, la detección de plagio o el perfilado de autor (*author profiling*).

Se han seleccionado los trabajos con mejor puntuación en cada una de las tareas propuestas, sumando un total de 40 contribuciones. Estos trabajos, en su mayoría de acceso público, están disponibles en el sitio web oficial del evento, salvo contadas excepciones debidamente indicadas en las respectivas revisiones.

Cada edición anual de PAN plantea tareas con un conjunto de datos común, disponible públicamente, y propone a los participantes la resolución de problemas mediante sus propios sistemas. Las soluciones se evalúan de forma estandarizada en función de métricas previamente definidas (por ejemplo, precisión, F1 o exactitud media), y los resultados se publican junto con un resumen técnico del enfoque utilizado por cada equipo. Estos documentos, organizados como working notes, están disponibles a través del repositorio oficial Webis [101].

El criterio de inclusión se ha basado en la selección de los trabajos mejor puntuados en cada una de las tareas específicas relacionadas con el perfilado de autor, alcanzando un total de 40 entradas distribuidas entre distintos idiomas, tipos de texto y formatos de tarea (clasificación, verificación, análisis de estilo, etc.). En todos los casos se han considerado únicamente aquellas contribuciones que ofrecen una descripción técnica suficiente de la metodología implementada, lo que ha permitido una clasificación detallada y un análisis comparativo entre enfoques.

3.2. Clasificación de los trabajos

Una vez delimitado el corpus, se procedió a la clasificación técnica de cada uno de los trabajos en función de las metodologías y modelos aplicados. Esta categorización ha permitido identificar no solo las soluciones más frecuentes, sino también la evolución del enfoque tecnológico a lo largo del período analizado. Para ello, se establecieron etiquetas temáticas basadas en el tipo de arquitectura utilizada, el nivel de preprocesamiento, la estrategia de entrenamiento y la naturaleza de las tareas abordadas (clasificación binaria, multiclase, verificación de autoría, entre otras).

Los trabajos fueron agrupados en cuatro grandes categorías: (i) enfoques tradicionales, como máquinas de vectores de soporte, n-gramas o regresión logística, así como técnicas de agrupamiento y árboles de decisión; (ii) redes neuronales clásicas (CNN, RNN, LSTM, RBF) o sistemas de redes que no utilizan transformadores; (iii) modelos de última generación, como *transformers* y *Large Language Models* (LLM); y (iv) modelos híbridos que combinen distintas tecnologías, planteamientos o modelos. Estos últimos reflejan una tendencia creciente hacia la integración de modelos para mejorar el rendimiento. Esta organización temática sienta las bases para un análisis más detallado del contexto técnico, que se expone en el apartado siguiente.

3.3. Metodología

El proceso de análisis se estructuró en torno a una revisión sistemática de cada trabajo seleccionado, orientada a identificar y documentar las características técnicas más relevantes. Se adoptó una estrategia mixta de análisis cualitativo, basada en la lectura comprensiva de las memorias técnicas, y cuantitativo, mediante la recopilación tabulada de variables comunes. Entre estas variables se incluyeron el tipo de modelo empleado, las técnicas de preprocesamiento de datos, el idioma de los textos, la naturaleza de la tarea y el sistema de evaluación utilizado.

La información fue registrada en una hoja de cálculo estructurada que permitió realizar comparaciones entre años, tecnologías y subtareas. En casos donde los autores empleaban modelos propios o no documentaban completamente el sistema, se procedió a una inferencia controlada basada en los ejemplos y las arquitecturas mencionadas. Se garantizó así una uniformidad analítica suficiente como para facilitar la comparación posterior, sin comprometer la integridad metodológica del estudio. Esta sistematización permitió construir una base sólida sobre la cual se fundamenta el análisis técnico presentado en el siguiente apartado.

Capítulo 4

Resultados

4.1. Resultados obtenidos

En este apartado se presentan los resultados derivados del análisis de los trabajos seleccionados de las competencias PAN (2020–2024). A partir de la clasificación previa de modelos, metodologías y tareas, se han identificado tendencias relevantes en cuanto al rendimiento, la evolución tecnológica y la recurrencia de determinadas estrategias. Los datos recogidos permiten observar no solo qué enfoques han obtenido mejores puntuaciones, sino también cómo se han consolidado ciertas arquitecturas (especialmente aquellas basadas en aprendizaje profundo) frente a técnicas más tradicionales. Los resultados se exponen de forma estructurada para facilitar su interpretación comparativa y su vinculación con los objetivos de análisis definidos previamente.

4.1.1. Comparación de rendimiento por tipo de modelo

El análisis de los 40 trabajos con mejor puntuación revela diferencias claras en el rendimiento según el tipo de modelo empleado. En general, los modelos basados en *transformers* han logrado las prestaciones más altas en las tareas de perfilado de autor recientes [102, 95, 91]. Desde 2021 en adelante, la mayoría de los equipos en lo alto del ránking incorporan *embeddings* (una representación numérica y densa) contextuales de arquitecturas como BERT (y variantes como RoBERTa, BERTweet o modelos multilingües) para representar los textos, superando por margen notable a enfoques más tradicionales [103, 104]. Por ejemplo, en el *shared task* de 2022, los enfoques con BERT alcanzaron precisiones superiores al 95 % e incluso rozando el 99 %, algo fuera del alcance de modelos previos. En 2023, de hecho, prácticamente todos los participantes optaron por ***transformers*** neuronales dada la escasez de datos (*few-shot learning*), y el sistema ganador usó DeBERTaV3 como núcleo de su solución [105].

Por otra parte, los **métodos tradicionales basados en rasgos estilométricos y clasificadores lineales** (SVM, regresión logística, *Random Forest*, etc.) mostraron un rendimiento competitivo, sobre todo en las primeras ediciones del período analizado. En 2020, estos enfoques clásicos dominaron el ránking: las seis mejores soluciones combinaron n-gramas (caracteres y palabras) con clasificadores SVM o regresión logística, superando a cualquier método neuronal profundo ese año [94]. De hecho, el primer modelo basado en *deep learning* apareció en la séptima posición de 2020. Este patrón resalta que, con conjuntos de datos relativamente pequeños o tareas muy enfocadas, los modelos sencillos con buenas características pueden igualar o superar a redes neuronales más complejas. Sin embargo, a medida que aumentó la complejidad de las tareas y la disponibilidad de arquitecturas preentrenadas, los métodos tradicionales perdieron terreno frente a los enfoques neuronales más avanzados.

Los modelos de **redes neuronales no-transformer** (p.ej., CNN, LSTM,

BiLSTM) ocuparon un espacio intermedio en cuanto a desempeño. En 2020, su adopción fue limitada (solo algunos equipos probaron redes neuronales profundas) y sus resultados quedaron por detrás de los clasificadores basados en rasgos. No obstante, en 2021, una red neuronal destacó del resto: el equipo ganador empleó una CNN alimentada con *embeddings* de 100 dimensiones entrenadas *ad hoc*, logrando la mayor precisión promedio (79 %) [106, 107]. Esto demuestra que las redes profundas especializadas aún podían ser efectivas con datos moderados. Aun así, fuera de este caso puntual, la tendencia general favoreció pronto a los modelos con lenguaje preentrenado. Para 2022 y 2023, los enfoques puramente con CNN/LSTM fueron minoría y típicamente se usaron en combinación con *transformers* o como extractores de características dentro de arquitecturas híbridas [104, 95].

Finalmente, destacan los **enfoques híbridos**, que combinan elementos de distintos paradigmas. Varios de los trabajos mejor clasificados integran lo mejor de ambos mundos: utilizan representaciones profundas (*embeddings* de BERT u otras) pero las alimentan a clasificadores tradicionales, o bien emplean *stacks* y ensamblajes de modelos heterogéneos, es decir, arquitecturas compuestas por distintas capas o módulos que cooperan en tareas complementarias dentro del sistema. En 2022 abundaron los *ensembles*: se reportan sistemas que apilan SVM, *Random Forest*, regresión logística y hasta métodos de AutoML, junto con capas neuronales, para meta-clasificación [95]. En esta edición, dos equipos quedaron terceros ecuánimes en puntuación, uno de enfoque híbrido y otro tradicional (97,22 %) [108]; el primero empleó BERT como generador de características, seguido de una red neuronal de múltiples capas (MLP) para la decisión final y el segundo usó un *Random Forest*, alimentado con unigramas [109]. Estos enfoques híbridos suelen alcanzar gran robustez, aprovechando la capacidad de generalización de los modelos profundos y la complementariedad de rasgos lingüísticos específicos.

En resumen, la comparación por tipos de modelo indica que el estado del arte se ha desplazado hacia soluciones centradas en *transformers* (a menudo enriquecidas con componentes adicionales), desplazando a segundo plano las soluciones puramente tradicionales o basadas únicamente en redes neuronales simples.

4.1.2. Tendencias por año (2020–2024)

Este apartado explora la evolución de las soluciones de estilometría y análisis de autoría observadas a lo largo de las competiciones anuales de PAN. Se puede apreciar una evolución notable en las estrategias ganadoras a lo largo de los años 2020 a 2024.

4.1.2.1. 2020: Enfoques Clásicos y Redes Neuronales

En 2020, las tareas de *author profiling* (**detección de propagadores de *fake news***) mostraron un claro dominio de métodos de aprendizaje automático tradicional. Los mejores sistemas de ese año emplearon representaciones sencillas (n-gramas de caracteres/palabras, estadísticas textuales) clasificadas mediante algoritmos lineales clásicos; las precisiones máximas alcanzadas rondaron 0.77–0.82 (77–82 %) [95]. Por ejemplo, el mejor resultado global de PAN 2020 (77.5 % de exactitud promedio en dos idiomas) fue logrado en empate por dos

enfoques puramente tradicionales: uno basado en SVM con n-gramas de caracteres/palabras (mejor en español con 0.820) [110] y otro basado en un conjunto de regresiones logísticas sobre rasgos textuales (mejor en inglés con 0.750) [111]. La conclusión de los organizadores fue que, en esa edición, "los enfoques tradicionales obtuvieron mayores *accuracies* que los de aprendizaje profundo" [94]. Pocos participantes probaron redes neuronales y quienes lo hicieron no lograron superar a las configuraciones de referencia con SVM o LR[1]. Este escenario puede atribuirse al tamaño limitado de los datos de entrenamiento (500 autores por idioma) y a que las señales estilométricas sencillas eran suficientemente informativas para la tarea binaria planteada (perfiles veraces vs. difusores de noticias falsas).

En la tarea de **perfilado de celebridades** de PAN 2020, el objetivo fue predecir la edad, el género y la ocupación de una celebridad utilizando únicamente los tuits de sus seguidores [94]. El equipo de Price y Hodge [112] obtuvo el mejor desempeño entre los participantes, alcanzando una puntuación de 0.5779. Su enfoque se basó en un clasificador de regresión logística para cada categoría demográfica. Las características clave que utilizaron incluyeron la longitud promedio de los tuits, los vectores de palabras promedio de los tuits de los seguidores y las proporciones de etiquetas de parte de la oración (POS *tags*), palabras vacías, tipos de entidades nombradas, enlaces, *hashtags*, menciones y emojis. Este estudio demostró la viabilidad del perfilado basado en seguidores, logrando un rendimiento comparable al de los modelos basados en los propios tuits de la celebridad, especialmente para la predicción de la ocupación.

En la tarea de **verificación de autoría** de PAN 2020, el desafío era determinar si dos textos, específicamente de *fanfiction*, fueron escritos por la misma persona [113]. El equipo ganador [65] logró el mejor rendimiento general en las pruebas con conjuntos de datos pequeños y grandes. Su sistema, una extensión de ADHOMINEM, integró la extracción de características neuronales con el modelado estadístico. Emplearon una red siamesa para mapear documentos a vectores de incrustación lingüísticos (LEVs) de longitud fija, que luego se procesaron con una capa de análisis discriminante lineal probabilístico (PLDA) para la puntuación del factor de Bayes. Implementaron estrategias de preprocesamiento como el enmascaramiento de temas y un enfoque de ventana deslizante con prefijo contextual (etiquetas de *fandom*), lo que mejoró la capacidad del modelo para reconocer variaciones estilísticas dentro del mismo dominio temático. Obtuvieron puntuaciones de 0.940 en el conjunto de datos pequeño y 0.969 en el grande en el conjunto de prueba. Esta edición reveló una tendencia híbrida y especializada en estilometría. Para la detección de cambios de estilo, el aprendizaje profundo resultó superior. PAN 2020 subraya la coexistencia de la solidez clásica y la exitosa expansión del aprendizaje profundo para tareas que se benefician de representaciones semánticas o estilísticas más ricas.

4.1.2.2. 2021: El Comienzo del Cambio

En 2021, se observa un punto de inflexión en las tendencias tecnológicas. La tarea de perfilado de propagadores de *hate speech* atrajo un mayor uso de modelos neuronales y, especialmente, de modelos *transformer* preentrenados. Si bien aún hubo soluciones basadas en rasgos manuales en los primeros puestos,

[1] Regresión Lógica

la mayoría de los participantes incorporó *embeddings* neuronales en su flujo de procesamiento. BERT y sus variantes se convirtieron en herramientas casi estándar, aprovechadas de diversas formas: ya sea *fine-tuning* directo para clasificar autores, como base para modelos híbridos o incluso adaptando métodos de verificación de autoría (un equipo afinó un transformador para replicar el método Impostor de verificación, logrando un segundo puesto *ex-aequo*) [103]. Paradójicamente, el ganador absoluto de 2021 [107] no empleó BERT sino una CNN "hecha a medida", alimentada con representaciones de *embeddings* de palabras de 100 dimensiones. Este enfoque logró una precisión global del 79 % y un destacado 85 % en español. Aun así, los siguientes clasificados ilustraron la transición: otros dos equipos en segundo lugar utilizaron, respectivamente, un *fine-tuning* de transformador combinado con verificación estilométrica (el equipo UO-UPV56 [114]), y un meta clasificador de n-gramas (el equipo MUCIC57 [115]). El mejor resultado en inglés, con un 75 % de precisión, lo obtuvo Dukić y Sović [116], mediante incrustaciones contextualizadas de BERT ajustadas y concatenadas con variables binarias indicadoras (presencia de *hashtag*, *retweets* y URLs). El cuarto puesto experimentó extensamente con BERT, RoBERTa, BERTweet y AutoML; y el resto del top 10 incluyó múltiples combinaciones de BERT+SVM, CNN+SVM, etc. En cuanto al rendimiento, 2021 mostró una ligera mejora sobre 2020 (precisión máxima $\tilde{0}.79$ vs 0.775), especialmente en español donde se alcanzó 85 % de acierto en la detección de haters. Esto sugiere que los modelos de lenguaje preentrenados empezaban a traducirse en ventajas medibles, pese a que la tarea seguía siendo desafiante.

En la tarea de **Verificación de Autoría** [117], que se presentó en un escenario más complejo de dominio cruzado y conjunto abierto, el sistema O2D2 (*Out-Of-Distribution Detector*) [118] obtuvo la puntuación global más alta (95.45 %). Este marco híbrido neuronal-probabilístico de extremo a extremo utilizó un *ensemble* de 21 modelos y un detector para casos indecidibles, demostrando la eficacia de modelar la incertidumbre. Sorprendentemente, se concluyó que la verificación de autoría a gran escala en un conjunto abierto no fue inherentemente más difícil que en un conjunto cerrado, siendo el tamaño del conjunto de datos de entrenamiento un factor clave para el éxito. El segundo mejor resultado [119] (93.59 %) propuso un enfoque innovador basado en una red siamesa basada en grafos para representar textos y extraer características, complementado con características estilométricas.

Finalmente, en la **Detección de Cambio de Estilo**, también se observaron soluciones impulsadas por *deep learning* [106]. Para la Subtarea 1 (clasificación de documentos de autor único vs. múltiples autores), la mejor puntuación [120] mantuvo un F1-score de 0.7954 usando un conjunto apilado (*stacking ensemble*) entrenado con incrustaciones de texto BERT y características estilísticas. Para las Subtareas 2 (posiciones de cambio de estilo) y 3 (atribución de autor a párrafos), las mejores puntuaciones en *F1-scores* fueron de 0.751 y 0.501, respectivamente [121]. El método, SCDWSS (*Style Change Detection based on Writing Style Similarity*), trató la detección de cambios de estilo como un problema de clasificación binaria basada en la similitud del estilo de escritura entre pares de párrafos, utilizando un modelo BERT preentrenado para extraer características y lograr las tres tareas bajo un marco unificado.

4.1.2.3. 2022: Consolidación de los *transformers*

En 2022, las técnicas basadas en *deep learning* y *transformers* consolidaron su predominio en las tareas de perfilado [108, 122, 123]. La competición de ese año, específicamente la tarea de perfilado de difusores de ironía y estereotipos (PAN-AP 2022, también conocida como IROSTEREO), registró resultados sobresalientes en términos absolutos, con varias soluciones logrando más de 95 % de precisión. Los organizadores reportaron que "diferentes *transformers* también han sido ampliamente utilizados para extraer características" [108], incluyendo BERT, SBERT y BERTweet, a menudo combinados con otros métodos. El mejor sistema de 2022 en esta tarea alcanzó una precisión de 99.44 %, considerada prácticamente perfecta. Este sistema, presentado por Yu et al.[124], empleó un modelo híbrido que integraba representaciones BERT con un modelo CNN y un método de *soft voting ensemble*. El segundo lugar (con 97.78 % de precisión) fue logrado por Tahaei et al.[125], quienes combinaron *embeddings* de Sentence-BERT (SBERT) con características de emojis. Su enfoque utilizó una capa de atención aditiva sobre las representaciones [CLS][2] de los *tweets* para obtener una representación vectorial por autor, demostrando la utilidad de añadir información semántica de emociones al modelo. También es notable que uno de los terceros lugares (con 97.22 % de precisión), presentado por Ikae [109], empleó únicamente técnicas tradicionales optimizadas: entrenó un *Random Forest* sobre unigramas cuidadosamente seleccionados. Estos unigramas fueron filtrados utilizando técnicas como Chi-cuadrado ($\chi 2$), PMI (*Pointwise Mutual Information*) y TF-IDF (*Term Frequency-Inverse Document Frequency*) para identificar los rasgos con mayor poder discriminatorio. Este resultado indica que, incluso en 2022, una solución bien diseñada con aprendizaje automático "clásico" podía competir con modelos complejos, aunque en las posiciones más altas, tales casos fueron menos frecuentes.

En la tarea de **Verificación de Autoría** (*Authorship Verification*) en PAN 2022, que consistía en determinar si dos textos de diferentes tipos de discurso (ensayos, correos electrónicos, mensajes de texto y notas comerciales) fueron escritos por el mismo autor, el equipo universitario de la Universidad Nacional Autónoma de México (UNAM) destacó por su enfoque innovador [63]. Este consistió en modelar el texto como un grafo y utilizar una arquitectura de Red Neuronal Siamesa (SNN), compuesta por dos redes neuronales convolucionales de grafos (GNN), para identificar características relevantes. Extrajeron características del texto basándose en la relación de las etiquetas de Parte de la Oración (POS) y la co-ocurrencia de palabras, lo que les permitió capturar información estructural del estilo de escritura independientemente del tipo de discurso. Aunque su puntuación general fue de 0.5856, la más alta en esta tarea obtuvo 0.587 [126]; esta propuesta empleó un modelo de lenguaje T5 como capa base de *embedding*, junto con CNN y un mecanismo de atención. Además, incorporó características estilísticas y gramaticales como etiquetas POS, emojis e información específica del autor y del tema.

Finalmente, en la tarea de **Detección de Cambio de Estilo** (Style Change Detection), el equipo universitario de la National Central University, Taiwán, fue el ganador entre los enfoques intrínsecos[127]. Propusieron una arquitectura unificada de redes neuronales de *ensemble*, afinando modelos *transformer* como

[2]Es el token especial que se coloca al principio de la secuencia de entrada en modelos como BERT

BERT, RoBERTa y ALBERT, y combinando las predicciones finales mediante un mecanismo de *ensemble* de votación por mayoría. Sus puntuaciones F1 promedio fueron de 0.7540 para la Subtarea 1, 0.5100 para la Subtarea 2 y 0.7156 para la Subtarea 3. El equipo de la Foshan University, China [128], obtuvo resultados sólidos con 0.7346 F1 en la Subtarea 1, 0.4687 F1 en la Subtarea 2 y 0.6720 F1 en la Subtarea 3. Emplearon ELECTRA junto con una red neuronal de capas totalmente conectadas (*Fully Connected Neural Network Classifier*), tratando las tareas como problemas de clasificación binaria o multiclase basados en la similitud del estilo de escritura. Otro equipo [129] logró 0.7471 F1 en la Subtarea 1, 0.4170 F1 en la Subtarea 2 y 0.6314 F1 en la Subtarea 3. Su método utilizó un modelo preentrenado BERT y una red neuronal convolucional unidimensional (Conv1D) para extraer información de características y clasificar la similitud de los textos.

En general, la tendencia de 2022 fue hacia sistemas sofisticados, frecuentemente ensamblando múltiples componentes. Por ejemplo, se observaron combinaciones como: *transformers* (BERT, T5, SBERT) con redes neuronales convolucionales; uso de mecanismos de atención para extraer características importantes y determinar relaciones entre *tokens*; y *ensembles* de clasificadores, que combinan modelos tradicionales (como SVM y *Random Forest*) con técnicas de *deep learning*. Las tareas de ese año aparentemente presentaban patrones identificables que los *transformers* capturaron muy eficazmente, lo cual se reflejó en las elevadas métricas globales.

4.1.2.4. 2023: Especialización y LLMs para *Data Augmentation*

En 2023, el panorama de las tareas de autor se orientó a nuevos desafíos, en particular el perfilado de influenciadores en criptomonedas con *few-shot learning*. A diferencia de años anteriores, aquí cada autor venía representado por muy pocos textos (máximo 10 *tweets* en la primera subtarea, y solo 1 *tweet* en las otras subtareas de interés e intención) [104], lo cual dificultó considerablemente la clasificación, ya que el entrenamiento de modelos de lenguaje con un gran número de parámetros se vuelve un reto con tan pocas instancias. Este cambio impulsó aún más la dependencia en modelos preentrenados potentes: la mayoría de los equipos aplicó grandes *transformers* ajustados a datos escasos. Estos modelos se pueden agrupar en tres categorías principales: modelos *encoder* (como BERT, RoBERTa, DeBERTa, ELECTRA, CryptoBERT), modelos *encoder-decoder* (como T5) y modelos *decoder* (como GPT y BLOOM) [130]. Algunos equipos exploraron estrategias de *prompting* o *fine-tuning* adaptativo [131]. El equipo ganador general del PAN 2023 en esta tarea fue el equipo NLP-CIMAT (Villa-Cueva et al.) [131]. Este equipo integró DeBERTaV3 para codificar los textos, obteniendo el mejor desempeño promedio en los tres subtemas. DeBERTaV3 mejora el modelo original DeBERTa al reemplazar el modelado de lenguaje enmascarado (MLM) con la detección de *tokens* reemplazados (RTD), una tarea de preentrenamiento más eficiente en cuanto a muestras, lo que optimiza la eficiencia de entrenamiento y la calidad del modelo [105]. Otros equipos también probaron T5 (*encoder-decoder*) con entrenamiento especial para pocos datos, como se observó en los *baselines* oficiales que emplearon T5-large con *bi-encoders* (para *zero-shot*) y *label tuning* (para *few-shot*). Para paliar la escasez de datos, se aplicaron técnicas de aumento de datos sintéticos, incluyendo el uso de ChatGPT para generar autores o tweets sintéticos, así como el modelo

Pegasus paraphrase y la traducción inversa (*back-translation*). Los resultados en 2023, medidos en Macro F1 debido a la naturaleza multiclase desequilibrada de las subtareas, fueron más modestos en términos numéricos, reflejando la dificultad intrínseca de perfilar con tan poca información por autor. Los mejores resultados individuales para cada subtarea fueron:

- Subtarea 1 (Influencia): 62.32 % obtenido por el equipo holo[130], que utilizó el modelo RoBERTuito.
- Subtarea 2 (Interés): 67.12 % logrado por el equipo stellar [131], que combinó el aumento de datos con un conjunto de modelos transformadores
- Subtarea 3 (Intención): 67.46 % conseguido por el equipo terra-classic[130], ajustando DeBERTaV3

Hay que destacar que casi el 46 % de las soluciones superaron al mejor *baseline* provisto por los organizadores, y solo una quedó por debajo de la asignación aleatoria. Estos *baselines* incluyeron una asignación aleatoria, T5-large con *bi-encoders* (*zero-shot*), T5-large con *label tuning* (*few-shot*), *n-grams* de caracteres con regresión logística, y el método LDSE (*Low-Dimensionality Statistical Embedding*). Esto evidencia un progreso: incluso en un entorno adverso de pocos ejemplos, las técnicas modernas (preentrenamiento masivo, adaptaciones *few-shot*) permitieron generalizar patrones útiles.

En definitiva, 2023 continuó la supremacía de los *transformers*, confirmando que son la herramienta central para afrontar tanto tareas clásicas de perfilado como variantes emergentes de bajo recurso.

4.1.2.5. 2024: Dominio de los LLMs y el Aprendizaje *Zero-Shot/Few-Shot*

Las tareas de PAN 2024 abordaron desafíos clave en el procesamiento del lenguaje natural, destacando el uso de LLMs y *fine-tuning* como enfoques estándar. Las innovaciones incluyeron la integración de técnicas avanzadas de ajuste, el aumento de datos, el procesamiento a nivel de oración y el manejo de "negaciones de generación". La tarea de **Verificación de Autoría de IA Generativa** tuvo como objetivo distinguir entre textos escritos por humanos y aquellos generados por IA, considerándolo un problema de clasificación binaria [91]. El equipo MarSan [132] se destacó al obtener el primer lugar, logrando un ROC-AUC (área bajo la curva) perfecto de 1.0 y una puntuación Brier[3] cercana a 1.0 en el conjunto de datos de prueba principal. Su solución, BinocularsLLM, integró el *fine-tuning* supervisado de LLMs como LLaMA2 y Mistral con una *classification head* y el *framework* "Binoculars", que evalúa la perplejidad y la entropía para la detección de texto generado por máquina. El equipo fOSU-STU [133] también abordó esta tarea utilizando un modelo BERT preentrenado y proponiendo el método *Tri-Sentence Analysis* (TSA) para capturar información contextual de grano fino, especialmente útil para textos cortos, e incorporando el método MPU (*Multiscale Positive-Unlabeled detection of AI-generated texts*) para mejorar la eficiencia y diferenciación en textos breves. El **Análisis de Pensamiento Opositor** se dividió en dos subtareas: la clasificación binaria de textos como

[3]Métrica de evaluación que se utiliza para evaluar la precisión de las predicciones probabilísticas de un modelo

conspiratorios o críticos (Subtarea 1) y la detección a nivel de *span* de elementos narrativos específicos (Subtarea 2) [91]. Los conjuntos de datos, en inglés y español, contenían 5,000 comentarios de Telegram cada uno, relacionados con la pandemia de COVID-19. Para la Subtarea 1, el equipo IUCL [134] obtuvo el primer lugar en inglés utilizando un modelo DeBERTa ajustado con una *sequence classification head* y aplicando un aumento de datos significativo, que quintuplicó el conjunto de entrenamiento original a aproximadamente 20,000 textos. El equipo SINAI [135] logró el primer lugar en español basándose en el uso de LLMs como GPT-3.5 y LLaMA3-8B-instruct, ajustados mediante *instruction fine-tuning*, siendo este crucial para que los modelos aprendieran las diferencias entre clases y observando que GPT-3.5 superó a LLaMA3-8B-instruct en español. En la Subtarea 2, los mejores resultados tanto en inglés como en español fueron obtenidos por el equipo Tulbure y Coll Ardanuy[136]. Su método consistió en ajustar modelos *transformer* y aplicar técnicas de aumento de datos mediante la sustitución de palabras por sinónimos o palabras semánticamente relacionadas. Trataron el problema como una tarea de clasificación de *tokens* y segmentaron el texto en oraciones para mitigar las limitaciones de longitud de los modelos *transformer*, asegurando que no se perdiera información. La tarea de **Desintoxicación de Texto Multilingüe** tuvo como objetivo cambiar el estilo de un texto de tóxico a no tóxico en nueve idiomas [137]. La evaluación final se basó en el juicio humano, considerando la transferencia de estilo, la similitud de contenido y la fluidez. El equipo SmurfCat [138] ocupó el primer lugar en la evaluación automática y el segundo en la evaluación humana final. Su solución se presenta mejor analizada en el apartado de **Casos relevantes y soluciones destacadas**. El equipo SomethingAwful [139] fue el mejor en la evaluación humana para varios idiomas. Utilizaron modelos LLaMa-3 (70B) "sin censura" con una estrategia de "*few-shot prompting*" (dar pocos ejemplos al modelo) y una técnica de "*jailbreaking*" de alineación para superar las "negaciones de generación" del modelo, es decir, cuando el modelo se niega a generar contenido por considerarlo sensible. Para el idioma amárico, utilizaron mT0-XL [140] debido a las limitaciones de LLaMa-3 con idiomas menos comunes. Finalmente, el **Análisis de Estilo de Escritura Multi-Autor** se centró en detectar los puntos de cambio de autoría. El documento a analizar estaba compuesto por textos de distintos usuarios, cuya dificultad variaba en función del grado de similitud temática entre los fragmentos [141]. El equipo NYCU-NLP [142] se clasificó en el primer lugar en el nivel "difícil" y segundo en el nivel "medio". Su sistema combinó el ajuste de modelos *transformer* preentrenados (RoBERTa, DeBERTa, ERNIE), un mecanismo de ensamblaje por votación y ajustes de similitud basados en incrustaciones de LaBSE (*Language-agnostic BERT sentence embedding*) para mejorar el rendimiento, especialmente cuando los párrafos tenían diferentes temas. El equipo fOSU-STU [143] utilizó LLaMA-3-8B con un ajuste supervisado por etiquetas (*label-supervised fine-tuning*) y *Low-Rank Adaptation* (LoRA), una técnica que permite ajustar el modelo de manera eficiente, reduciendo los costos de entrenamiento y despliegue. Observaron un rendimiento menor en el conjunto de datos "fácil" debido a su desequilibrio de clases.

En resumen, los resultados de PAN 2024 reflejan una consolidación de tendencias que venían gestándose en ediciones anteriores: el dominio creciente de los grandes modelos de lenguaje, la sofisticación en las estrategias de *fine-tuning* y ensamblaje, y una atención cada vez mayor a aspectos complejos del lenguaje como el estilo, la moralidad o la autoría generada por IA. La incorporación de

técnicas de aumento de datos y la adaptación a idiomas menos representados evidencian una madurez.

4.1.3. Observaciones según el tipo de tarea (clasificación vs. verificación vs. cambios de estilo)

Además de las tendencias generales, es importante distinguir las diferencias de enfoque entre las tareas de clasificación de autor (propias del *author profiling* tradicional), verificación de autoría, análisis de moralidad y análisis de escritura de múltiples autores. En las tareas de *author profiling* consideradas (2020–2023), el problema se formula típicamente como una clasificación supervisada, ya sea binaria o multiclase: a cada autor se le asigna una etiqueta de perfil (por ejemplo, propagador frente a no propagador de desinformación, tipo de influenciador, etc.) en base a sus textos. Este enfoque permite entrenar clasificadores discriminativos a partir de ejemplos etiquetados por clase, y las soluciones mejor posicionadas suelen aprovechar al máximo esta estructura: emplean modelos entrenados de extremo a extremo para predecir la categoría del autor, optimizando métricas globales como *accuracy* o F1. Como se ha observado, la incorporación de grandes modelos de lenguaje ha potenciado notablemente la efectividad de estas tareas, alcanzando en algunos casos precisiones superiores al 95 %, especialmente cuando las señales estilísticas son evidentes, como en los casos de ironía.

Por el contrario, las tareas de verificación de autoría (*Authorship Verification*) —también recurrentes en PAN entre 2020 y 2023— plantean un desafío diferente: determinar si dos documentos han sido escritos por la misma persona, generalmente sin disponer de muchas muestras por autor. Se trata de un problema de tipo *one-class* o pareado, más abierto, que algunos investigadores consideran "más fundamental y, por lo general, más exigente" que la clasificación cerrada [117]. En este contexto, un sistema debe captar la similitud estilística entre dos textos, a veces incluso pertenecientes a dominios o registros distintos, como se ha evaluado en PAN 2020–2023. Este enfoque complica el uso de clasificadores tradicionales, ya que no existen clases fijas ni ejemplos suficientes por autor en el entrenamiento. En consecuencia, los métodos más exitosos en verificación tienden a diferir de los usados en perfilado: se han utilizado redes neuronales siamesas o modelos de distancia —que generan *embeddings* estilísticos y calculan similitudes—, así como técnicas de *fine-tuning* específicas para pares de texto. Por ejemplo, en PAN 2021, el mejor sistema de verificación *cross-domain* combinó múltiples representaciones, incluyendo *transformers* ajustados y modelos de *ensemble*, optimizados en función de métricas *ad hoc* como AUC, F1 y C@1 [118]. Es habitual el uso de umbrales de decisión —propios de tareas de *matching*— en lugar de una clasificación categórica directa.

De forma análoga, las tareas de detección de cambio de estilo (*Style Change Detection*, SCD) han evolucionado significativamente entre 2020 y 2024. Estas tareas buscan identificar transiciones estilísticas dentro de un documento, lo que puede implicar detectar múltiples autores, reconocer fragmentos con estilos distintos o incluso atribuir fragmentos específicos. Inicialmente, el análisis se enfocaba en el nivel de párrafo, pero progresivamente ha avanzado hacia el análisis a nivel de oración, en contextos textuales más cohesionados y con menor variabilidad temática. Las soluciones mejor puntuadas comparten una base metodológica común: el uso de modelos de lenguaje preentrenados, espe-

cialmente *transformers*. La incorporación de técnicas como *fine-tuning*, modelos en *ensemble* y estrategias de *data augmentation* ha sido clave para mejorar el rendimiento. En años como 2022 y 2021, se consolidó el uso de representaciones semánticas profundas mediante modelos como BERT o Electra; incluso en 2020, ya se alcanzaban buenos resultados mediante *embeddings* combinados con clasificadores como *Random Forest*. Aunque la dificultad de estas tareas ha variado según factores como la homogeneidad temática o la granularidad de los segmentos, la mayoría de los sistemas han superado las líneas base establecidas, lo que evidencia un progreso sostenido. En conjunto, los resultados sugieren que la combinación de modelos preentrenados, estrategias de adaptación y enfoques híbridos ha sido determinante para abordar eficazmente las tareas de estilo, si bien persisten retos en los escenarios más complejos.

Una observación relevante es que, a pesar de las diferencias metodológicas entre tareas, se ha producido una convergencia progresiva en el uso de técnicas, especialmente en torno a los modelos basados en *transformers*. Varios equipos de *author profiling* han adoptado estrategias propias de la verificación de autoría para mejorar la clasificación, como el uso de modelos ajustados mediante el método Impostor, adaptados luego a tareas de odio o desinformación. De forma inversa, en las tareas de verificación se han empleado modelos previamente entrenados para la clasificación de autor, reutilizados para evaluar similitud estilística entre pares de textos. Esta simbiosis técnica ha sido posible gracias a la versatilidad de las representaciones generadas por los *transformers*, que permiten capturar rasgos estilísticos útiles tanto para identificar perfiles como para comparar estilos entre textos. No obstante, los resultados cuantitativos evidencian que la dificultad varía notablemente entre tareas: mientras que la clasificación de autor alcanza con frecuencia métricas superiores al 90 % de precisión en escenarios bien definidos, las tareas de verificación suelen presentar mayor variabilidad, especialmente en entornos *cross-domain*, con valores de F1 que rara vez superan el 0.80. En paralelo, las tareas de detección de cambio de estilo —a medio camino entre clasificación y verificación— han mejorado de forma sostenida gracias a enfoques híbridos y técnicas de refinamiento contextual, aunque siguen representando un reto técnico, sobre todo en niveles de dificultad más altos. En definitiva, aunque cada tarea plantea desafíos específicos, los avances metodológicos y la reutilización flexible de modelos han permitido mejorar el rendimiento en todo el espectro de tareas evaluadas en PAN.

4.1.4. Casos relevantes y soluciones destacadas

A lo largo de las ediciones de PAN entre 2020 y 2024, se han presentado numerosas propuestas que no solo han obtenido buenos resultados, sino que también reflejan enfoques representativos de las distintas arquitecturas empleadas: desde modelos tradicionales y redes neuronales, hasta *transformers* y soluciones híbridas. Este apartado recoge algunos de los casos más relevantes, seleccionados por su rendimiento sobresaliente y por ejemplificar de manera clara las fortalezas, limitaciones e innovaciones asociadas a cada tipo de modelo. Estas soluciones permiten ilustrar cómo han evolucionado las estrategias técnicas en *author profiling* y qué patrones metodológicos se consolidan en cada edición.

4.1.4.1. Mejor Solución basada en *transformers*: SmurfCat (PAN 2024 - Detoxificación de Texto Multilingüe)

Su solución[138] se basó en el modelo mT0-XL, mejorado con aumento de datos (mediante traducción automática y un filtrado especial) y la aplicación de la técnica de alineación ORPO (*Odds Ratio Preference Optimization*), un algoritmo que no requiere un modelo de referencia adicional para optimizar el comportamiento del modelo. El equipo SmurfCat presentó esta solución en la tarea de Detoxificación de Texto Multilingüe (Tarea 2) en PAN 2024. Esta tarea tenía como objetivo principal "desintoxicar" un corpus multilingüe, es decir, reescribir textos tóxicos de manera no tóxica mientras se preservaba el contenido principal, aplicándose en 7 idiomas.

El enfoque central del equipo SmurfCat se basó en la familia de modelos mT0, que son modelos *transformer* de secuencia a secuencia, inicializados a partir de mT5. Las técnicas clave empleadas fueron:

- Ajuste Fino (*Fine-tuning*): Los modelos mT0 seleccionados fueron ajustados finamente para cada uno de los idiomas de la competición.
- Aumento de Datos: Se aplicaron diversas técnicas de aumento de datos para enriquecer el conjunto de entrenamiento.
- Búsqueda por Haz Diversa (*Diverse Beam Search*): Durante la inferencia, se generaron 10 hipótesis y se seleccionaron las 5 más probables utilizando una búsqueda por haz diversa. La mejor candidata se eligió basándose en una métrica de relevancia, calculada como el producto de la similitud (utilizando *embeddings* LaBSE) y las puntuaciones de toxicidad (utilizando un clasificador de toxicidad xlm-roberta-large).
- Optimización por Preferencia de Razón de Probabilidades (ORPO): Para un ajuste adicional y la mejora del rendimiento, se aplicó la alineación ORPO. Es un algoritmo de optimización de preferencias monolítico que no requiere un modelo de referencia, a diferencia de otros métodos. Para el entrenamiento ORPO, se generaron hipótesis utilizando la búsqueda por haz diversa y se anotaron con puntuaciones de relevancia, seleccionando las de mayor relevancia como preferidas y las demás como rechazadas[144].

La solución de SmurfCat logró el primer lugar en la evaluación automática y el segundo lugar en la evaluación manual realizada por humanos. Su modelo mostró el mejor rendimiento en la evaluación automática para todos los idiomas, superando a modelos más grandes como mT0-XXL (13 mil millones de parámetros). Este enfoque se clasifica como puramente basado en *transformers* porque su componente central es una arquitectura basada en mT0, y todas las mejoras posteriores, como el ajuste fino y la alineación ORPO, se aplican directamente para potenciar las capacidades de este modelo.

4.1.4.2. Mejor Solución Híbrida: ADHOMINEM de Boenninghoff (PAN 2020/2021 - Verificación de Autoría)

El enfoque ADHOMINEM del equipo de Boenninghoff destacó en la tarea de Verificación de Autoría en PAN 2020, y su rendimiento se mantuvo como uno de los mejores en 2021[65, 118]. Esta solución se describió como una "extensión

sustancial" de su enfoque ADHOMINEM anterior, combinando extracción de características neuronales con modelado estadístico. Las características clave de su metodología incluyen:

- Extracción de Características Neuronales: El sistema emplea redes neuronales para extraer características de los textos. Estas características neuronales se interpretan tanto desde un punto de vista métrico como probabilístico.
- Modelado Estadístico: Las características obtenidas se integran en un marco de modelado estadístico para la decisión final de clasificación.
- Aumento de Heterogeneidad de Datos: Un aspecto crucial fue la recombinación de pares de documentos después de cada época de entrenamiento, lo que aumentó significativamente la heterogeneidad de los datos de entrenamiento y, por lo tanto, la robustez del modelo.
- Gestión de Incertidumbre: El sistema permite la opción de no dar una respuesta (un "*non-response*") en casos de alta incertidumbre, por ejemplo, cuando la probabilidad de predicción se acerca a 0.5.

El enfoque ADHOMINEM obtuvo puntuaciones de rendimiento excepcionales, superando a todos los demás sistemas participantes en la tarea de Verificación de Autoría de PAN 2020 en ambos desafíos (conjuntos de datos pequeños y grandes). En 2020, alcanzó una puntuación de 0.9281. Su rendimiento siguió siendo robusto en 2021, con una puntuación de 0.950, manteniendo su posición de liderazgo. Este enfoque se considera híbrido debido a su combinación explícita y efectiva de técnicas de redes neuronales (para la extracción de características) y métodos de modelado estadístico (para la clasificación final), aprovechando las fortalezas de ambas aproximaciones para lograr un rendimiento superior. En la **figura** 4.1, podemos apreciar los 5 módulos que componen la solución:

1. Extracción de Características Neuronales y Aprendizaje Métrico Profundo (Vectores de Representación Lingüística junto LSTM)
2. Evaluación Profunda mediante Factor de Bayes
3. Modelado de Incertidumbre y Adaptación
4. Inferencia por Conjunto de Modelos (como último paso)
5. Detector de Datos Fuera de Distribución (O2D2) [145]

4.1.4.3. Mejor Solución Tradicional: Pizarro (PAN 2020 - Perfilado de Autores de Noticias Falsas)

La solución presentada por Pizarro [110] en la tarea de Perfilado de Autores de Noticias Falsas de PAN 2020 destaca como la mejor entre aquellas basadas en enfoques tradicionales. Su propuesta se apoyó en técnicas clásicas de aprendizaje automático y procesamiento del lenguaje natural, utilizando representaciones textuales basadas en n-gramas —tanto de caracteres como de palabras— y aplicando máquinas de soporte vectorial (SVM) como clasificador principal.

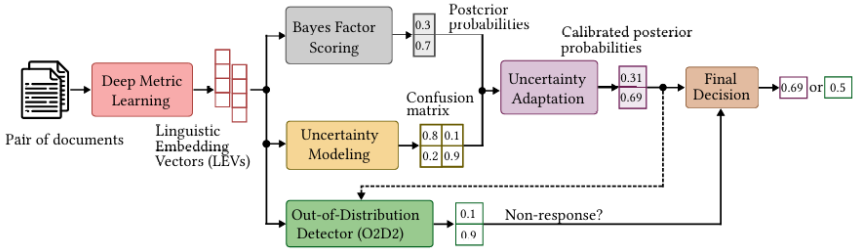


Figura 4.1: Arquitectura híbrida - *O2D2: Out-Of-Distribution Detector to Capture Undecidable Trials in Authorship Verification* [118].

Esta combinación permitió identificar patrones estilométricos relevantes sin necesidad de recurrir a arquitecturas neuronales complejas, lo que resultó en un rendimiento altamente competitivo dentro del contexto de la tarea. En concreto, la metodología logró el mejor resultado para el idioma español, alcanzando una precisión del 82 % y superando otras aproximaciones, como la solución de ensamble de Buda y Bolonyai [111], que obtuvo el mejor desempeño en inglés. El enfoque de Pizarro se considera plenamente tradicional, dado que se basa exclusivamente en técnicas de ingeniería de características manual y en algoritmos convencionales de clasificación, sin integrar redes neuronales profundas ni modelos *transformer*, los cuales definen gran parte de las soluciones más recientes en el campo.

4.1.4.4. Consideraciones sobre la interpretación de los resultados y análisis de mejores resultados

Es importante subrayar que las puntuaciones obtenidas en cada tarea PAN no deben interpretarse de forma aislada ni como indicadores absolutos de superioridad técnica. Tal como se señala en el pie de nota correspondiente, una puntuación alta o baja no implica necesariamente que una solución sea mejor o peor, sino que debe analizarse en función del nivel de dificultad, la naturaleza de la tarea y las condiciones del conjunto de datos. Las comparaciones solo son válidas cuando las tareas mantienen su estructura, objetivos y métricas fundamentales. En consecuencia, la lectura de los resultados debe realizarse desde un enfoque contextualizado, que tenga en cuenta la complejidad de la tarea, el grado de ambigüedad lingüística, la calidad del *dataset* y la cantidad de ejemplos disponibles para el entrenamiento.

En este marco, las **Tablas** 4.1, 4.2, 4.3, 4.4 resumen de manera estructurada los resultados más relevantes obtenidos por los equipos participantes entre 2020 y 2024, segmentados por tipo de tarea: análisis multiautor, verificación de autoría, *author profiling* y tareas de análisis moral. En cada tabla, se destacan las soluciones con mejor rendimiento dentro de su contexto específico, resaltando aquellas que han liderado cada edición. Esta disposición permite observar patrones de progresión tecnológica y consolidación de enfoques —como el predominio de *transformers* en etapas recientes—, así como excepciones notables, como el caso de sistemas tradicionales que, bajo ciertas condiciones, logran competir eficazmente con modelos de última generación.

La disposición cronológica y temática de las tablas permite también identificar cómo algunas tareas, como la verificación de autoría, tienden a presentar puntuaciones más moderadas debido a su complejidad intrínseca, mientras que otras, como el perfilado de emisores de ironía o estereotipos, alcanzan cifras excepcionalmente altas gracias a una mayor regularidad estilística en los textos analizados. Asimismo, los resultados reflejan el impacto de técnicas complementarias como el *data augmentation*, los modelos híbridos y los mecanismos de votación o ensamblaje, que han resultado especialmente eficaces en tareas de perfilado con pocos ejemplos por clase.

Este enfoque analítico no solo permite comparar tecnologías, sino que proporciona criterios más sólidos para evaluar su aplicabilidad en contextos reales, donde el rendimiento técnico debe ir acompañado de fiabilidad, equidad y explicabilidad.

Año	Subtipo de tarea	Trabajo	Puntuación
2024	Style Change Detection	[143]	**0,9027**
2024	Style Change Detection	[142]	**0,8947**
2023	Style Change Detection	[146]	0,8797
2023	Style Change Detection	[147]	0,8780
2022	Style Change Detection	[127]	0,6532
2022	Style Change Detection	[128]	0,6222
2022	Style Change Detection	[129]	0,5934
2021	Style Change Detection	[121]	0,6683
2021	Style Change Detection	[120]	0,6420
2020	Style Change Detection	[148]	0,7484
2020	Style Change Detection	[57]	0,6489

Tabla 4.1: *Multi-author Analysis* 2020–2024

Año	Subtipo de tarea	Trabajo	Puntuación
2024	Generative AI Detection	[132]	0,9240
2024	Generative AI Detection	[149]	0,9210
2023	Authorship Verification	[150]	0,6227
2023	Authorship Verification	[151]	0,6143
2022	Authorship Verification	[63]	0,5788
2022	Authorship Verification	[126]	0,5728
2021	Authorship Verification	[118]	**0,9545**
2021	Authorship Verification	[119]	**0,9359**
2020	Authorship Verification	[65]	0,9350
2020	Authorship Verification	[152]	0,9016

Tabla 4.2: *Author Identification* 2020–2024

Año	Subtipo de tarea	Trabajo	Puntuación
2023	Profiling Cryptocurrency...	(*Overview* 2023)[130]	0,6392
2023	Profiling Cryptocurrency...	[131]	0,6334
2023	Profiling Cryptocurrency...	(*Overview* 2023)[130]	0,6054
2022	IROSTEREO	[124]	**0,9944**
2022	IROSTEREO	[125]	**0,9778**
2021	Profiling Hate...	[107]	0,8500
2021	Profiling Hate...	[116]	0,7500
2020	Profiling Fake...	[110]	0,7775
2020	Profiling Fake...	[111]	0,7775
2020	Celebrity Profiling	(*Baseline Overview*)[153]	0,6460
2020	Celebrity Profiling	[112]	0,5993
2020	Celebrity Profiling	[154]	0,5353

Tabla 4.3: *Author Profiling* 2020–2024

4.2. Objetivos logrados

El presente apartado tiene como propósito evaluar en qué medida se han alcanzado los objetivos definidos al inicio del proyecto, tanto a nivel conceptual como metodológico y técnico. A lo largo de este trabajo se propuso analizar el estado actual del *author profiling* en el marco de las tareas PAN, con especial atención a su evolución tecnológica, sus aplicaciones prácticas, los modelos em-

Año	Subtipo de tarea	Trabajo	Puntuación
2024	Oppositional Thinking Analysis	[134]	**0,8380**
2024	Oppositional Thinking Analysis	[135]	**0,8290**
2024	Oppositional Thinking Analysis	[136]	0,6279
2024	Multilingual Text Detoxification	[139]	0,7740
2024	Multilingual Text Detoxification	[138]	0,7410
2023	Trigger Detection	[155]	0,5500
2023	Trigger Detection	[156]	0,5460

Tabla 4.4: *Morality* 2020–2024

pleados, y las implicaciones del uso de inteligencia artificial en tareas de análisis de autoría. Cada uno de estos objetivos ha sido abordado desde un enfoque sistemático, combinando una revisión del estado del arte con un análisis comparativo de los trabajos más relevantes publicados entre 2020 y 2024.

4.2.1. Identificar patrones tecnológicos predominantes

Este objetivo se cumple de manera exhaustiva en la sección de resultados, que desglosa la evolución tecnológica año por año, desde 2020 hasta 2024.

Adopción progresiva de redes neuronales profundas: El análisis demuestra claramente esta tendencia. En 2020, se observa un dominio de métodos de aprendizaje automático tradicional en tareas de perfilado de autor, con pocos participantes experimentando con redes neuronales, y sin que estas superaran a las configuraciones clásicas. Sin embargo, ya en 2021, la detección de cambio de estilo mostró soluciones impulsadas por *deep learning*. Para 2022, se consolida el predominio de las técnicas basadas en *deep learning* y *transformers* en tareas de perfilado, logrando precisiones superiores al 95 %. La conclusión general para 2023 es la "supremacía de los *transformers*, confirmando que son la herramienta central".

Aparición de modelos basados en *transformers* y LLM (*Large Language Models*): Se rastrea el origen de los *transformers* hasta el artículo "*Attention is All You Need*" de 2017 y la aparición de BERT en 2018. A partir de 2021, se señala que los modelos de lenguaje preentrenados comenzaron a mostrar "ventajas medibles". En 2022, se reporta que "diferentes *transformers* también han sido ampliamente utilizados para extraer características", incluyendo BERT, SBERT y BERTweet. Para 2024, se destaca el "dominio de los LLMs" en tareas como la verificación de autoría de IA generativa, con modelos como LLaMA2 y Mistral logrando resultados sobresalientes. La sección 4.1.4.1 también detalla la solución SmurfCat[138] (PAN 2024) basada en *transformers* (mT0) y técnicas de alineación.

Uso de arquitecturas híbridas y técnicas tradicionales frente a soluciones basadas en aprendizaje profundo: Se realiza una comparación explícita. Queda subrayado que en 2020, los enfoques tradicionales como SVM y regresión logística con n-gramas "obtuvieron mayores *accuracies* que los de

aprendizaje profundo" en el perfilado de autores de noticias falsas. Sin embargo, se presenta a ADHOMINEM como la "Mejor Solución Híbrida" en 2020/2021 para verificación de autoría, la cual "integró la extracción de características neuronales con el modelado estadístico". Además, la "Mejor Solución Tradicional" se ejemplifica con Pizarro [110] (PAN 2020), que usó n-gramas y Máquinas de Soporte Vectorial (SVM), logrando el mejor resultado en español para su tarea. Esta dicotomía y la emergencia de enfoques híbridos son analizadas con detalle.

4.2.2. Evaluar el rendimiento relativo de las soluciones

Este objetivo se aborda directamente a través de las comparaciones de puntuaciones y la discusión de las métricas empleadas en cada tarea y edición.

Tipo de tecnología empleada y naturaleza de la tarea: Se dedica la sección 4.1.3 a "Observaciones según el tipo de tarea (clasificación vs. verificación vs. cambios de estilo)". Aquí se contrasta el rendimiento de diferentes tecnologías según la tarea. Por ejemplo, en tareas de clasificación de autor (como perfilado de *haters*), se observa que los modelos de lenguaje han permitido alcanzar precisiones superiores al 95 %. En contraste, las tareas de verificación de autoría, consideradas más complejas por su naturaleza de "*one-class* o pareado" y su frecuente configuración *cross-domain*, suelen tener una "mayor variabilidad" en los resultados y sus valores de F1 "rara vez superan el 0.80". La detección de cambio de estilo, aunque desafiante, ha mostrado "progreso sostenido" gracias a enfoques híbridos y técnicas de refinamiento contextual, con la combinación de modelos preentrenados, adaptación y técnicas híbridas siendo "determinante". Por otro lado, el multilingüismo se ejemplifica con la tarea de Detoxificación de Texto Multilingüe (PAN 2024).

Evolución de un mismo enfoque técnico en diferentes ediciones: Se traza la trayectoria de ciertas soluciones. Por ejemplo, el sistema ADHOMINEM[118] mantuvo su "posición de liderazgo" desde 2020 (con 0.928) hasta 2021 (con 0.950) en la verificación de autoría, demostrando la robustez de su enfoque híbrido. Asimismo, se observa la evolución de las soluciones en la Detección de Cambio de Estilo, donde los enfoques basados en *deep learning* fueron "superiores" en 2020, y para 2021 se detallan soluciones que combinan *stacking ensembles* con *embeddings* de texto BERT y características estilísticas, con mejoras notables. Para 2024, se menciona el continuo desarrollo en detección de cambio de estilo con modelos *transformer* ajustados y ensamblaje por votación.

4.2.3. Detectar estrategias recurrentes y su mejora a lo largo del tiempo

Quedan definidos los patrones de continuidad y mejora en las estrategias de los equipos.

Mantenimiento de líneas continuas de trabajo: La recurrencia de equipos y metodologías es un tema subyacente. Se menciona la continuidad de las tareas en PAN, señalando que "continúan y avanzan tareas compartidas pasadas", y que la edición de 2022 "continuó en la misma línea". El caso de ADHOMINEM[118] es el más claro, con su éxito sostenido y su "extensión sustancial" a lo largo de los años.

Optimización y reconfiguración de soluciones: Se proporcionan varios ejemplos: La solución SmurfCat[138] en 2024 no solo utilizó *transformers*, sino

que los optimizó con aumento de datos (mediante traducción automática) y la técnica de alineación ORPO, que mejora el comportamiento del modelo. El equipo SINAI[135] para el Análisis de Pensamiento Opositor en 2024 ajustó LLMs como GPT-3.5 y LLaMA3-8B-instruct, aplicando aumento de datos y segmentando el texto para superar limitaciones. El equipo NYCU-NLP [142] en la tarea de Análisis de Estilo de Escritura Multi-Autor en 2024 combinó el ajuste fino de *transformers* (RoBERTa, DeBERTa, ERNIE) con un mecanismo de ensamblaje por votación y ajustes de similitud basados en embeddings de LaBSE. Estos ejemplos ilustran cómo las estrategias se refinan y adaptan a las nuevas capacidades tecnológicas.

4.2.4. Comparar tareas similares entre distintas ediciones

Comparaciones cualitativas y cuantitativas: El documento estructura sus resultados por año y por tipo de tarea, lo que facilita la comparación. Por ejemplo, en Verificación de Autoría, se compara explícitamente el rendimiento entre PAN 2020 (*fanfiction*) y PAN 2021 (*cross-domain*), y luego con PAN 2022 (*cross-discourse types*), mostrando cómo la tarea se ha vuelto más desafiante con la diversidad de tipos de discurso. En el Perfilado de Autores de Noticias Falsas (2020) y Perfilado de Difusores de Hate Speech (2021), se observa la transición de la dominancia tradicional a la ventaja de los *transformers*. La tabla 4.1 también ofrece una comparación cuantitativa de los resultados de *Multi-author Analysis* desde 2020 hasta 2024. Se señala que, a pesar de las diferencias metodológicas, se ha producido una "convergencia progresiva en el uso de técnicas, especialmente en torno a los modelos basados en *transformers*", y analiza cómo las dificultades varían entre clasificación y verificación, con la primera logrando métricas superiores al 90 % y la segunda con mayor variabilidad (rara vez $>$0.80 F1)

4.2.5. Proponer una tipología de soluciones técnicas aplicadas al author profiling sobre un escenario

Este objetivo se aborda mediante la categorización de las soluciones más destacadas y la contextualización del estudio dentro de problemas sociales relevantes. **Tipología de soluciones técnicas:** Se dedica la sección 4.1.4 a los "Casos relevantes y soluciones destacadas", donde clasifica explícitamente las tres "mejores" soluciones según el enfoque:

- Mejor Solución basada en *transformers*
- Mejor Solución Híbrida
- Mejor Solución Tradicional

Aplicación sobre un escenario (impacto moral en redes sociales): Aunque no se desarrolla una aplicación concreta bajo el apartado de resultados, sí se enmarca su estudio en la relevancia social y política de la identificación de autores y el análisis de contenido en redes sociales. En la introducción se establece el contexto de "bulos, mentiras, discursos de odio" y la necesidad de "sistemas de verificación de contenido y la regulación de su consumo". El *author profiling* se presenta como una herramienta clave para "identificar perfiles demográficos

de usuarios en Twitter, Reddit y otras plataformas", "detectar comportamientos anómalos" y contribuir al análisis de la "intencionalidad de los contenidos falsos". La solución de Pizarro[110], por ejemplo, se aplica directamente al "Perfilado de Autores de Noticias Falsas", un escenario con claras implicaciones morales. Por lo tanto, el análisis de los distintos enfoques se realiza con la perspectiva de su potencial aplicación en estos escenarios de impacto moral, validando las técnicas que podrían emplearse.

4.3. Problemas encontrados

El desarrollo de este trabajo ha supuesto un desafío técnico considerable, especialmente debido a la complejidad inherente al campo del *machine learning* y su aplicación al procesamiento del lenguaje natural. Si bien mi formación en ingeniería del software proporciona una base sólida, abordar con profundidad modelos avanzados como redes neuronales recurrentes, arquitecturas *transformer* o grandes modelos de lenguaje ha requerido una dedicación adicional para alcanzar una comprensión adecuada de sus fundamentos, funcionamiento y aplicabilidad.

Una de las principales dificultades ha sido familiarizarme con conceptos y metodologías propios de disciplinas como la estadística aplicada, la representación semántica del lenguaje o la evaluación de modelos en entornos multiclase y multilingües. No obstante, la documentación técnica disponible en los artículos revisados ha sido de gran ayuda para acercarme progresivamente a una comprensión funcional de las estrategias implementadas, permitiéndome identificar patrones, estructuras y criterios de evaluación con mayor seguridad a medida que avanzaba el análisis.

Además del componente técnico, uno de los retos más significativos ha sido incorporar al trabajo una dimensión crítica que no se limitara a valorar el rendimiento de los modelos, sino que también señalara sus implicaciones sociales y éticas. Esta intención de contextualizar tecnológicamente los resultados desde una perspectiva responsable ha supuesto un ejercicio de reflexión constante, particularmente en la identificación de aspectos como los sesgos de representación, la equidad en la inferencia o el papel de estas tecnologías en la configuración del discurso público.

A pesar de estas dificultades iniciales, la experiencia ha sido altamente enriquecedora. He adquirido conocimientos básicos pero sólidos sobre los fundamentos de la inteligencia artificial moderna aplicada al lenguaje, y he desarrollado competencias para interpretar, comparar y evaluar sistemas complejos con mayor criterio. Al mismo tiempo, considero especialmente valioso haber podido articular —aunque sea de forma introductoria— una mirada crítica sobre cuestiones que, en mi opinión, son centrales en el debate contemporáneo sobre tecnología y sociedad. Señalar estos problemas no solo ha sido un ejercicio académico, sino también una tarea profundamente interesante y formativa.

Capítulo 5

Conclusiones y trabajos futuros

Este apartado final presenta una síntesis de los hallazgos más relevantes del estudio, así como una reflexión crítica sobre su alcance e implicaciones. En primer lugar, se recogen las conclusiones generales extraídas a partir de los resultados. A continuación, se evalúa el posible impacto social y medioambiental del uso de IA en este ámbito, especialmente en relación con su despliegue en contextos digitales sensibles. Por último, se plantean varias líneas futuras que podrían guiar nuevas investigaciones, incluidas propuestas para el diseño de nuevas tareas centradas en la dimensión moral y política del contenido en redes sociales.

5.1. Conclusiones

Este trabajo ha logrado satisfactoriamente los objetivos analíticos que se propuso, proporcionando una revisión del estado del arte en las competiciones PAN entre 2020 y 2024. Identifica y analiza con precisión los patrones tecnológicos predominantes, destacando el auge de las redes neuronales profundas, *transformers* y LLMs, y contrastándolos con las soluciones híbridas y tradicionales. Evalúa el rendimiento relativo de estas soluciones en función de la tecnología y la naturaleza de las tareas (clasificación, verificación, cambio de estilo, multilingüismo), ofreciendo comparaciones cuantitativas y cualitativas que muestran la evolución y la dificultad inherente de cada desafío. Además, detecta estrategias recurrentes y cómo los equipos las han optimizado con el tiempo, aportando ejemplos concretos. Los principales hallazgos se pueden resumir de la siguiente manera:

- Consolidación de los modelos basados en *transformers* y LLMs: se ha observado una adopción progresiva y consistente de arquitecturas de red neuronal profunda, especialmente *transformers* y grandes modelos de lenguaje, que han pasado a ocupar un lugar central en las soluciones más exitosas, especialmente en tareas de *author profiling*.
- Relevancia persistente de enfoques tradicionales e híbridos: a pesar del auge del *deep learning*, los métodos clásicos —basados en rasgos estilométricos y clasificadores lineales— así como las soluciones híbridas, han demostrado ser estrategias robustas, particularmente en los primeros años analizados, y siguen ofreciendo resultados competitivos en ciertos escenarios.
- Diferenciación por tipo de tarea: el rendimiento de los modelos varía en función de la naturaleza del problema abordado. Mientras que las tareas de clasificación de autor han alcanzado precisiones superiores al 95 %,

las tareas de verificación de autoría y de detección de cambio de estilo presentan un mayor grado de dificultad, aunque también muestran una mejora constante gracias a la flexibilidad de los modelos preentrenados.

En conjunto, este estudio evidencia la consolidación del campo del *author profiling* como una disciplina técnicamente madura, con modelos cada vez más sofisticados y adaptables. Las soluciones presentadas en las ediciones recientes de PAN reflejan una capacidad creciente para afrontar desafíos complejos en el análisis de textos digitales, destacando especialmente su utilidad en contextos críticos como la identificación de desinformación, la detección de discurso de odio o la trazabilidad de contenidos generados por IA.

5.2. Impacto social y medioambiental

El presente trabajo de fin de grado, centrado en el análisis del *author profiling* mediante IA a partir de las competiciones PAN (2020–2024), tiene un impacto social y medioambiental que debe ser valorado en términos tanto de sus aportaciones como de sus implicaciones.

Desde el punto de vista social, este estudio contribuye a una mejor comprensión de cómo las tecnologías de perfilado de autor pueden aplicarse para identificar discursos dañinos, manipulación informativa o patrones ideológicos en entornos digitales. En un momento de especial sensibilidad ante la desinformación, la polarización y los discursos de odio en redes sociales, este trabajo ofrece un análisis riguroso sobre los modelos, estrategias y resultados que están configurando el estado del arte en esta área. Al sintetizar y evaluar las soluciones presentadas por los principales equipos de investigación a nivel internacional, se aporta una base sólida para futuras decisiones académicas, técnicas y legislativas sobre el uso ético de la IA en contextos comunicativos.

Además, este trabajo destaca los dilemas éticos asociados al uso de modelos de lenguaje para inferir características personales, como el riesgo de sesgos algorítmicos, la falta de transparencia y la posible vulneración del derecho a la privacidad. Al incluir una reflexión explícita sobre estos aspectos, se promueve una aproximación crítica a la tecnología, fomentando la alfabetización digital y el pensamiento ético en torno al desarrollo y aplicación de sistemas inteligentes. Este enfoque puede resultar especialmente útil para otros estudiantes, investigadores y responsables de políticas que busquen integrar principios de justicia, equidad y transparencia en sus propios proyectos.

En términos medioambientales, aunque este trabajo no ha desarrollado modelos nuevos ni ha realizado entrenamiento intensivo de redes neuronales, sí se apoya en el análisis de arquitecturas que conllevan altos costes computacionales, como los *transformers* y los LLMs. Al estudiar de forma crítica la evolución hacia modelos cada vez más grandes y complejos, el trabajo invita implícitamente a reflexionar sobre la sostenibilidad del desarrollo tecnológico. La recopilación y comparación de soluciones más eficientes o híbridas aporta una perspectiva útil para considerar alternativas que equilibren rendimiento y consumo energético.

En conjunto, el impacto de este trabajo se materializa en tres dimensiones: (1) como aportación técnica y académica al conocimiento del estado del arte en *author profiling*; (2) como impulso al debate ético sobre los límites del análisis automatizado de autoría; y (3) como catalizador de prácticas más sostenibles

y responsables en el diseño de soluciones basadas en IA. Aunque sus efectos no son inmediatos ni directos, sienta las bases para una investigación crítica, informada y alineada con los retos sociales y medioambientales del presente.

5.3. Lineas futuras

A partir de los resultados obtenidos y del análisis de las tareas propuestas por PAN en los últimos años, se abre una línea de investigación especialmente prometedora orientada al estudio del impacto moral y el consumo de contenido político en redes sociales. Aunque se han abordado problemas cercanos —como la detección de discurso de odio, el análisis de ironía o la desintoxicación textual—, aún no se ha consolidado una tarea específica que explore de forma sistemática cómo los usuarios interactúan con contenidos de carga ideológica, ni cómo estos influyen en la polarización, la radicalización o el posicionamiento moral. Proponer una nueva tarea PAN que combine técnicas de *author profiling* con evaluación moral contextual permitiría avanzar en la comprensión de los mecanismos discursivos que configuran la opinión pública en entornos digitales, y contribuiría a desarrollar herramientas más responsables para la moderación de contenido, la educación mediática o la prevención de la manipulación algorítmica. Esta línea representa una intersección clave entre IA, análisis sociopolítico y ética computacional.

5.3.1. Propuesta de Tarea Futura PAN: Análisis del Impacto Moral y Ético en el Discurso Digital (AI-MEDD)

Las competiciones PAN , en el marco de la conferencia CLEF, se han consolidado como un referente en el análisis forense y estilométrico de textos digitales. Históricamente, PAN ha abordado desafíos computacionales críticos como la verificación de autoría, la detección de plagio, el perfilado de autor (incluyendo la identificación de edad, género, personalidad, bots y propagadores de noticias falsas o discursos de odio) y la detección de cambios de estilo. Los enfoques tecnológicos predominantes han evolucionado desde métodos clásicos (como n-gramas y SVM) hasta la dominación de modelos basados en transformadores y grandes modelos de lenguaje (LLM). Esta evolución ha permitido abordar tareas cada vez más complejas y con menos datos, como el perfilado con *few-shot learning* o la detección de texto generado por IA. Las fuentes ya reconocen la existencia de dilemas éticos asociados a estas tecnologías. Se mencionan explícitamente cuestiones legales y éticas en la recopilación y análisis de datos de redes sociales, así como el riesgo de que los modelos de IA perpetúen sesgos algorítmicos si se entrenan con datos no representativos, lo que puede conducir a resultados discriminatorios. Además, se subraya la necesidad de una aplicación ética y responsable de la IA que integre principios de privacidad, transparencia y no discriminación. La difusión de desinformación, discurso de odio y estereotipos, que son focos actuales de PAN, ya tiene un impacto social y ético significativo en la sociedad.

5.3.1.1. Justificación de la Nueva Tarea

Aunque PAN ya aborda la identificación de contenido dañino, la dimensión moral y ética subyacente a estos fenómenos no siempre se analiza de forma explícita. El discurso digital no solo es una cuestión de veracidad o polarización, sino también de los valores y principios morales que se invocan, manipulan o subvierten. Las fuentes señalan cómo las emociones moldean la difusión de contenido moralizado y cómo los fundamentos morales pueden correlacionarse con la propagación del discurso de odio. La creciente complejidad del entorno informativo, exacerbada por la IA generativa, hace imperativo ir más allá de la mera detección de contenido problemático. Necesitamos comprender cómo los actores digitales (humanos y máquinas) construyen narrativas con implicaciones morales, qué valores apelan o atacan, y cuál es el impacto ético de sus comunicaciones. Una tarea dedicada a este fin permitiría desarrollar tecnologías más sensibles a la dimensión humana del lenguaje, fundamentales para fomentar un entorno digital más ético y una ciudadanía con mayor pensamiento crítico.

5.3.1.2. Propuesta Detallada: Análisis del Impacto Moral y Ético en el Discurso Digital (AIMEDD)

El objetivo de este proyecto es desarrollar y evaluar tecnologías que permitan identificar, clasificar y analizar las dimensiones morales y éticas presentes en el discurso digital. De esta forma, buscamos comprender su impacto en la opinión pública y el comportamiento social, con la finalidad última de informar el desarrollo de aplicaciones que determinen el impacto moral de los mensajes en redes sociales e instituciones.

En primer lugar, una de las tareas clave del proyecto es la identificación de juicios morales y su clasificación. Basándonos en la clasificación de juicios morales propuesta por Bertram F. Malle[34], se identificará el tipo de juicio moral expresado en un texto, atendiendo a la cantidad de información procesada para establecerlo. Este enfoque nos permitirá categorizar el texto según el tipo de juicio moral dominante, ayudando a contextualizar los valores cognitivos que subyacen en los juicios expresados y a identificar las diferentes complejidades en los juicios morales según el contexto.

En paralelo, abordaremos la medición de la intensidad moral del lenguaje. Esta tarea va más allá del análisis tradicional de sentimientos, ya que se enfocará en la carga ética o moral de las palabras y frases, evaluando su capacidad para generar una respuesta emocional y su influencia en el comportamiento moral de los usuarios. La medición de la intensidad moral será esencial para entender cómo ciertos mensajes con carga ética elevada pueden incrementar la propagación de contenido viral, especialmente en contextos como redes sociales donde el impacto emocional es un factor clave para la interacción.

Además, se incluirá la detección de manipulación del juicio y sesgos algorítmicos. En este sentido, se identificarán técnicas que distorsionan la realidad o manipulan los juicios morales del lector, como el uso de "nudges" que influyen sin ser percibidos. También se evaluará si el contenido generado por sistemas de IA perpetúa sesgos algorítmicos que pueden resultar en discriminación o desinformación. Esta tarea es crucial para abordar los dilemas ético-tecnológicos surgidos por el uso de la IA y las preocupaciones sobre la falta de transparencia en la forma en que los algoritmos manejan la información.

La caracterización ético-moral del autor será otra de las tareas fundamentales. A partir de una colección de textos de un autor, se analizará su tendencia a expresar ciertos juicios morales y si su estilo de escritura refleja patrones que puedan contribuir a la difusión de desinformación o discursos de odio. Este perfilado de autor complementará las investigaciones previas en PAN, ayudando a identificar a los propagadores de noticias falsas y estereotipos, lo que mejorará la trazabilidad de la información y permitirá una clasificación más precisa de los mensajes.

Por último, se llevará a cabo una auditoría ética de contenido generado por IA. Esta tarea tiene como objetivo extender la verificación de autoría para evaluar si los textos producidos por modelos de lenguaje generativo presentan sesgos morales, promueven desinformación ética o carecen de la sensibilidad moral presente en los textos humanos. En este sentido, se definirá un conjunto de criterios claros para evaluar la "alineación ética" de la IA, permitiendo garantizar un desarrollo más responsable de estas tecnologías.

Para llevar a cabo estas tareas, será necesario contar con *datasets* ricos en anotaciones de juicios morales y dilemas éticos, los cuales podrían derivarse de corpora existentes, como los de *fake news* o debates argumentativos. Además, se deben incorporar datos multilingües para asegurar la aplicabilidad del análisis a diferentes contextos culturales. Los modelos utilizados incluirán redes neuronales avanzadas y modelos basados en transformadores como BERT, RoBERTa, GPT, entre otros. Técnicas como el *fine-tuning* y el aumento de datos serán claves para optimizar la capacidad de los modelos y mejorar su desempeño en tareas complejas de análisis moral.

El impacto de este proyecto no se limita solo a mejorar la capacidad para identificar y mitigar la desinformación o los discursos de odio. También contribuirá a fomentar la alfabetización digital y el pensamiento crítico entre los usuarios. El análisis de cómo los algoritmos afectan la conformación de percepciones morales será fundamental para entender cómo los sistemas digitales influyen en la sociedad. Además, este proyecto permitirá mejorar el diseño de sistemas de IA responsables, en especial en lo que respecta a la moderación de contenido y la generación de texto.

En resumen, esta propuesta busca tender un puente sólido entre el análisis computacional del lenguaje y ámbitos socioculturales, reconociendo que el impacto del discurso digital va más allá de la información, afectando profundamente la esfera moral y ética de los individuos y las comunidades.

Bibliografía

[1] Webis Group. *PAN*. https://pan.webis.de/. Accessed: April 26, 2025. 2025.

[2] CLEF 2025 Conference y Labs of the Evaluation Forum. *CLEF 2025 Conference and Labs of the Evaluation Forum*. https://clef2025.clef-initiative.eu/. Accessed: April 26, 2025. 2025.

[3] Cadena SER. *Hay un ecosistema que potencia la desinformación y los discursos de odio: "El periodismo tiene que frenarlo"*. Accedido el 10 de mayo de 2025. 2025. URL: https://cadenaser.com/galicia/2025/03/05/bulos-redes-sociales-y-medios-hay-un-ecosistema-que-potencia-la-desinformacion-y-los-discursos-de-odio-el-periodismo-tiene-que-frenarlo-radio-galicia/.

[4] El Periódico. *El Gobierno detecta un aumento de discursos de odio en redes sociales contra menores migrantes no acompañados*. Accedido el 10 de mayo de 2025. 2024. URL: https://www.elperiodico.com/es/sociedad/20240702/gobierno-detecta-redes-sociales-odio-menores-migrantes-feijoo-bulos-104831155.

[5] Público. *Los discursos de odio se normalizan: necesitamos herramientas para confrontarlos*. Accedido el 10 de mayo de 2025. 2025. URL: https://www.publico.es/opinion/columnas/discursos-odio-normalizan-necesitamos-herramientas-confrontarlos.html.

[6] El Español. *La inteligencia artificial mina la confianza online y promueve una regresión hacia la conexión con el mundo real*. Accedido el 10 de mayo de 2025. 2024. URL: https://www.elespanol.com/invertia/disruptores/grandes-actores/tecnologicas/20241204/inteligencia-artificial-mina-confianza-online-promueve-regresion-conexion-mundo-real/906159463_0.html.

[7] La Vanguardia. *Los desafíos de la inteligencia artificial: los bulos en redes agudizan la polarización*. Accedido el 10 de mayo de 2025. 2025. URL: https://www.lavanguardia.com/vida/20250205/10353130/desafios-inteligencia-artificial-bulos-redes-agudizan-polarizacion.html.

[8] El País. *Cómo la nueva inteligencia artificial puede manipularte como votante*. Accedido el 10 de mayo de 2025. 2023. URL: https://elpais.com/ideas/2023-09-17/como-la-nueva-inteligencia-artificial-puede-manipularte-como-votante.html.

[9] Casey L Addis y Thomas Lum. «US initiatives to promote global Internet freedom: Issues, policy, and technology». En: *Congressional Research Service* (2011).

[10] S. Kelly y S. Cook. «Freedom on the net 2011 : a global assessment of internet and digital media.» En: *Freedom House* (2011). URL: http://www.freedomhouse.org/uploads/fotn/2011/FOTN2011.pdf.

[11] Wikipedia contributors. *Freedom of speech — Wikipedia, The Free Encyclopedia*. [Online; accessed 6-December-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Freedom_of_speech&oldid=1259398726.

[12] Bulent Tarman y Mehmet Fatih Yigit. «The Impact of Social Media on Globalization, Democratization and Participative Citizenship». eng ; ger. En: *Journal of social science education* 12.1 (2013), págs. 75-80. ISSN: 1618-5293.

[13] Allie Funk, Hjalti Vesteinsson y Kian Baker. «The Struggle for Trust Online». En: *Freedom on the Net 2024*. Ed. por Allie Funk et al. Accessed June 2025. Freedom House, 2024. URL: https://freedomonthenet.org.

[14] Kirill Solovev y Nicolas Pröllochs. «Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media». En: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022, págs. 3706-3717. ISBN: 9781450390965. DOI: 10.1145/3485447.3512266. URL: https://doi.org/10.1145/3485447.3512266.

[15] Sherly Haristya. «The efficacy of civil society in global internet governance». En: *Internet Histories* 4.3 (2020), págs. 252-270. DOI: 10.1080/24701475.2020.1769892. eprint: https://doi.org/10.1080/24701475.2020.1769892. URL: https://doi.org/10.1080/24701475.2020.1769892.

[16] Linda Klebe Trevino. «Ethical Decision Making in Organizations: A Person-Situation Interactionist Model». En: *The Academy of Management Review* 11.3 (1986), págs. 601-617. ISSN: 03637425. URL: http://www.jstor.org/stable/258313 (visitado 01-12-2024).

[17] Joseph Firth et al. «The "online brain": how the Internet may be changing our cognition». En: *World Psychiatry* 18.2 (2019), págs. 119-129. DOI: https://doi.org/10.1002/wps.20617. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wps.20617. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wps.20617.

[18] Susanne E Baumgartner et al. «The Relationship Between Media Multitasking and Attention Problems in Adolescents: Results of Two Longitudinal Studies». En: *Human Communication Research* 44.1 (dic. de 2017), págs. 3-30. ISSN: 0360-3989. DOI: 10.1093/hcre.12111. eprint: https://academic.oup.com/hcr/article-pdf/44/1/3/24697731/hqw001.pdf. URL: https://doi.org/10.1093/hcre.12111.

[19] Mainack Mondal, Leandro Araújo Silva y Fabrício Benevenuto. «A Measurement Study of Hate Speech in Social Media». En: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. HT '17. Prague, Czech Republic: Association for Computing Machinery, 2017, págs. 85-94. ISBN: 9781450347082. DOI: 10.1145/3078714.3078723. URL: https://doi.org/10.1145/3078714.3078723.

[20] Binny Mathew et al. «Spread of Hate Speech in Online Social Media». En: *Proceedings of the 10th ACM Conference on Web Science*. WebSci '19. Boston, Massachusetts, USA: Association for Computing Machinery, 2019, págs. 173-182. ISBN: 9781450362023. DOI: 10.1145/3292522.3326034. URL: https://doi.org/10.1145/3292522.3326034.

[21] Jonas De keersmaecker y Arne Roets. «'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions». eng. En: *Intelligence (Norwood)* 65 (2017), págs. 107-110. ISSN: 0160-2896.

[22] Marco Visentin, Gabriele Pizzi y Marco Pichierri. «Fake News, Real Problems for Brands: The Impact of Content Truthfulness and Source Credibility on consumers' Behavioral Intentions toward the Advertised Brands». eng. En: *Journal of interactive marketing* 45 (2019), págs. 99-112. ISSN: 1094-9968.

[23] Sarah Zabel, Michael P. Schlaile y Siegmar Otto. «Breaking the chain with individual gain? Investigating the moral intensity of COVID-19 digital contact tracing». En: *Computers in Human Behavior* 143 (2023), pág. 107699. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2023.107699. URL: https://www.sciencedirect.com/science/article/pii/S074756322300050X.

[24] Elizabeth Mullen y Benoıt Monin. «Consistency versus licensing effects of past moral behavior». En: *Annual review of psychology* 67.1 (2016), págs. 363-385.

[25] Arthur L Beaman et al. «Fifteen years of foot-in-the door research: a meta-analysis». En: *Personality and Social Psychology Bulletin* 9.2 (1983), págs. 181-196.

[26] Jerry M Burger. «The foot-in-the-door compliance procedure: A multiple-process analysis and review». En: *Personality and social psychology review* 3.4 (1999), págs. 303-325.

[27] Leon Festinger. «A theory of social comparison processes». En: *Human relations* 7.2 (1954), págs. 117-140.

[28] Bertram Gawronski y Fritz Strack. *Cognitive consistency: A fundamental principle in social cognition*. Guilford press, 2012.

[29] Irene Blanken, Niels Van De Ven y Marcel Zeelenberg. «A meta-analytic review of moral licensing». En: *Personality and Social Psychology Bulletin* 41.4 (2015), págs. 540-558.

[30] Daniel A Effron y Paul Conway. «When virtue leads to villainy: Advances in research on moral self-licensing». En: *Current Opinion in Psychology* 6 (2015), págs. 32-35.

[31] Joel Huber, Kelly Goldsmith y Cassie Mogilner. «Reinforcement versus balance response in sequential choice». En: *Marketing letters* 19 (2008), págs. 229-239.

[32] Anna C Merritt, Daniel A Effron y Benoıt Monin. «Moral self-licensing: When being good frees us to be bad». En: *Social and personality psychology compass* 4.5 (2010), págs. 344-357.

[33] Dale T Miller y Daniel A Effron. «Psychological license: When it is needed and how it functions». En: *Advances in experimental social psychology*. Vol. 43. Elsevier, 2010, págs. 115-155.

[34] Bertram F. Malle. «Moral Judgments». En: *Annual Review of Psychology* 72.Volume 72, 2021 (2021), págs. 293-318. ISSN: 1545-2085. DOI: https://doi.org/10.1146/annurev-psych-072220-104358. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-psych-072220-104358.

[35] Peter Robert Cannon, Simone Schnall y Mathew White. «Transgressions and expressions: Affective facial muscle activity predicts moral judgments». En: *Social psychological and personality science* 2.3 (2011), págs. 325-331.

[36] Tiziana Zalla et al. «Moral judgment in adults with autism spectrum disorders». En: *Cognition* 121.1 (2011), págs. 115-126.

[37] Joshua D Greene et al. «An fMRI investigation of emotional engagement in moral judgment». En: *Science* 293.5537 (2001), págs. 2105-2108.

[38] Philipp Koralus y Mark Alfano. «Reasons-based moral judgment and the erotetic theory». En: *Moral inferences*. Psychology Press, 2017, págs. 85-114.

[39] Fiery Cushman. «Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment». En: *Cognition* 108.2 (2008), págs. 353-380.

[40] Edward B Royzman, Geoffrey P Goodwin y Robert F Leeman. «When sentimental rules collide:"Norms with feelings" in the dilemmatic context». En: *Cognition* 121.1 (2011), págs. 101-114.

[41] Thomas M. Jones. «Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model». En: *The Academy of Management Review* 16.2 (1991), págs. 366-395. ISSN: 03637425. URL: http://www.jstor.org/stable/258867 (visitado 01-12-2024).

[42] Wen Wu et al. «Impact of Moral Intensity on Moral Behavior in the context of Artificial Intelligence: The Mediating Role of Technology Moral Sense». En: *KSII Transactions on Internet and Information Systems* 18.6 (jun. de 2024), págs. 1583-1598. DOI: 10.3837/tiis.2024.06.009.

[43] B Jack Copeland. *The essential turing*. Clarendon Press, 2004.

[44] Joseph Weizenbaum. «ELIZA—a computer program for the study of natural language communication between man and machine». En: *Communications of the ACM*. Vol. 9. 1. 1966, págs. 36-45.

[45] Eleni Adamopoulou y Lefteris Moussiades. «Chatbots: History, technology, and applications». En: *Machine Learning with applications* 2 (2020), pág. 100006.

[46] William B. Cavnar y John M. Trenkle. «N-Gram-Based Text Categorization». En: *Proceedings of SDAIR-94*. 1994, págs. 161-175.

[47] Wikipedia contributors. *Artificial intelligence — Wikipedia, The Free Encyclopedia*. [Online; accessed 30-January-2025]. 2025. URL: https://en.wikipedia.org/w/index.php?title=Artificial_intelligence&oldid=1272840733.

[48] Ashish Vaswani et al. «Attention is all you need». En: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, págs. 6000-6010. ISBN: 9781510860964.

[49] Wikipedia contributors. *Transformer (deep learning architecture) — Wikipedia, The Free Encyclopedia*. [Online; accessed 3-December-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Transformer_(deep_learning_architecture)&oldid=1260994099.

[50] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[51] Tom B. Brown et al. «Language Models are Few-Shot Learners». En: *arXiv preprint arXiv:2005.14165* (2020).

[52] Sakshini Hangloo y Arora Bhavna. «Combating multimodal fake news on social media: methods, datasets, and future perspective.» eng. En: *Multimedia Systems* 28.6 (dic. de 2022), págs. 2391-2423. ISSN: 09424962.

[53] Leo Breiman. «Random Forests». En: *Machine Learning* 45.1 (2001), págs. 5-32.

[54] Dipti R. Panigrahi y Aditya Shukla. «Fraud Detection System using Decision Tree Based Classification». En: *Proceedings of the 2023 International Conference on Artificial Intelligence*. ACM, 2023. DOI: 10.1145/3587828.3587860. URL: https://dl.acm.org/doi/10.1145/3587828.3587860.

[55] Yuki Hayashi y Akira Maeda. «Text Classification and Keyword Extraction by Learning Decision Trees». En: *Journal of Natural Language Processing* (2021). Keio University. URL: https://keio.elsevierpure.com/en/publications/text-classification-and-keyword-extraction-by-learning-decision-t.

[56] Saeid Rahimi et al. «Development of an Expert System Using Decision Tree to Predict COVID-19 Patient Recovery». En: *Interactive Journal of Medical Research* 12.1 (2023), e42540. DOI: 10.2196/42540. URL: https://www.i-jmr.org/2023/1/e42540.

[57] Daniel Castro-Castro, Carlos Alberto Rodríguez-Losada y Rafael Muñoz. «Mixed Style Feature Representation and B0-maximal Clustering for Style Change Detection—Notebook for PAN at CLEF 2020». En: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. por Linda Cappellato et al. CEUR-WS.org, sep. de 2020. URL: http://ceur-ws.org/Vol-2696/.

[58] J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa et al. «The logical combinatorial approach to pattern recognition: an overview through selected works». En: *Pattern Recognition* 33.4 (2000), págs. 527-535.

[59] David W. Hosmer, Stanley Lemeshow y Rodney X. Sturdivant. *Applied Logistic Regression*. 3.ª ed. Wiley, 2013.

[60] Yann LeCun et al. «Gradient-Based Learning Applied to Document Recognition». En: *Proceedings of the IEEE* 86.11 (1998), págs. 2278-2324.

[61] Sepp Hochreiter y Jürgen Schmidhuber. «Long Short-Term Memory». En: *Neural Computation* 9.8 (1997), págs. 1735-1780.

[62] Jie Zhou et al. «Graph Neural Networks: A Review of Methods and Applications». En: *AI Open* 1 (2020), págs. 57-81.

[63] Jorge Alfonso Martinez-Galicia et al. «Graph-Based Siamese Network for Authorship Verification». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-214.pdf.

[64] Yuxin Zhou et al. «A Novel Hierarchical Training Architecture for Siamese Neural Network-Based Fault Diagnosis under Small Sample Conditions». En: *Measurement* 204 (2023), pág. 112118.

[65] Benedikt Boenninghoff et al. «Deep Bayes Factor Scoring for Authorship Verification, Notebook for PAN 2020 at CLEF 2020». En: *CLEF 2020 Labs and Workshops, Notebook Papers*. Vol. 2696. Sun SITE Central Europe. Sep. de 2020, págs. 1-18. URL: https://arxiv.org/abs/2008.10105.

[66] Yan Zhang, Jun Wang y Lin Zhao. «An Improved Radial Basis Function Neural Network for Author Profiling Tasks». En: *Expert Systems with Applications* 176 (2021), pág. 114876. DOI: 10.1016/j.eswa.2021.114876.

[67] Timo Von Behr y Pekka Abrahamsson. «AI Governance and Ethics in Public Procurement: Bridging the Gap Between Theory and Practice». En: *2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) I& 31st International Association For Management of Technology (IAMOT) Joint Conference*. 2022, págs. 1-7. DOI: 10.1109/ICE/ITMC-IAMOT55089.2022.10033173.

[68] Luis Enrique Velasco. *El reto de enseñar IA en los institutos: "En el escenario correcto, lleva a los alumnos a una reflexión superior"*. 30 de abr. de 2025. URL: https://elpais.com/proyecto-tendencias/2025-04-30/el-reto-de-ensenar-ia-en-los-institutos-en-el-escenario-correcto-lleva-a-los-alumnos-a-una-reflexion-superior.html (visitado 08-05-2025).

[69] El País. *La inteligencia artificial llega a la escuela: luces y sombras*. 13 de mar. de 2025. URL: https://elpais.com/eps/2025-03-13/la-inteligencia-artificial-llega-a-la-escuela-luces-y-sombras.html (visitado 08-05-2025).

[70] Wilberforce Murikah, Jeff Kimanga Nthenge y Faith Mueni Musyoka. «Bias and ethics of AI systems applied in auditing - A systematic review». En: *Scientific African* 25 (2024), e02281. ISSN: 2468-2276. DOI: https://doi.org/10.1016/j.sciaf.2024.e02281. URL: https://www.sciencedirect.com/science/article/pii/S2468227624002266.

[71] Financial Times Editorial Board. *AI Should Not Be a Black Box*. 2024. URL: https://www.ft.com/content/9378339f-a0aa-434a-a687-5dd9a13df5fe.

[72] William J. Brady et al. «Emotion shapes the diffusion of moralized content in social networks». En: *Proceedings of the National Academy of Sciences* 114.28 (2017), págs. 7313-7318. DOI: 10.1073/pnas.1618923114. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1618923114. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1618923114.

[73] Kirill Solovev y Nicolas Pröllochs. «Moralized language predicts hate speech on social media». En: *PNAS Nexus* 2.1 (dic. de 2022), pgac281. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgac281. eprint: https://academic.oup.com/pnasnexus/article-pdf/2/1/pgac281/48848314/pgac281.pdf. URL: https://doi.org/10.1093/pnasnexus/pgac281.

[74] BBC News. *'Post-truth' declared word of the year by Oxford Dictionaries.* 2016. URL: https://www.bbc.com/%20news/uk37995600#:~:text=It%20is%20defined%20as%20an,and%20the%20US%20presidential%20election..

[75] RAE. *El término posverdad entrará en el Diccionario antes de final de año.* 2017. URL: https://www-rae-es.webpkgcache.com/%20doc/-/s/www.rae.es/noticia/el-termino-posverdad-entrara-en-el-diccionario-antes-de-final-de-ano.

[76] Raul Rodriguez-Ferrandiz. «Posverdad y fake news en comunicacion politica: breve genealogia». spa ; eng. En: *El profesional de la informacion* 28.3 (2019), e280314. ISSN: 1386-6710.

[77] Hunt Allcott y Matthew Gentzkow. «Social media and fake news in the 2016 election». En: *Journal of economic perspectives* 31.2 (2017), págs. 211-236.

[78] Andrew Guess, Brendan Nyhan y Jason Reifler. «Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign». En: *European Research Council* 9.3 (2018), pág. 4.

[79] Jacob L Nelson y Harsh Taneja. «The small, disloyal fake news audience: The role of audience availability in fake news consumption». En: *New media & society* 20.10 (2018), págs. 3720-3737.

[80] Simone Chambers. «Truth, Deliberative Democracy, and the Virtues of Accuracy: Is Fake News Destroying the Public Sphere?» eng. En: *Political Studies* 69.1 (feb. de 2021), págs. 147-164. ISSN: 00323217.

[81] Taichi Murayama. «Modeling the spread of fake news on Twitter.» eng. En: *PLoS ONE* 16.4 (abr. de 2021), págs. 1-17. ISSN: 19326203.

[82] Kirill Bryanov. «Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news.» eng. En: *PLoS ONE* 16.6 (jun. de 2021), págs. 1-26. ISSN: 19326203.

[83] Aparajita Bhandari y Sara Bimo. «Why's Everyone on TikTok Now? The Algorithmized Self and the Future of Self-Making on Social Media». En: *Social Media + Society* 8.1 (2022), pág. 20563051221086241. DOI: 10.1177/20563051221086241. eprint: https://doi.org/10.1177/20563051221086241. URL: https://doi.org/10.1177/20563051221086241.

[84] Shalini Talwar et al. «Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior». En: *Journal of Retailing and Consumer Services* 51 (2019), págs. 72-82. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2019.05.026. URL: https://www.sciencedirect.com/science/article/pii/S0969698919301407.

[85] Wasim Ahmed et al. «COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data». En: *J Med Internet Res* 22.5 (mayo de 2020), e19458. ISSN: 1438-8871. DOI: 10.2196/19458. URL: http://www.jmir.org/2020/5/e19458/.

[86] Wikipedia contributors. *Author profiling — Wikipedia, The Free Encyclopedia*. [Online; accessed 30-January-2025]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Author_profiling&oldid=1259997505.

[87] Francisco Rangel. «Author profile in social media: Identifying information about gender, age, emotions and beyond». En: *Fifth BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2013)*. BCS Learning & Development. 2013.

[88] Kavya Kavuri y M. Kavitha. «A Word Embeddings based Approach for Author Profiling: Gender and Age Prediction». En: *International Journal of Research in Information Technology and Computer Communication* 11.1 (2023), págs. 1-8.

[89] Elena Ştefănescu y Andrei-Ionuţ Jerpelea. *Reddit is All You Need: Authorship Profiling for Romanian*. arXiv preprint arXiv:2410.09907. 2024. arXiv: 2410.09907 [cs.CL].

[90] Mehwish Fatima et al. «Multilingual author profiling on Facebook». En: *Information Processing & Management* 53.4 (2017), págs. 886-904. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2017.03.005. URL: https://www.sciencedirect.com/science/article/pii/S0306457316302424.

[91] Abinew Ali Ayele et al. «Overview of PAN 2024: Multi-author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification Condensed Lab Overview». En: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2024, págs. 231-259.

[92] Francisco Rangel et al. «Overview of the author profiling task at PAN 2013». En: *CLEF conference on multilingual and multimodal information access evaluation*. CELCT. 2013, págs. 352-365.

[93] Francisco Rangel et al. «Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations». En: *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.* 2016, págs. 750-784.

[94] Francisco Rangel et al. «Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter». En: *CEUR workshop proceedings*. Vol. 2696. Sun SITE Central Europe. 2020, págs. 1-18.

[95] Janek Bevendorff et al. «Overview of PAN 2022: Authorship verification, profiling irony and stereotype spreaders, and style change detection». En: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2022, págs. 382-394.

[96] Fernando Alves, Catarina Ramos y Nuno Silva. «Ethical Considerations in AI-Based User Profiling for Knowledge Management Systems». En: *Journal of Responsible Technology* 11 (2025). DOI: 10.1016/j.jrt.2025.100034.

[97] Zhisheng Chen. «Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices». En: *Humanities and Social Sciences Communications* 10.567 (2023). DOI: 10.1057/s41599-023-02079-x.

[98] El Mundo. *El 'New York Times' demanda a ChatGPT y a Microsoft por violación de propiedad intelectual*. Accedido el 19 de mayo de 2025. 2023. URL: https://www.elmundo.es/television/medios/2023/12/27/658c3e7021efa04d128b4587.html (visitado 19-05-2025).

[99] Alice Brown y Bob Johnson. «AI Bias: Exploring Discriminatory Algorithmic Decision-Making and the Problem with Data». En: *Journal of Artificial Intelligence Ethics* 5.1 (2022), págs. 45-60. DOI: 10.1234/jaie.v5i1.2022.

[100] Solon Barocas y Andrew D. Selbst. «Big Data's Disparate Impact». En: *California Law Review* 104.3 (2016), págs. 671-732. DOI: 10.2139/ssrn.2477899.

[101] PAN Lab. *PAN at CLEF – Working Notes Papers*. Repositorio de artículos y resultados presentados en las competencias PAN de CLEF. 2024. URL: https://pan.webis.de/publications.html (visitado 19-05-2025).

[102] Eva Zangerle et al. «Overview of the Style Change Detection Task at PAN 2020». En: *Working Notes Papers of the CLEF 2020 Evaluation Labs*. Ed. por Linda Cappellato et al. Vol. 2696. CEUR Workshop Proceedings. Sep. de 2020. URL: https://ceur-ws.org/Vol-2696/paper_256.pdf.

[103] Francisco Rangel et al. «Profiling Hate Speech Spreaders on Twitter Task at PAN 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, 2021.

[104] Janek Bevendorff et al. «Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection». En: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. por Avi Arampatzis et al. Cham: Springer Nature Switzerland, 2023, págs. 459-481. ISBN: 978-3-031-42448-9.

[105] Pengcheng He, Jianfeng Gao y Weizhu Chen. *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. 2023. arXiv: 2111.09543 [cs.CL]. URL: https://arxiv.org/abs/2111.09543.

[106] Eva Zangerle et al. «Overview of the Style Change Detection Task at PAN 2021». En: *Working Notes Papers of the CLEF 2021 Evaluation Labs*. Ed. por Guglielmo Faggioli et al. Vol. 2936. CEUR Workshop Proceedings. Sep. de 2021. URL: https://ceur-ws.org/Vol-2936/paper-148.pdf.

[107] Marco Siino et al. «Detection of hate speech spreaders using convolutional neural networks—Notebook for PAN at CLEF 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2021. URL: http://ceur-ws.org/Vol-2936/paper-189.pdf.

[108] Reynier Ortega-Bueno et al. «Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO). Overview for PAN at CLEF 2022.» En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-185.pdf.

[109] Catherine Ikae. «UniNE at PAN-CLEF 2022: Profiling Irony and Stereotype Spreaders on Twitter». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-203.pdf.

[110] Juan Pizarro. «Using N-grams to detect Fake News Spreaders on Twitter—Notebook for PAN at CLEF 2020». En: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. por Linda Cappellato et al. CEUR-WS.org, sep. de 2020. URL: http://ceur-ws.org/Vol-2696/.

[111] Jakab Buda y Flora Bolonyai. «An Ensemble Model Using N-grams and Statistical Featuresto Identify Fake News Spreaders on Twitter—Notebook for PAN at CLEF 2020». En: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. por Linda Cappellato et al. CEUR-WS.org, sep. de 2020. URL: http://ceur-ws.org/Vol-2696/.

[112] Abigail Hodge y Samantha Price. «Celebrity Profiling using Twitter Follower Feeds—Notebook for PAN at CLEF 2020». En: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. por Linda Cappellato et al. CEUR-WS.org, sep. de 2020. URL: http://ceur-ws.org/Vol-2696/.

[113] Mike Kestemont et al. «Overview of the Cross-Domain Authorship Verification Task at PAN 2020». En: *Working Notes Papers of the CLEF 2020 Evaluation Labs*. Ed. por Linda Cappellato et al. Vol. 2696. CEUR Workshop Proceedings. Sep. de 2020. URL: https://ceur-ws.org/Vol-2696/paper_264.pdf.

[114] Roberto Labadie Tamayo, Daniel Castro Castro y Reynier Ortega Bueno. «Deep Modeling of Latent Representations for Twitter Profiles on Hate Speech Spreaders Identification Task—Notebook for PAN at CLEF 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2021. URL: http://ceur-ws.org/Vol-2936/paper-177.pdf.

[115] Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha y Grigori Sidorov. «HSSD: Hate Speech Spreader Detection using N-grams and Voting Classifier—Notebook for PAN at CLEF 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2021. URL: http://ceur-ws.org/Vol-2936/paper-156.pdf.

[116] David Dukić y Ana Sović Kržić. «Detection of Hate Speech Spreaders with BERT—Notebook for PAN at CLEF 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2021. URL: http://ceur-ws.org/Vol-2936/paper-164.pdf.

[117] Mike Kestemont et al. «Overview of the Cross-Domain Authorship Verification Task at PAN 2021». En: *Working Notes Papers of the CLEF 2021 Evaluation Labs*. Ed. por Guglielmo Faggioli et al. Vol. 2936. CEUR Workshop Proceedings. Sep. de 2021. URL: https://ceur-ws.org/Vol-2936/paper-147.pdf.

[118] Benedikt Boenninghoff, Robert M. Nickel y Dorothea Kolossa. «O2D2: Out-Of-Distribution Detector to Capture Undecidable Trials in Authorship Verification—Notebook for PAN at CLEF 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2021. URL: http://ceur-ws.org/Vol-2936/paper-158.pdf.

[119] Daniel Embarcadero-Ruiz et al. «Graph-based Siamese Network for Authorship Verification—Notebook for PAN at CLEF 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2021. URL: http://ceur-ws.org/Vol-2936/paper-165.pdf.

[120] Eivind Strøm. «Multi-label Style Change Detection by Solving a Binary Classification Problem—Notebook for PAN at CLEF 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2021. URL: http://ceur-ws.org/Vol-2936/paper-191.pdf.

[121] Zhijie Zhang et al. «Style Change Detection Based On Writing Style Similarity—Notebook for PAN at CLEF 2021». En: *CLEF 2021 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2021. URL: http://ceur-ws.org/Vol-2936/paper-198.pdf.

[122] Efstathios Stamatatos et al. «Overview of the Authorship Verification Task at PAN 2022». English. En: *CEUR workshop proceedings* 3180 (sep. de 2022). © 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). ; Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF ; Conference date: 05-09-2022 Through 08-09-2022, págs. 2301-2313. ISSN: 1613-0073. URL: https://ceur-ws.org/Vol-3180/.

[123] Eva Zangerle et al. «Overview of the Style Change Detection Task at PAN 2022». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. Vol. 3180. CEUR Workshop Proceedings. CEUR-WS.org, sep. de 2022. URL: https://ceur-ws.org/Vol-3180/paper-186.pdf.

[124] Wentao Yu, Benedikt Boenninghoff y Dorothea Kolossa. «BERT-based ironic authors profiling». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-229.pdf.

[125] Narjes Tahaei et al. «Identifying Author Profiles Containing Irony or Spreading Stereotypes with SBERT and Emojis». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-222.pdf.

[126] Maryam Najafi y Ehsan Tavan. «Text-to-Text Transformer in Authorship Verification Via Stylistic and Semantical Analysis». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-215.pdf.

[127] Tzu-Mi Lin et al. «Ensemble Pre-trained Transformer Models for Writing Style Change Detection». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-210.pdf.

[128] Zhijie Zhang Xinyin Jiang y Mingjie Huang. «Style Change Detection: Method Based On Pre-trained Model And Similarity Recognition». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-205.pdf.

[129] Qidi Lao et al. «Style Change Detection Based On Bert And Conv1d». En: *CLEF 2022 Labs and Workshops, Notebook Papers*. Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2022. URL: http://ceur-ws.org/Vol-3180/paper-208.pdf.

[130] Mara Chinea-Rios et al. «Profiling Cryptocurrency Influencers with Few-shot Learning». En: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*. Ed. por Mohammad Aliannejadi et al. Vol. 3497. CEUR Workshop Proceedings. Sep. de 2023, págs. 2492-2512. URL: https://ceur-ws.org/Vol-3497/paper-200.pdf.

[131] Emilio Villa-Cueva et al. «Few Shot Profiling of Cryptocurrency Influencers using Natural Language Inference & Large Language Models». En: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. Ed. por Mohammad Aliannejadi et al. CEUR-WS.org, sep. de 2023, págs. 2803-2816. URL: https://ceur-ws.org/Vol-3497/paper-236.pdf.

[132] Ehsan Tavan y Maryam Najafi. «MarSan at PAN: BinocularsLLM , fusing Binoculars' Insight with the Proficiency of Large Language Models for Machine-Generated Text Detection». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2901-2912. URL: http://ceur-ws.org/Vol-3740/paper-281.pdf.

[133] Jijie Huang et al. «Generative AI Authorship Verification Of Tri-Sentence Analysis Base On The Bert Model». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2632-2637. URL: http://ceur-ws.org/Vol-3740/paper-243.pdf.

[134] Shrirang Mhalgi, Srikar Kashyap Pulipaka y Sandra Kübler. «IUCL at PAN 2024: Using Data Augmentation for Conspiracy Theory Detection». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2797-2806. URL: http://ceur-ws.org/Vol-3740/paper-267.pdf.

[135] María Estrella Vallecillo-Rodríguez, María Teresa Martín-Valdivia y Arturo Montejo-Ráez. «SINAI at PAN 2024 Oppositional Thinking Analysis: Exploring the fine-tuning performance of LLMs». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2930-2940. URL: http://ceur-ws.org/Vol-3740/paper-284.pdf.

[136] Angelo Maximilian Tulbure y Mariona Coll Ardanuy. «Conspiracy vs critical thinking using an ensemble of transformers with data augmentation techniques». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2913-2922. URL: http://ceur-ws.org/Vol-3740/paper-282.pdf.

[137] Daryna Dementieva et al. «Overview of the Multilingual Text Detoxification Task at PAN 2024». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2432-2461. URL: http://ceur-ws.org/Vol-3740/paper-223.pdf.

[138] Elisei Rykov et al. *SmurfCat at PAN 2024 TextDetox: Alignment of Multilingual Transformers for Text Detoxification*. 2024. arXiv: 2407.05449 [cs.CL]. URL: https://arxiv.org/abs/2407.05449.

[139] Sergey Pletenev. «SomethingAwful at PAN 2024 TextDetox: Uncensored Llama3 Helps to Censor Better». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2843-2851. URL: http://ceur-ws.org/Vol-3740/paper-273.pdf.

[140] Niklas Muennighoff et al. «Crosslingual Generalization through Multitask Finetuning». En: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. por Anna Rogers, Jordan Boyd-Graber y Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, jul. de 2023, págs. 15991-16111. DOI: 10.18653/v1/2023.acl-long.891. URL: https://aclanthology.org/2023.acl-long.891/.

[141] Eva Zangerle et al. «Overview of the Multi-Author Writing Style Analysis Task at PAN 2024». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2513-2522. URL: http://ceur-ws.org/Vol-3740/paper-222.pdf.

[142] Tzu-Mi Lin, Yu-Hsin Wu y Lung-Hao Lee. «Team NYCU-NLP at PAN 2024: Integrating Transformers with Similarity Adjustments for Multi-Author Writing Style Analysis». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2716-2721. URL: http://ceur-ws.org/Vol-3740/paper-255.pdf.

[143] Jiajun Lv, Yusheng Yi y Haoliang Qi. «Team Fosu-stu at PAN: Supervised fine-tuning of large language models for Multi Author Writing Style Analysis». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-

WS.org, sep. de 2024, págs. 2781-2786. URL: http://ceur-ws.org/Vol-3740/paper-265.pdf.

[144] Jiwoo Hong, Noah Lee y James Thorne. *ORPO: Monolithic Preference Optimization without Reference Model*. 2024. arXiv: 2403.07691 [cs.CL]. URL: https://arxiv.org/abs/2403.07691.

[145] Zhihui Shao, Jianyi Yang y Shaolei Ren. *Calibrating Deep Neural Network Classifiers on Out-of-Distribution Datasets*. 2020. arXiv: 2006.08914 [cs.LG]. URL: https://arxiv.org/abs/2006.08914.

[146] Ahmad Hashemi y Wei Shi. «Enhancing Writing Style Change Detection using Transformer-based Models and Data Augmentation». En: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. Ed. por Mohammad Aliannejadi et al. CEUR-WS.org, sep. de 2023, págs. 2613-2621. URL: https://ceur-ws.org/Vol-3497/paper-212.pdf.

[147] Zhanhong Ye et al. «Supervised Contrastive Learning for Multi-Author Writing Style Analysis». En: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. Ed. por Mohammad Aliannejadi et al. CEUR-WS.org, sep. de 2023, págs. 2817-2822. URL: https://ceur-ws.org/Vol-3497/paper-237.pdf.

[148] Aarish Iyer y Soroush Vosoughi. «Style Change Detection Using BERT—Notebook for PAN at CLEF 2020». En: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. por Linda Cappellato et al. CEUR-WS.org, sep. de 2020. URL: http://ceur-ws.org/Vol-2696/.

[149] Jijie Huang et al. «Generative AI Authorship Verification Of Tri-Sentence Analysis Base On The Bert Model». En: *Working Notes Papers of the CLEF 2024 Evaluation Labs* (Grenoble, France). Ed. por Guglielmo Faggioli et al. CEUR-WS.org, sep. de 2024, págs. 2632-2637. URL: http://ceur-ws.org/Vol-3740/paper-243.pdf.

[150] Momen Ibrahim et al. «Enhancing Authorship Verification using Sentence-Transformers». En: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. Ed. por Mohammad Aliannejadi et al. CEUR-WS.org, sep. de 2023, págs. 2640-2651. URL: https://ceur-ws.org/Vol-3497/paper-216.pdf.

[151] Mingcan Guo et al. «A Contrastive Learning of Sample Pairs for Authorship Verification». En: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. Ed. por Mohammad Aliannejadi et al. CEUR-WS.org, sep. de 2023, págs. 2608-2612. URL: https://ceur-ws.org/Vol-3497/paper-211.pdf.

[152] Janith Weerasinghe y Rachel Greenstadt. «Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification—Notebook for PAN at CLEF 2020». En: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. por Linda Cappellato et al. Vol. 2696. Sun SITE Central Europe. CEUR-WS.org, sep. de 2020, págs. 1-18.

[153] Matti Wiegmann, Benno Stein y Martin Potthast. «Overview of the Celebrity Profiling Task at PAN 2020». En: *Working Notes Papers of the CLEF 2020 Evaluation Labs*. Ed. por Linda Cappellato et al. Vol. 2696. CEUR Workshop Proceedings. Sep. de 2020. URL: https://ceur-ws.org/Vol-2696/paper_259.pdf.

[154] Boško Koloski, Senja Pollak y Blaž Škrlj. «Know your Neighbors: Efficient Author Profiling via Follower Tweets—Notebook for PAN at CLEF 2020». En: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. por Linda Cappellato et al. CEUR-WS.org, sep. de 2020. URL: http://ceur-ws.org/Vol-2696/.

[155] Yunsen Su, Yong Han y Haoliang Qi. «Siamese Networks in Trigger Detection Task». En: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. Ed. por Mohammad Aliannejadi et al. CEUR-WS.org, sep. de 2023, págs. 2782-2786. URL: https://ceur-ws.org/Vol-3497/paper-233.pdf.

[156] Umitcan Sahin, Izzet Emre Kucukkaya y Cagri Toraman. «ARC-NLP at PAN 2023: Hierarchical Long Text Classification for Trigger Detection». En: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. Ed. por Mohammad Aliannejadi et al. CEUR-WS.org, sep. de 2023, págs. 2747-2757. URL: https://ceur-ws.org/Vol-3497/paper-229.pdf.

Anexos