**INTRODUCCIÓN**

En las últimas décadas, el avance de las tecnologías de la información y la comunicación ha transformado profundamente la interacción humana, dando lugar a un nuevo entorno social en línea. Este espacio digital, caracterizado por su accesibilidad y alcance global, ha permitido la creación de nuevas formas de comunicación y expresión. Sin embargo, también ha propiciado el surgimiento de comportamientos transgresivos, un fenómeno que ha captado la atención de investigadores, educadores y legisladores debido a sus implicaciones sociales y psicológicas (Pavesi, 2020).

El vocablo transgresión proviene del latín *transgressio* y significa “acción y efecto de hacer daño a alguien”. El verbo transgredir alude a “no respetar, saltarse, pasar más allá de normas o costumbres” (Anders, 2025). El comportamiento transgresivo se define como la violación de normas sociales establecidas, viola o desafía los estándares de comportamiento, creencia o moralidad. A menudo implica una forma de violencia verbal o física, acoso, actos sexuales no deseados, intimidación, racismo, misoginia o discriminación. Puede presentarse en diferentes ámbitos de nuestra vida, desde lo social y ético hasta lo legal, por individuos o grupos a través de actos deliberados o inconscientes (Carranza Coello and Carpio Ramírez, 2019).

El comportamiento transgresor en línea ha aumentado con rapidez. En el contexto digital, este tipo de conductas incluye desde el ciberacoso y el discurso de odio hasta la difusión de información y actividades ilícitas. La preferencia por el acoso en línea, en contraposición al acoso presencial se da debido al distanciamiento físico y emocional, sensación de anonimato, falta de regulación legal, sin supervisión adulta para menores de edad, en virtud de este efecto de desinhibición en línea los niños y adolescentes son particularmente propensos a involucrarse en conductas transgresoras en línea tanto como perpetradores o como víctimas (Devarakonda et al., 2020), de hecho en 2019 se realizó una encuesta periódica de población por el Ministerio de Justicia, Seguridad y Estadística de los Países Bajos, donde tres de cada diez jóvenes declararon haber participado en lo que consideraban un comportamiento transgresor en línea (Statistiek, 2020).

La urgencia de abordar los comportamientos transgresores en las redes sociales no puede ser subestimada. La Asamblea del Consejo de Derechos Humanos de la ONU (“OHCHR,” 2025), la Plataforma EDVAW (“EDVAW Platform,” 2022) y Amnistía Internacional (“Amnistía Internacional,” 2024) han destacado la recurrencia de la violencia y la discriminación en línea, exacerbando la desigualdad de género, el racismo y otras formas de agresión.

A partir de 2019, el acceso a internet en Cuba se masificó, alcanzando a 7 700 000 de usuarios en línea en 2021, de ellos 6 000 280 utilizaban las redes sociales (Camacho Costales, 2021). El marco legal (“Código de Familia,” 2022; “Código Penal,” 1987; “Cuba 2019 Constitución,” 2019; “Ley de Comunicación,” 2023) establece la igualdad y reconoce y prohíbe la transgresión en línea; las redes son fenómenos novedosos, por lo que la invisibilidad de los comportamientos transgresores en línea es ordinaria. Incluso entre entidades legales, educativas y de salud, faltan protocolos, estándares y estrategias institucionales para su atención sistemática. Aunque existen estructuras internas, regulaciones universitarias y leyes nacionales que intentan abordar el problema, las universidades cubanas como instituciones educativas no cuentan con una estrategia dirigida a gestionar comportamientos transgresores. Su detección, clasificación, atención legal y psicológica carecen de mecanismos, legislación y protocolos que contribuyan a su prevención y atención desde una perspectiva interseccional.

En este ámbito universitario, las conductas transgresoras se reproducen en espacios físicos y digitales debido a sesgos interseccionales (“UNDP,” 2025), que afectan a los jóvenes. Las universidades cubanas ofrecen mejor infraestructura tecnológica y acceso a espacios digitales, lo que aumenta los riesgos de recibir o ejercer formas de violencia o discriminación en línea. La Universidad Central de Marta Abreu de Las Villas (UCLV), bajo la jurisdicción del Ministerio de Educación Superior de Cuba, es la universidad más multidisciplinaria del país. Es la universidad con el mayor acceso nacional a internet y cubre geográficamente a estudiantes de cinco provincias, un factor que contribuye a su diversidad demográfica, de género y étnica. Los grupos vulnerables, principalmente motivados por género y raza, cohabitan en el contexto universitario y son, por lo tanto, susceptibles a ataques en espacios digitales. La presencia de estos comportamientos entre los estudiantes impacta los procesos de interacción social dentro de la institución (UCLV, 2025).

La UCLV cuenta con la Comisión Universitaria de Prevención, el Centro Universitario de Bienestar (CBU), la Clínica Legal y la Cátedra de Género y Derechos Humanos, que apoyan el trabajo de precaución y atención a los comportamientos transgresores en línea. La Comisión Universitaria de Prevención es responsable del desarrollo e implementación de estrategias de prevención; el CBU proporciona apoyo y asesoramiento a las personas afectadas y ofrece atención psicológica gratuita; la Clínica Legal ofrece asistencia legal; y la Cátedra de Género y Desarrollo Humano se centra en temas relacionados con el género. Sin embargo, carecen de los protocolos, políticas y estrategias que lo hagan posible, siendo la articulación entre ellos un elemento que mejora la gestión universitaria para abordar este problema (“Tackling Online Transgressive Behaviours- Strengthening Institutional Policies on Digital Safety in Cuban University Community,” 2024).

En este contexto, se presenta la necesidad de ofrecer un estudio de los factores que contribuyen al comportamiento transgresivo en línea y sus implicaciones para los usuarios en las plataformas digitales mediante la recopilación y análisis de datos de interacciones en redes sociales como Twitter, Instagram, Facebook, entre otras. La carencia de un sistema que detecte dicha conducta en nuestra comunidad universitaria es un atraso, un factor crítico que agrava el problema. Sin un marco institucional que identifique y gestione estos actos, se corre el riesgo de que la violencia y el acoso en línea se normalicen, creando un ambiente hostil que afecta a la salud mental y emocional de los estudiantes. Esto no solo perjudica su bienestar individual, sino que también impacta negativamente en su rendimiento académico y en su participación en actividades sociales y educativas.

Tomando en cuenta lo analizado, se formulan las siguientes preguntas de investigación:

1. ¿Cómo detectar comportamientos transgresivos en línea entre los usuarios de redes sociales vinculados a la UCLV?
2. ¿Qué modelo de procesamiento del lenguaje natural es el más adecuado para detectar automáticamente comportamientos transgresivos en textos en español publicados por usuarios de redes sociales vinculados a la UCLV?
3. ¿Cuáles son las estrategias más efectivas para realizar minería de datos en las redes sociales utilizadas por la comunidad de la UCLV?

Objetivo general:

Desarrollar una herramienta computacional que permita la detección de comportamientos transgresivos en línea utilizando técnicas de procesamiento de lenguaje natural y minería de opiniones.

Objetivos específicos:

1. Identificar frases y sinónimos que indiquen algún tipo de comportamiento transgresivo en línea.
2. Utilizar técnicas de minería de opinión para la polarización de los textos analizados.
3. Extraer agentes o frases de comportamiento transgresivo a partir de los reportes estadísticos que se generan en la herramienta computacional diseñada.
4. Evaluar la herramienta propuesta mediante la aplicación de pruebas unitarias a un caso de estudio real.

El presente trabajo se encuentra estructurado en introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas y los anexos.

En el capítulo 1 se exponen los principales aspectos teóricos relacionados con el procesamiento del lenguaje natural y sus diferentes técnicas. Se analizan y caracterizan los diferentes tipos de comportamientos transgresivos en línea. Sentará las bases de los posibles modelos de inteligencia artificial a utilizar y herramientas de software disponibles en este contexto, y con el propósito de determinar las potencialidades de cada una es necesario un análisis comparativo de las mismas.

En el capítulo 2 se aborda…

En el capítulo 3 se presenta ….

**CAPÍTULO 1. ASPECTOS TEÓRICOS SOBRE TÉCNICAS DE PROCESAMIENTO DEL LENGUAJE NATURAL**

Este capítulo constituye un marco de referencia esencial para comprender los fundamentos teóricos y metodológicos del procesamiento del lenguaje natural, con especial énfasis en su aplicación al análisis de comportamientos transgresivos en entornos digitales, incorpora un análisis de los modelos de lenguaje masivo, evaluando su adaptabilidad a las peculiaridades del idioma español y se explica la necesidad de trabajar el idioma español ante que el inglés. Se exponen los diferentes tipos de conductas transgresoras en línea y se realiza un estudio detallado de las diferentes herramientas computacionales para PLN.

* 1. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN), conocido internacionalmente como Natural Language Processing (NLP) (Ramanathan, 2025), representa una de las disciplinas más innovadoras dentro del campo de la inteligencia artificial (IA). Esta tecnología busca cerrar la brecha de comunicación entre humanos y máquinas, dotando a los sistemas computacionales de la capacidad de comprender, interpretar y generar lenguaje humano. El PLN permite a las computadoras procesar y analizar grandes volúmenes de datos en lenguaje natural, extrayendo significado, contexto e incluso intencionalidad. Esta capacidad va mucho más allá del simple reconocimiento de palabras, abarcando aspectos complejos como la semántica, la pragmática y los matices culturales del lenguaje. Entre sus funciones principales destacan la comprensión del lenguaje (capacidad para interpretar textos y discursos), la generación de contenido lingüístico coherente, la traducción automática entre idiomas y el análisis de sentimientos y emociones expresadas en el lenguaje.

La relevancia del PLN adquiere una dimensión particularmente crítica en el ámbito de la ciberseguridad (“Cómo aplicar el procesamiento del lenguaje natural a la ciberseguridad,” 2023), donde se ha convertido en una herramienta indispensable para la protección digital. En este campo, el PLN ofrece capacidades transformadoras que permiten a los sistemas de seguridad automatizar procesos complejos y detectar amenazas de manera proactiva. Su aplicación abarca diversas áreas estratégicas de la seguridad informática, al procesar y dar sentido a registros de eventos, informes técnicos y fuentes abiertas de inteligencia sobre amenazas, identificando patrones que podrían indicar actividades sospechosas o ataques en curso. Esta capacidad es particularmente valiosa considerando el volumen masivo de datos que deben manejar los equipos de seguridad modernos.

El PLN integra múltiples técnicas para lograr que las máquinas o modelos comprendan, interpreten y generen lenguaje humano de manera efectiva. Su funcionamiento se basa en un conjunto de procesos interconectados, cada uno enfocado en un aspecto diferente del lenguaje (Josep, 2023). A continuación, se describen las etapas fundamentales que componen este sistema:

1. Tokenización (Fragmentación del Texto): El primer paso consiste en descomponer el texto en unidades mínimas llamadas tokens, que pueden ser palabras completas, sílabas o incluso caracteres individuales, dependiendo del enfoque utilizado. Este proceso es esencial para que el sistema pueda manipular y analizar el contenido de manera estructurada (Canorea, 2022).
2. Análisis Morfológico (Estructura de las Palabras): En esta fase, se examina la composición interna de cada término, identificando raíces, prefijos, sufijos y variaciones gramaticales (como género, número o tiempo verbal). Esto permite normalizar palabras diferentes pero relacionadas (por ejemplo, "corriendo" y "correr") para un procesamiento más eficiente (Daniel Llamas, 2019).
3. Análisis Sintáctico (La Gramática de las Oraciones): Aquí se estudia cómo se organizan las palabras dentro de una frase para establecer relaciones jerárquicas entre ellas. Se determinan elementos como el sujeto, el predicado y los complementos, lo que ayuda a entender la estructura lógica del mensaje (“Syntactic Analysis in Natural Language Processing,” 2025).
4. Análisis Semántico (El Significado Oculto): Más allá de la gramática, esta etapa busca descifrar el sentido real detrás de las palabras. Implica interpretar el contexto, resolver ambigüedades (como palabras con múltiples significados) e incluso inferir intenciones no explícitas en el texto (Rodrigo, 2025a).
5. Generación de Texto (De Datos a Lenguaje Natural): La capacidad inversa —crear contenido comprensible— requiere combinar todos los niveles anteriores. El sistema debe seleccionar términos adecuados, aplicar reglas gramaticales y mantener coherencia temática para producir respuestas fluidas y naturales (Rodrigo, 2025b).

Estos componentes no funcionan de manera aislada, sino que interactúan en un flujo continuo. Por ejemplo, para responder a una pregunta, un sistema de PLN debe tokenizar el input, analizar su sintaxis, comprender su semántica y finalmente generar una respuesta bien formada. Esta interdependencia es lo que permite aplicaciones avanzadas, como chatbots inteligentes o sistemas de traducción automática. Esta tecnología es versátil ya que se puede aplicar en diversos idiomas, lo que permite su uso global en múltiples contextos. Esta capacidad es fundamental, ya que el lenguaje humano es diverso y varía significativamente entre culturas y regiones. La adaptabilidad del PLN a diferentes idiomas implica no solo la traducción automática, sino también la comprensión de las particularidades gramaticales, semánticas y culturales de cada lengua.

La evolución del PLN ha dado lugar a modelos de lenguaje avanzados, como GPT-4, Google Translate, Siri de Apple, Alexa de Amazon y Google Assistant, que ofrecen capacidades de comprensión contextual más sofisticadas (SCILabs, 2025). Una tendencia clave es la integración del PLN con otras tecnologías emergentes, como los grafos de conocimiento, para crear sistemas de detección de amenazas más completos y precisos. Asimismo, existe un creciente énfasis en desarrollar modelos explicables (XAI) que permitan entender el proceso de toma de decisiones de estos sistemas, un aspecto crítico para su adopción en entornos de seguridad sensibles (Delgado, 2024). Sistemas equipados con esta tecnología pueden analizar automáticamente correos electrónicos, mensajes y publicaciones en redes sociales para identificar patrones asociados a actividades maliciosas. Esto incluye la detección de conductas transgresoras, donde el sistema puede reconocer señales de anomalías gramaticales características de estos ataques (Iglesias Álvarez, 2024).

El PLN presenta tres enfoques diferentes para su implementación (Stryker and Holdsworth, 2024). Las primeras implementaciones del PLN se fundamentaban en árboles de decisión simples que requerían un conjunto de reglas predefinidas. Estas aplicaciones solo podían ofrecer respuestas a comandos específicos, como la versión inicial de Moviefone (“Moviefone,” 2025), que contaba con capacidades básicas de generación de lenguaje natural. La ausencia de aprendizaje automático o inteligencia artificial en este tipo de PLN limita significativamente su funcionalidad y escalabilidad. Posteriormente, surgió el PLN estadístico, que se encarga de extraer, clasificar y etiquetar automáticamente los elementos de datos en texto y voz, asignando una probabilidad estadística a cada interpretación posible. Este enfoque se basa en el aprendizaje automático, lo que permite un análisis más detallado de la lingüística, como el etiquetado de partes del discurso. El PLN estadístico introdujo la técnica crucial de representar elementos del lenguaje, como palabras y reglas gramaticales, en forma de vectores y sentó las bases para desarrollos iniciales en PLN, tales como correctores ortográficos. En tiempos recientes, los modelos de aprendizaje profundo han dominado el campo del PLN, aprovechando grandes volúmenes de datos no estructurados, tanto de texto como de voz, para mejorar su precisión, puede considerarse una evolución del PLN estadístico, pero emplea redes neuronales para su funcionamiento (Stryker and Holdsworth, 2024). Dentro de este ámbito, existen varias subcategorías de modelos:

1. Modelos secuencia a secuencia (seq2seq): Basados en redes neuronales recurrentes (RNN) (“Recurrent neural network,” 2024), estos modelos son principalmente utilizados para la traducción automática, facilitando la conversión de frases de un idioma a otro (Sanz, 2020).
2. Modelos transformadores: Estos modelos utilizan la tokenización del lenguaje y la autoatención para evaluar las relaciones entre diferentes partes del texto. Se pueden entrenar de manera efectiva mediante aprendizaje auto-supervisado en grandes bases de datos textuales. Un avance significativo en este campo fue el desarrollo de BERT (Bidirectional Encoder Representations from Transformers) de Google, que se ha convertido en un pilar del motor de búsqueda de Google (Erick, 2024).
3. Modelos autorregresivos: Este tipo de modelo transformador está diseñado para predecir la siguiente palabra en una secuencia, lo que representa un avance importante en la generación de texto. Ejemplos de estos modelos incluyen GPT, Llama, Claude y el modelo de código abierto Mistra (“Modelo autorregresivo,” 2024).
4. Modelos fundacionales: Estos modelos pre-entrenados y seleccionados pueden acelerar la implementación de proyectos de PLN y aumentar la confianza en su rendimiento. Por ejemplo, los modelos fundacionales de IBM Granite son versátiles y aplicables en múltiples sectores, permitiendo tareas como la generación de contenido y la extracción de información (AWS, 2024).
   1. RNA (Redes Neuronales Artificiales)

Las redes neuronales artificiales (RNA) son sistemas computacionales inspirados en el funcionamiento del cerebro humano, diseñados para reconocer patrones y aprender a partir de datos. Están compuestas por unidades básicas llamadas neuronas artificiales, que se organizan en capas: una capa de entrada (que recibe los datos), una o varias capas ocultas (donde se procesa la información) y una capa de salida (que entrega el resultado) (ver Figura1). Estas redes utilizan algoritmos de aprendizaje automático, como el descenso de gradiente y la retropropagación, para ajustar sus parámetros (pesos y sesgos) y minimizar el error en las predicciones, optimizando así su precisión en tareas como clasificación, predicción o generación de contenido. Su capacidad para manejar datos complejos y no lineales las hace especialmente útiles en problemas donde las reglas tradicionales de programación resultan insuficientes (Leal et al., 2021).

Las RNA tienen aplicaciones transformadoras en diversos campos: en salud, permiten diagnósticos médicos más precisos, como la detección temprana de tumores en imágenes radiológicas (Shahid et al., 2019); en tecnología cotidiana, son la base del reconocimiento de voz en asistentes virtuales como Siri o Alexa y de los sistemas de recomendación personalizados en plataformas como Netflix o Spotify (Molina, 2022); y en industrias como la climatología, ayudan a modelar patrones climáticos para predecir fenómenos extremos (Buendía-Rodríguez et al., 2002; Servín-Palestina et al., 2022; Tirado Picado, 2024). Sin embargo, su eficacia depende en gran medida de la cantidad y calidad de los datos de entrenamiento, así como de una arquitectura bien diseñada. Además, enfrentan desafíos importantes, como su alto costo computacional, la dificultad para interpretar sus decisiones al ser consideradas "cajas negras" (Haskel-Ittah, 2023) y el riesgo de sesgos si los datos utilizados no son diversos o representativos. A pesar de estas limitaciones, las RNA representan una herramienta poderosa que está transformando múltiples áreas del conocimiento y la industria, siempre que se utilicen con responsabilidad ética y sostenibilidad.

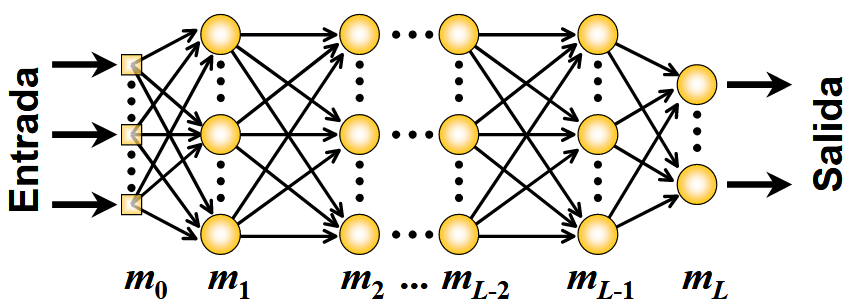


Figura 1 Estructura interna de una red neuronal artificial multicapa

* + 1. RNN (Redes Neuronales Recurrentes)

Las RNN son un tipo de arquitectura de redes neuronales diseñadas para procesar datos secuenciales o temporales, como series de tiempo, texto o audio, donde el orden de los elementos es crucial (“Recurrent neural network,” 2024). A diferencia de las RNA tradicionales, las RNNs incorporan bucles que permiten la persistencia de información entre pasos, manteniendo un estado oculto que actúa como memoria contextual (Stryker, 2021) (ver Figura 2). Permiten que la red recuerde información pasada alimentando la salida de un paso al siguiente paso. Esto ayuda a la red a comprender el contexto de lo que ya ha sucedido y a hacer mejores predicciones basadas en eso. Por ejemplo, al predecir la siguiente palabra en una oración, la RNN usa las palabras anteriores para ayudar a decidir qué palabra es más probable que venga a continuación (Arras et al., 2017).

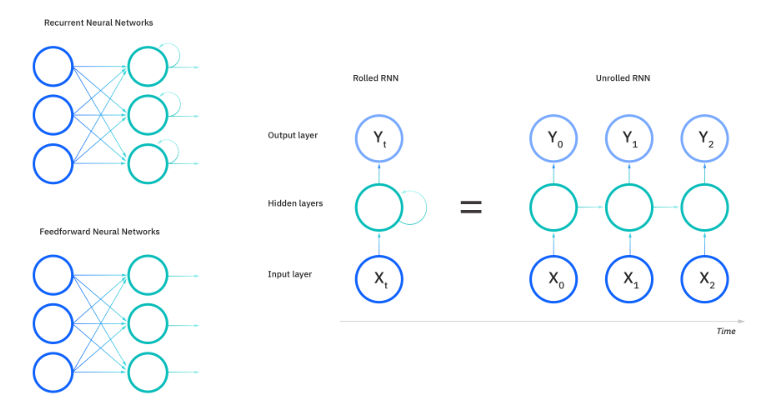


Figura 2 Redes neuronales pre-alimentadas contra redes neuronales recurrentes

Entre las ventajas de las RNN se encuentran su capacidad para procesar secuencias de longitud variable y su flexibilidad en aplicaciones como el PLN, el análisis de series temporales y el reconocimiento de voz. En el ámbito del PLN, las RNN son fundamentales para tareas como la generación de texto, la traducción automática y el análisis de sentimientos (Arras et al., 2017). Además, en el análisis de series temporales, ayudan a predecir valores futuros, como los niveles de inundación o las tendencias del mercado (Amalou et al., 2022).

Sin embargo, las RNN también enfrentan desafíos, como la inestabilidad de los gradientes durante el entrenamiento, lo que puede dificultar el aprendizaje. Tienen dificultades para recordar información a largo plazo en secuencias muy largas. Para superar estas limitaciones, se han desarrollado variantes como las LSTM (Long Short-Term Memory) (Hamad, 2023) y las GRU (Gated Recurrent Unit) (Anishnama, 2023), que mejoran la capacidad de recordar información a largo plazo y son más eficientes computacionalmente (Nosouhian et al., 2021).

La recurrencia que caracteriza a estas redes se ha visto reemplazadas por un mecanismo de autoatención mediante los Transformers, que han revolucionado el procesamiento de lenguaje natural, lo que permite procesar secuencias completas simultáneamente.

* + 1. Transformers

Tras explorar las RNA y RNN, surge una arquitectura que supera muchas de sus limitaciones, los Transformers. Introducidos en 2017 con el artículo "Attention is All You Need" (Vaswani et al., 2017), los Transformers abandonan por completo la recurrencia y las convoluciones, basándose en el mecanismo de auto-atención (self-attention) (Wydmanski, 2022), que permite a la red analizar todas las partes de una secuencia simultáneamente y determinar cómo se relacionan entre sí, sin importar su distancia.

Los Transformers están compuestos por bloques repetidos que incluyen capas de auto-atención seguidas de capas totalmente conectadas. Cada capa incorpora mecanismos de normalización y conexiones residuales que estabilizan y aceleran el aprendizaje, ver Figura 3. Dado que no procesan la información de manera secuencial, utilizan codificaciones posicionales para incorporar información sobre el orden de los elementos en la secuencia. Estas codificaciones pueden ser sinusoidales o aprendidas, y permiten que el modelo entienda la posición relativa de las palabras o tokens (“Transformador (modelo de aprendizaje automático),” 2025). Han demostrado un rendimiento sobresaliente en tareas de traducción automática, generación de texto, reconocimiento de voz, visión por computadora y más. Su capacidad para escalar a grandes volúmenes de datos y parámetros ha sido clave para el desarrollo de modelos de lenguaje masivos como GPT, BERT y sus derivados.

En el campo de PLN, la estructura de los transformer contiene componentes esenciales que le permiten procesar y generar texto de manera efectiva y eficiente (de la Torre Lorente, 2024). Entre estos elementos se encuentra el embedding o incrustación que convierte cada palabra o token en un vector de alta dimensión que representa su significado y contexto, permitiendo captar relaciones semánticas complejas.

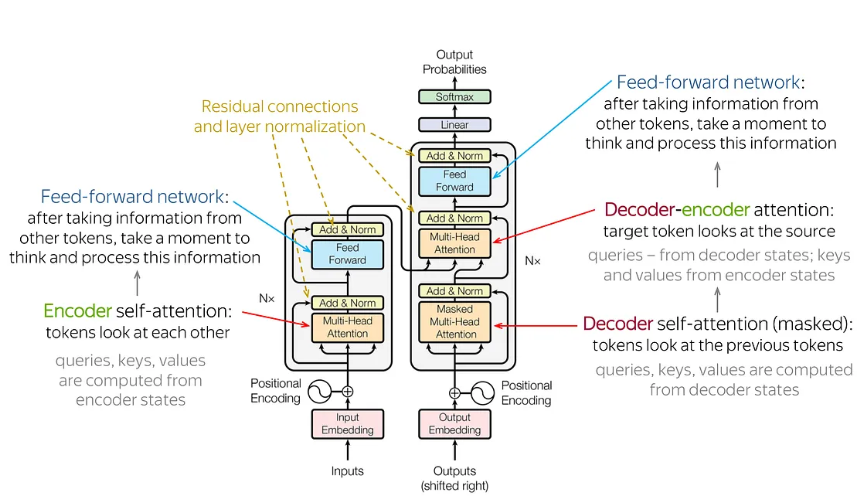


Figura 3 Arquitectura completa del modelo transformador (Petriconi, 2022)

La auto-atención evalúa la importancia de cada palabra respecto a las demás en la misma secuencia, capturando dependencias a largo plazo y la atención cruzada se utiliza en tareas como traducción, alineación de palabras entre secuencias de entrada y salida. Las Redes Neuronales Feedforward (FFNN) (Cota, 2023) procesan los vectores enriquecidos por la atención para refinar las representaciones mediante capas densas, la normalización y conexiones residuales mejoran la estabilidad y velocidad del entrenamiento, facilitando el flujo de información y evitando la degradación del gradiente, ver Figura 4.

Durante el entrenamiento, el Transformer aprende a predecir la siguiente palabra en una secuencia basándose en el contexto previo, ajustando sus parámetros para minimizar errores. Este aprendizaje permite al modelo captar patrones y relaciones complejas en el lenguaje. Una vez entrenado, el modelo puede realizar diversas tareas de procesamiento de lenguaje natural, como traducción automática o generación de texto coherente a partir de una entrada inicial (de la Torre Lorente, 2024). Gracias a su capacidad para analizar todas las partes de una secuencia simultáneamente y manejar múltiples dependencias en paralelo, los Transformers ofrecen un rendimiento superior en tareas complejas de lenguaje natural, superando a modelos anteriores y ampliando las posibilidades en inteligencia artificial.

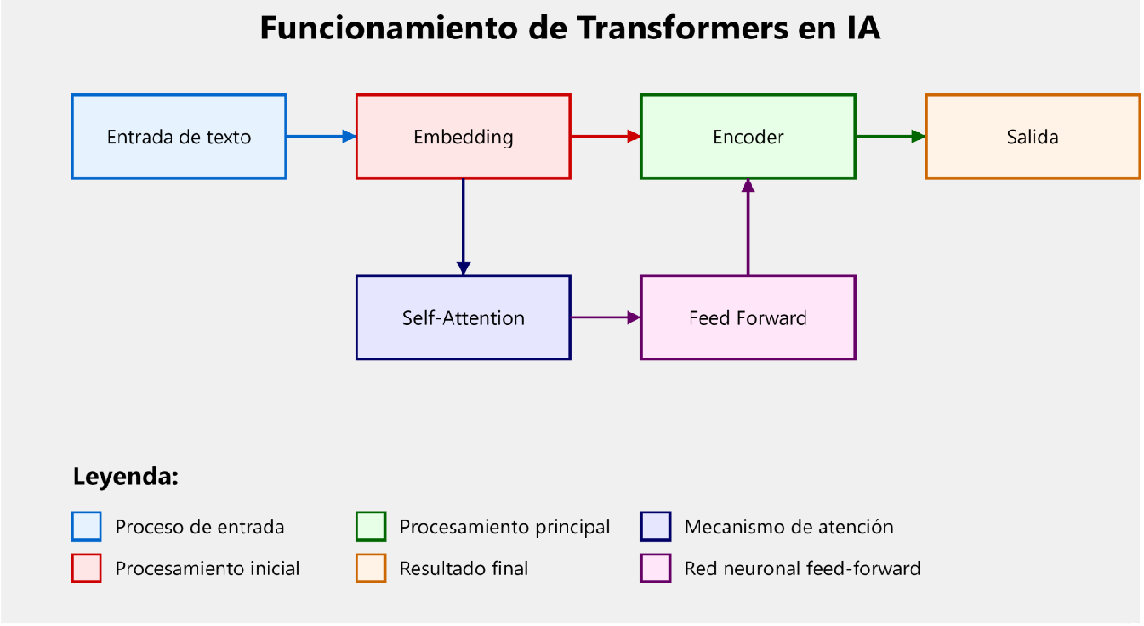


Figura 4 Componentes de la estructura de un Transformer

En el ámbito del PLN, los modelos de lenguaje basados en Transformers, desarrollados en el siguiente epígrafe han mostrado resultados notables en inglés y español. La evolución de la tecnología y el acceso a grandes volúmenes de datos en diferentes idiomas ha permitido que los modelos de PLN se entrenen de manera más eficiente. Esto asegura que tanto hablantes de inglés como de español, y otros idiomas, puedan beneficiarse de las innovaciones en este campo, mejorando así la comunicación y la interacción entre humanos y máquinas.

* 1. Grandes Modelos de Lenguaje

Los Grandes Modelos de Lenguaje (LLM por sus siglas en inglés) representan la vanguardia en inteligencia artificial, emplean arquitecturas masivas de aprendizaje profundo entrenadas con volúmenes extraordinarios de datos. Estos sistemas se basan en los transformer, que son el corazón de los LLM modernos, permitiéndoles entender y generar de manera eficiente y escalable (Jimenez Melo, 2023), ya que son una arquitectura compuesta por dos componentes neurales fundamentales, un módulo codificador y otro decodificador, ambos dotados de sofisticados mecanismos de auto-atención. El codificador actúa como un analizador exhaustivo, procesando la totalidad del texto de entrada (como una oración completa) para construir una representación matemática compacta que encapsula su significado esencial. Posteriormente, el decodificador toma esta representación abstracta y la transforma progresivamente en una secuencia de salida, permitiendo aplicaciones como la traducción automática entre idiomas.

Su enfoque de auto-atención permite a los sistemas dominar aspectos fundamentales del lenguaje como gramática, sintaxis y conocimiento semántico sin necesidad de intervención humana directa (Bergmann, 2024). A diferencia de las RNN, que analizan la información secuencialmente, los transformadores procesan datos lingüísticos en paralelo, aprovechando al máximo la capacidad de procesamiento de unidades GPU y reduciendo drásticamente los tiempos de entrenamiento.

La arquitectura transformadora de las LLM destaca especialmente por su escalabilidad, permitiendo la creación de modelos con una complejidad sin precedentes. Estas redes neuronales pueden llegar a incorporar cientos de miles de millones de parámetros ajustables, lo que explica su extraordinaria capacidad para comprender y generar lenguaje humano con un nivel de sofisticación cada vez más cercano al humano. Estos modelos pueden incorporar cantidades masivas de datos, a menudo de Internet, pero también de fuentes como Common Crawl (“Common Crawl,” 2024), que comprende más de 50 000 millones de páginas web, y Wikipedia, que tiene aproximadamente 57 millones de páginas.

Los LLM pueden considerar miles de millones de parámetros con muchos usos posibles (Aryan, 2023) (ver Figura 5). El modelo GPT-3 de OpenAI (Li, 2020) tiene 175 000 millones de parámetros, ChatGPT, puede identificar patrones a partir de datos y generar resultados naturales y legibles. Si bien no sabemos el tamaño de Claude 2, puede aceptar entradas con hasta 100 000 tokens en cada indicación, lo que significa que puede funcionar en cientos de páginas de documentación técnica o, incluso, en un libro completo. El modelo Jurassic-1 de AI21 Labs (Padnos, 2023) tiene 178 000 millones de parámetros y un vocabulario simbólico de partes de 250 000 palabras y capacidades de conversación similares. El modelo Command de Cohere tiene capacidades similares y puede funcionar en más de 100 idiomas diferentes (Kayid and Reimers, 2022).

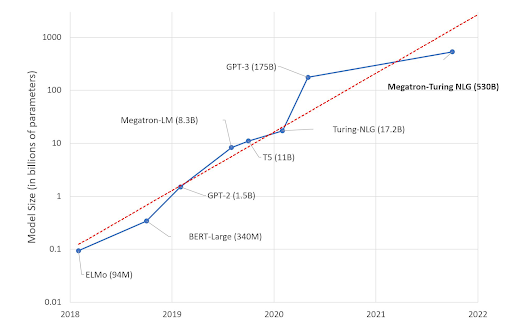


Figura 5 Cantidad de parámetros (en billones) de algunos de los LLM con más conocidos desde el 2018 al 2022

1.3.1 BERT (Bidirectional Encoder Representations from Transformers)

La investigación (Devlin et al., 2019) introdujo a BERT (Bidirectional Encoder Representations from Transformers), un modelo LLM revolucionario en el campo del PNL. Gracias a su arquitectura basada en transformadores y su entrenamiento bidireccional, BERT ha establecido nuevos estándares en tareas como comprensión de texto, análisis de sentimientos y respuesta a preguntas. Su capacidad para capturar el contexto completo de una palabra al considerar tanto el texto precedente como el subsiguiente lo distingue de enfoques anteriores, que procesaban el lenguaje de manera secuencial y unidireccional.

BERT fue pre-entrenado utilizando dos estrategias clave:

1. Masked Language Modeling (MLM): Oculta aleatoriamente el 15% de las palabras en una oración y entrena el modelo para predecirlas, permitiéndole aprender representaciones contextuales profundas (Murel Ph.D. and Kavlakoglu, 2024).
2. Next Sentence Prediction (NSP): Entrena al modelo para determinar si dos oraciones están relacionadas, mejorando su comprensión de la coherencia textual (Malingan, 2023).

El modelo se entrenó inicialmente en dos versiones:

1. BERT-Base: 110 millones de parámetros (12 capas de transformadores).
2. BERT-Large: 340 millones de parámetros (24 capas).

Ambas variantes utilizaron corpus masivos como Wikipedia y el BookCorpus, lo que les permitió generalizar eficazmente en múltiples dominios.

En este contexto, la representación de entrada en BERT (Narein T, 2021) juega un papel fundamental como se ilustra en la Figura 6, las incrustaciones (embeddings) son una representación numérica que convierte datos como palabras, imágenes o categorías en vectores de números reales en un espacio continuo, son la suma de las incrustaciones de los tokens, las incrustaciones de segmentación y las incrustaciones de posición. En el contexto del PLN, transforman palabras en vectores densos que capturan relaciones semánticas y contextuales entre ellas (Narein T, 2021). Esto permite a las máquinas entender el significado y las conexiones entre palabras más allá de su forma literal. Las que utiliza el modelo son una combinación de tres componentes principales (Metzger, 2022) que forma la entrada final que lo alimenta:

1. Token Embeddings (Canorea, 2022): Representaciones vectoriales que codifican cada palabra o token en el texto.
2. Segment Embeddings (Raman, 2021): Indicadores que diferencian segmentos dentro del texto (por ejemplo, dos oraciones distintas).
3. Position Embedding (Preetham, 2024): Representaciones que incorporan información posicional para preservar el orden secuencial de los tokens.

Esta estructura permite a BERT procesar información contextual rica y bidireccionalmente, lo cual es esencial para tareas como el MLM y la NSP.

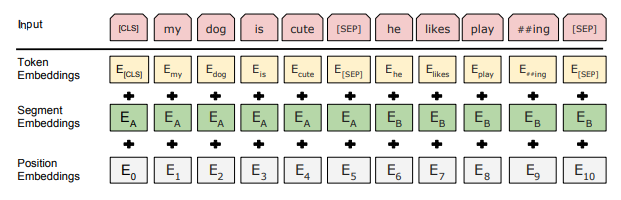


Figura 6 La representación de entrada de BERT.

Su diseño multilingüe generalista limitaba su eficacia en idiomas específicos como el español, donde fenómenos lingüísticos (morfología, giros culturales) requieren adaptaciones especializadas. Esto impulsó el desarrollo de modelos derivados como mBERT (Multilingual BERT) (Google Bert, 2024) desarrollado por el equipo de HuggingFace (“Hugging Face,” 2025) y pre-entrenado con 104 lenguajes, también se crearon adaptaciones entrenadas para el lenguaje de español llamadas es-BERT (Spanish BERT) como BETO (Cañete et al., 2023) haciendo ajuste fino (fine-tuning) (Haponik, 2023) al modelo multilenguaje de BERT mostrando mejores resultados al ser entrenado con casos específicamente en español, BERTin (Rosa et al., 2022) el cual utilizó muestreo de perplejidad, disminuyendo el pre-entrenamiento del modelo en aproximadamente la mitad de pasos, y con una quinta parte de los datos normalmente necesarios. Paralelamente, RoBERTa (Robustly Optimized BERT Approach) emergió como otra optimización de BERT, eliminando la tarea de NSP e introduciendo técnicas como el enmascaramiento dinámico y entrenamiento con datasets diez veces mayores (160GB de texto) (Liu et al., 2019), lo que mejoró su robustez para tareas multilingües. Sin embargo, incluso estas mejoras mostraban limitaciones en español, lo que motivó adaptaciones especializadas como RoBERTuito (Pérez et al., 2022), pre-entrenado con un corpus de cerca de 500 millones de tweets de más de 400 usuarios aleatorios logró capturar peculiaridades del lenguaje informal, como jerga, emoticonos, hashtags y estructuras gramaticales no convencionales. Al realizar análisis de sentimiento y detección de emociones (García Vega et al., 2020a) en las tareas 1 y 2 de TASS 2020 (García Vega et al., 2020b), detección de ironía en los datasets del Fórum de Evaluación de Lenguajes Ibéricos (IberLEF 2019) (García Cumbreras et al., 2019) y detección de discurso de odio en la prueba del HateEval del 2019 (Basile et al., 2019), se obtienen resultados superiores a los demás modelos de dominio público en tareas de clasificación en español (ver Figura 7). Dichos resultados se expresan como la puntuación promedio de Macro F1 obtenida en 10 ejecuciones de los experimentos de clasificación. En negrita se indican los modelos con mejor rendimiento.

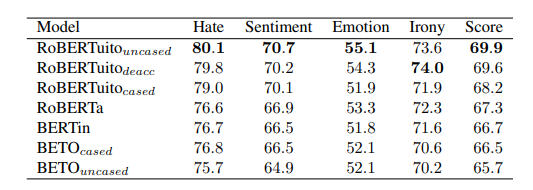


Figura 7 Resultados de evaluación para tareas de clasificación en español.

La investigación de (Cañete et al., 2023; Pérez et al., 2022) destaca el rendimiento superior de los modelos pre-entrenados BETO y RoBERTuito en tareas en español, gracias a su entrenamiento y optimización de manera especializada, particularmente para el análisis de sentimientos. Por ello, se eligen estos modelos para analizar sentimientos en tweets en español, con el fin de evaluar su eficacia en la identificación de polaridad sentimental en textos de redes sociales y comparar su desempeño en esta tarea.

1.3.2 BETO

BETO es el primer modelo BERT pre-entrenado exclusivamente para español, utilizando un gran corpus de textos en español provenientes de fuentes como Wikipedia y el proyecto OPUS (“OPUS Home - Open and Universal Science (OPUS) Project,” 2022). El volumen de datos empleado es comparable al corpus original de BERT. BETO también incorpora técnicas de entrenamiento exitosas de modelos anteriores, como el dynamic masking (Ivezic and Ivezic, 2022) y el enmascaramiento de palabras completas Whole-Word Masking (WWM) (Yeolekar, 2024) (ver Tabla 1) garantiza que todos los subtokens de una palabra seleccionada se enmascaren de manera conjunta. Ha demostrado un rendimiento excepcional en diversas tareas, superando consistentemente a mBERT y a es-BERT y logrando resultados de vanguardia en conjuntos de datos estándar para español como POS (Part of Speech) y MLDoc (Cañete, 2025).

|  |  |
| --- | --- |
| **Método** | **Oración enmascarada** |
| Estilo de BERT | The quick [MASK] ##wn fox [MASK] ##s over the lazy dog |
| Whole Word | The quick [MASK] [MASK] fox [MASK] over the lazy dog |

Tabla 1 Enmascaramiento de la oración: “The quick brown fox jumps over the lazy dog” utilizando WWM.

El proceso de WWM (ver Figura 8) comienza con una frase que se divide en tokens. Los tokens pueden ser palabras completas o sub-palabras. Se eligen ciertas palabras para enmascarar. Las palabras seleccionadas se reemplazan con el token. Sin embargo, las sub-palabras mantienen su forma original. Todos los tokens que pertenecen a las palabras seleccionadas se enmascaran juntos.

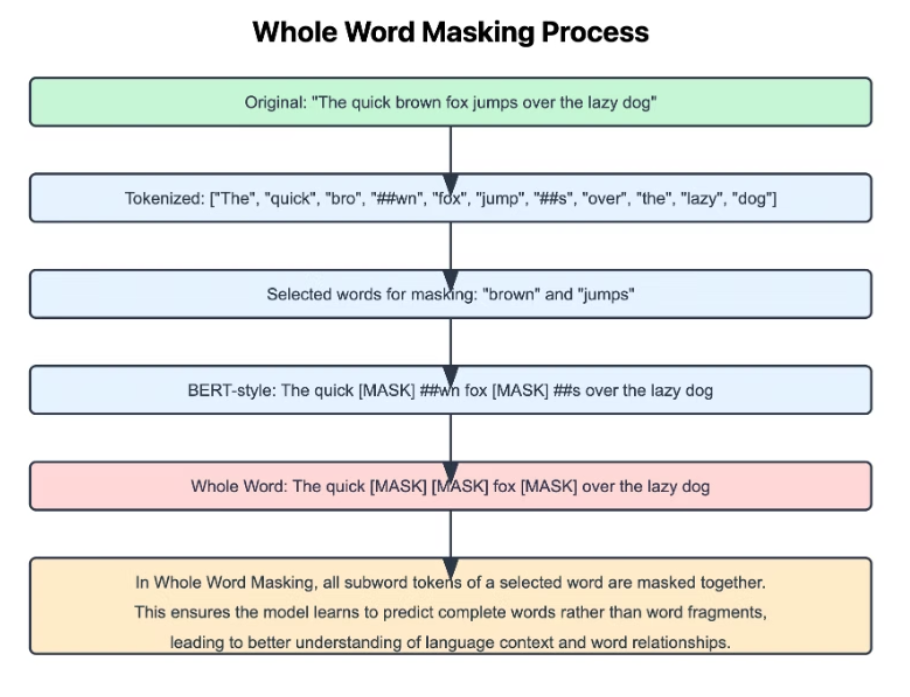


Figura 8 Funcionamiento de Whole-Word Masking

El objetivo es asegurar que el modelo aprenda a predecir palabras completas en lugar de fragmentos de palabras, lo que mejora la comprensión del contexto del lenguaje y las relaciones entre las palabras (Yeolekar, 2024).

1.3.3 RoBERTuito

RoBERTuito es presentado como un modelo especializado en textos de redes sociales en español, está optimizado para contenido generado por usuarios en esa lengua. Es basado en la arquitectura RoBERTa (Liu et al., 2019), con un enfoque BERT optimizado de manera robusta y entrenado con enmascaramiento dinámico, FULL-SENTENCES (oraciones completas) sin pérdida por NSP, Mini-lotes grandes para mejorar la estabilidad y eficiencia, un Byte-level BPE (Wang et al., 2019) más amplio que le permite manejar caracteres especiales y errores ortográficos comunes en redes sociales (Liu et al., 2019). RoBERTuito adoptó hiperparámetros exitosos de modelos anteriores utilizando un tamaño de lote grande. En tareas de clasificación con datos generados por usuarios en español como análisis de sentimientos y de emociones utilizando los conjuntos de datos TASS 2020 (García Vega et al., 2020b), detección de sarcasmo utilizando el conjunto de datos IrosVa 2019 (Ortega-Bueno et al., 2019) y detección de discurso de odio utilizando el conjunto de datos HatEval del SemEval 2019, Tarea 5 (“Multilingual detection of hate speech against immigrants and women in Twitter (hatEval),” 2019), RoBERTuito supera significativamente a otros modelos pre-entrenados con datos del mismo lenguaje, destacándose especialmente en la detección de discurso de odio y el análisis de sentimientos.

Aunque tanto BETO como RoBERTuito están diseñados para español, sus fundamentos arquitectónicos y metodologías de entrenamiento difieren (Pérez et al., 2022). La arquitectura de BETO se alinea estrechamente con BERT, centrándose en el Masked Language Modeling (Murel Ph.D. and Kavlakoglu, 2024) y en la comprensión de relaciones entre oraciones mediante mecanismos de atención. RoBERTuito en cambio prioriza entrenamientos con lotes más grandes y datos más extensos, optimizando el modelo para contenido dinámico e informal típico de redes sociales.

Estos modelos consideran la gramática, los patrones de vocabulario y las estructuras de textos de redes sociales en español, mejorando su capacidad para entender y procesar el idioma. Los modelos pre-entrenados generales pueden no adaptarse igual de bien a los matices específicos del español, por lo que usar modelos dedicados como BETO y RoBERTuito suele generar resultados superiores, como en (Molero et al., 2023) que se utiliza RoBERTuito con el dataset MeOffendES (Plaza-del-Arco et al., 2021) para detección de lenguaje ofensivo en español, se muestran resultados para cada categoría: NO (no ofensivo), NOM (no ofensivo con lenguaje explícito), OFG (ofensivo hacia grupos) y OFP (ofensivo hacia personas), junto con los promedios macro y micro, ver Figura 9.

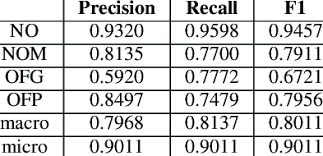


Figura 9 Resultados de clasificación del modelo RoBERTuito en el dataset MeOffendES para detección de lenguaje ofensivo en español.

La evolución de los LLM, como BERT, BETO y RoBERTuito, ha revolucionado el PLN al ofrecer capacidades avanzadas de comprensión y generación de texto. Sin embargo, su implementación efectiva requiere de herramientas computacionales especializadas que permitan entrenar, ajustar y desplegar estos modelos en escenarios reales, como plataformas, bibliotecas y frameworks clave que facilitan el desarrollo de aplicaciones basadas en PLN.

* 1. Herramientas computaciones para PLN

La detección de comportamientos transgresivos en redes requiere del uso de herramientas computacionales basadas en PLN que hagan posible el procesamiento de grandes cantidades de datos de manera automatizada. Estas herramientas se han convertido en esenciales para identificar, clasificar y mitigar dichas conductas de manera automatizada y eficiente. A diferencia de la moderación manual, que resulta insuficiente por su lentitud y subjetividad, los sistemas de PLN permiten analizar grandes volúmenes de texto en tiempo real, detectar patrones lingüísticos asociados a comportamientos nocivos y priorizar contenido para revisión humana. El uso de modelos avanzados de aprendizaje automático y aprendizaje profundo ha mejorado la precisión en la detección de contextos complejos, como el sarcasmo, las variaciones culturales del lenguaje o el uso de eufemismos. La implementación de estas herramientas no solo beneficia a plataformas digitales y empresas, sino también a investigadores y autoridades que buscan comprender fenómenos sociales en línea. Estas herramientas combinan técnicas de aprendizaje automático, modelos de lenguaje y bases de datos léxicas para mejorar la precisión en la identificación de conductas nocivas.

* + 1. ATLAS.ti

ATLAS.ti es un software de pago especializado en el análisis cualitativo de datos, ampliamente empleado por investigadores en áreas como las ciencias sociales, la antropología y la psicología. Esta herramienta permite organizar, analizar y visualizar grandes volúmenes de información cualitativa, incluyendo entrevistas, grupos focales, documentos textuales, imágenes, audios y videos, lo que la convierte en una plataforma versátil para la investigación académica y profesional. Entre sus principales funcionalidades destacan la codificación avanzada de datos, la creación de redes conceptuales y la generación de representaciones visuales de los hallazgos, facilitando así la interpretación y comunicación de resultados complejos (Friese, 2021). Este procedimiento es fundamental para identificar patrones, tendencias y relaciones dentro de los datos, contribuyendo a interpretaciones más profundas y matizadas. El software incorpora herramientas de auto-codificación asistidas por inteligencia artificial, que agilizan el etiquetado y análisis de grandes conjuntos de datos, así como funciones para comparar y recuperar fácilmente segmentos codificados, lo que resulta especialmente útil para el seguimiento de cambios a lo largo del tiempo o entre distintas fuentes (Friese, 2021).

Otra característica destacada de ATLAS.ti es la posibilidad de construir redes visuales que representan las relaciones entre códigos, categorías y conceptos, proporcionando un mapa conceptual que facilita la identificación de conexiones, solapamientos y vacíos en la información analizada. Estas visualizaciones permiten a los investigadores obtener una comprensión integral de los datos y descubrir nuevos conocimientos que pueden no ser evidentes mediante un análisis textual tradicional (Justicia, 2003). Adicionalmente, ATLAS.ti ofrece opciones avanzadas de visualización, como diagramas, nubes de palabras y mapas interactivos, que enriquecen la presentación y el análisis de los resultados.

* + 1. Brand24

Brand24 es una potente herramienta de pago, se enfoca en el monitoreo en línea que permite a las empresas rastrear y analizar en tiempo real las conversaciones sobre su marca en internet. A través de Brand24, las compañías pueden supervisar plataformas de redes sociales, blogs, foros, sitios de noticias y otras fuentes digitales para obtener información valiosa sobre cómo su marca es percibida por su público objetivo (Casas, 2025). La plataforma proporciona una visión integral de las menciones en línea, análisis de sentimientos y alcance en redes sociales, lo que facilita a las empresas mantenerse al tanto de su reputación digital y tomar decisiones informadas para mejorar la imagen de su marca.

Una de las funciones clave de Brand24 es la capacidad de monitorear palabras claves, por ejemplo, relacionadas con un negocio o producto, lo que permite conocer la frecuencia con la que se menciona la marca y el sentimiento asociado a esas menciones (Bernzweig, 2023). Al analizar estos datos, las empresas obtienen conclusiones valiosas sobre la percepción de sus clientes y pueden identificar áreas de mejora. Ofrece análisis detallados e informes que ayudan a medir el impacto de las campañas de marketing digital y a tomar decisiones basadas en datos para potenciar la presencia en línea de la marca.

Más allá del monitoreo, Brand24 facilita la interacción en tiempo real con la audiencia, permitiendo responder a publicaciones, comentarios y reseñas en redes sociales. Esta capacidad de respuesta fortalece la relación con los clientes, mejora la satisfacción y fomenta la lealtad hacia la marca. La herramienta proporciona los recursos necesarios para participar activamente en las conversaciones digitales y aprovechar las redes sociales como un canal estratégico para la gestión de la reputación (Casas, 2025).

* + 1. SentiWordNet

WordNet es una extensa base de datos léxica desarrollada originalmente para el idioma inglés por la Universidad de Princeton, aunque actualmente existen versiones en diversos idiomas. Esta herramienta se ha convertido en un recurso fundamental para la lingüística computacional y el PLN. La característica distintiva de WordNet es su organización conceptual, en lugar de ordenar alfabéticamente las palabras como un diccionario tradicional, WordNet agrupa sustantivos, verbos, adjetivos y adverbios en conjuntos de sinónimos cognitivos llamados synsets, donde cada uno expresa un concepto distinto (Princeton University, 2010).

Estos synsets son la unidad básica de información en WordNet, cada uno representa un concepto específico y se codifica mediante un número único de ocho dígitos denominado offset (Ferreyra (comp), 2018). Esta estructura permite una desambiguación semántica precisa, diferenciando a WordNet de los diccionarios de sinónimos tradicionales, ya que no solo articula formas de palabras como cadenas de letras, sino sentidos específicos de las palabras. WordNet establece relaciones explícitas entre los synsets, creando una compleja red semántica navegable donde las palabras cercanas están semánticamente relacionadas.

La estructura de WordNet se basa en relaciones semánticas y léxicas bien definidas. La relación principal entre palabras es la sinonimia, pero también incluye otras relaciones conceptuales que conectan los aproximadamente 117 000 synsets entre sí (Princeton University, 2010). Cada synset contiene una breve definición y, en la mayoría de los casos, una o más frases cortas que ilustran el uso de sus miembros. Las palabras con varios significados distintos están representadas en múltiples synsets, permitiendo una gran precisión en la representación del conocimiento léxico.

SentiWordNet es un recurso léxico que extiende WordNet para proporcionar información sobre polaridad sentimental. Mientras WordNet se centra en relaciones semánticas objetivas, SentiWordNet añade una capa de valoración subjetiva, asignando a cada synset de WordNet valores de positividad, negatividad y objetividad (Esuli and Sebastiani, 2006). Esta extensión surge como respuesta a la necesidad de analizar computacionalmente el lenguaje subjetivo en aplicaciones como la minería de opiniones y el análisis de sentimientos.

La estructura de SentiWordNet mantiene la organización en synsets de WordNet, pero añade tres valores numéricos a cada uno. La positividad que define cuán positivo es el término, negatividad refiriéndose a cuan negativo es y objetividad que define cuán objetivo o neutral es. Estos tres valores suman 1, representando la distribución probabilística de las tres categorías de polaridad para cada synset. Esta anotación permite derivar valores de sentimiento para palabras, frases y textos completos, proporcionando una base cuantitativa para la interpretación automática de opiniones (Husnain et al., 2021).

La combinación de SentiWordNet con técnicas de aprendizaje automático, particularmente con modelos de word embeddings (Barnard, 2023) como Word2Vec o GloVe (Yadav, 2024), ha mostrado mejoras significativas en la captura de relaciones semánticas y en la ejecución de tareas como clasificación de texto y sistemas de recomendación. Por ejemplo, en el análisis de sentimientos, los word embeddings pueden ayudar a detectar distintas denominaciones de una misma entidad (target) que aparecen en contextos similares, complementando así la información proporcionada por SentiWordNet (Moré, 2019).

* + 1. TweetNLP

TweetNLP se posiciona como una herramienta especializada en el análisis de texto generado en redes sociales, particularmente en Twitter, donde la informalidad lingüística y la dinámica comunicativa requieren enfoques adaptativos. Desarrollado como un ecosistema modular, combina modelos transformer optimizados para lenguaje social con funciones de preprocesamiento específicos, permitiendo aplicaciones avanzadas en minería de opiniones, detección de contenido ofensivo y análisis semántico contextualizado (Camacho-collados et al., 2022).

El sistema integra técnicas avanzadas para normalizar texto no convencional como la traducción de emojis a descriptores semánticos (🔥 a "emoji fuego"), la descomposición de hashtags mediante separación morfológica (#VamosPorMás a "vamos por más"), la normalización de jerga digital (ej. "xq" a "porque", "tb" a "también") y la detección de entidades dinámicas (menciones, URLs, trending topics). Este proceso prepara el texto para modelos lingüísticos, reduciendo la ambigüedad inherente a los mensajes sociales. TweetNLP utiliza arquitecturas basadas en RoBERTa y BERT reentrenadas con 550 millones de tweets, optimizando su capacidad para capturar contextos efímeros y referencias culturales, interpretar ironía y sarcasmo mediante patrones léxico-sintácticos y manejar neologismos y variaciones dialectales en tiempo real. Estos modelos, de tamaño moderado con aproximadamente 110 millones de parámetros, permiten ejecución en hardware estándar sin sacrificar precisión. La plataforma ofrece módulos especializados para nueve tareas clave (Camacho-collados et al., 2022) ver Figura 10

Pese a sus avances, TweetNLP enfrenta retos en precisión en la clasificación de texto en lenguajes específicos como el español, interpretación de referencias culturales hiperlocales, análisis de threads conversacionales extensos y sesgos en lenguaje inclusivo/no binario.

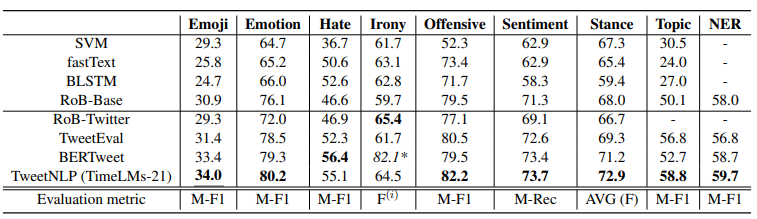


Figura 10 Resultados de pruebas en las 9 tareas soportadas por TweetNLP

* + 1. Librerías de Procesamiento de Lenguaje Natural

El PLN ha experimentado un avance revolucionario en los últimos años, gracias al desarrollo de bibliotecas especializadas que democratizan el acceso a modelos avanzados y técnicas de vanguardia. Desde el análisis de sentimientos hasta la generación de texto, estas herramientas permiten a investigadores, desarrolladores y empresas implementar soluciones complejas con pocas líneas de código.

A continuación, se explorarán las bibliotecas más relevantes para PLN, cada una con sus fortalezas y casos de uso específicos. Desde opciones generalistas como spaCy y NLTK, hasta frameworks modernos basados en transformers como Hugging Face Transformers y soluciones como PySentimiento, especializadas para idiomas como el español.

* + - 1. spaCy

spaCy es una biblioteca de código abierto para Python, diseñada específicamente para el PLN avanzado en contextos industriales y de investigación. Se distingue por su eficiencia, robustez y facilidad de integración en sistemas de producción, características que la han posicionado como una de las herramientas más utilizadas en el ámbito del PLN contemporáneo. Su estructura se basa en un modelo de datos centralizado que representa documentos como objetos, con anotaciones pre-calculadas para análisis morfológico, sintáctico y semántico. La filosofía de diseño de spaCy prioriza la velocidad de ejecución y la facilidad de uso, ofreciendo una API intuitiva y modelos pre-entrenados optimizados para diversas tareas de PLN en múltiples idiomas (Vasiliev, 2020).

El núcleo funcional de spaCy representa el texto procesado como una secuencia de tokens enriquecidos con anotaciones lingüísticas, tales como lemas, etiquetas gramaticales, dependencias sintácticas y entidades nombradas. El procesamiento del texto se realiza de manera modular y conFigurable, donde cada componente ejecuta tareas específicas, como la tokenización, el etiquetado de partes del discurso, el análisis de dependencias, la lematización y el reconocimiento de entidades nombradas. Esta arquitectura permite personalizar y ampliar, integrando componentes propios o ajustando los existentes para adaptarse a los requisitos de cada proyecto (Vasiliev, 2020).

spaCy se utiliza ampliamente en diversas aplicaciones de PLN, desde la extracción de información y la clasificación de textos hasta la construcción de chatbots y el análisis de sentimientos. Su eficiencia y su API intuitiva lo convierten en una herramienta popular entre investigadores y desarrolladores que buscan construir sistemas de PLN robustos y escalables. La documentación de spaCy proporciona ejemplos detallados y guías prácticas para implementar diversas tareas de PLN, facilitando el aprendizaje y la adopción de la biblioteca. Los modelos preentrenados de spaCy abarcan múltiples idiomas y dominios, permitiendo abordar una amplia variedad de problemas de PLN con un mínimo esfuerzo de conFiguración.

* + - 1. NLTK (VADER)

Natural Language Toolkit (NLTK) es una plataforma poderosa para construir programas en Python que trabajan con datos de lenguaje humano (Bird et al., 2009a). Ofrece interfaces fáciles de usar para más de 50 colecciones estructuradas de texto y recursos léxicos como WordNet, junto con un conjunto de bibliotecas para procesamiento de texto que incluyen clasificación, tokenización, derivación, etiquetado, análisis sintáctico y razonamiento semántico (Bird et al., 2009b). NLTK es ampliamente utilizado para análisis de texto y tareas de PLN, siendo una herramienta indispensable para investigadores, desarrolladores y estudiantes en el campo de la lingüística computacional.

VADER (Valence Aware Dictionary and sEntiment Reasoner) es un léxico y herramienta basada en reglas para análisis de sentimientos, especialmente adaptada para los sentimientos expresados en redes sociales. Está incluida en la biblioteca NLTK y está diseñada para analizar e interpretar el sentimiento del texto de manera más precisa y matizada que las herramientas tradicionales. VADER puede reconocer e interpretar tanto la polaridad como la intensidad de los sentimientos expresados en el texto con implementaciones muy sencillas como muestra la Figura 11, lo que la convierte en una herramienta valiosa para el análisis de sentimientos en datos de redes sociales y en análisis de retroalimentación de clientes (Geetha, 2023).

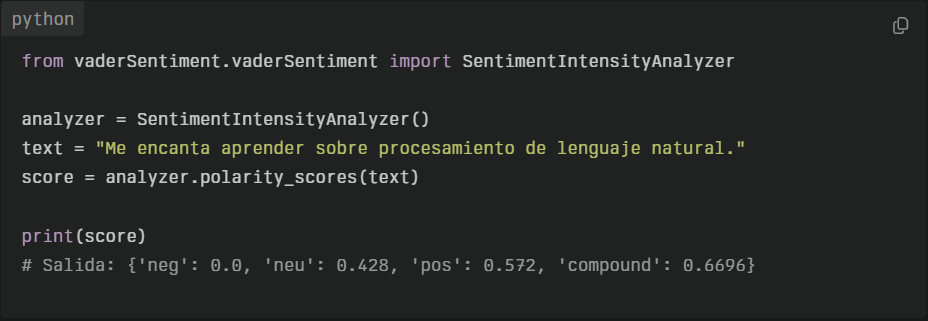


Figura 11 Ejemplo básico de análisis de sentimiento con VADER

Una de las ventajas clave de NLTK y VADER es su facilidad de uso e integración con el lenguaje de programación Python. Ambas son herramientas de código abierto que pueden ser fácilmente instaladas mediante el gestor de paquetes de Python, lo que facilita a los desarrolladores incorporarlas en sus proyectos de PLN. Además, NLTK y VADER ofrecen documentación detallada y ejemplos que ayudan a los usuarios a entender e implementar sus diversas funcionalidades, haciéndolas accesibles incluso para principiantes en el campo del procesamiento de lenguaje natural.

* + - 1. TextBlob

TextBlob es una biblioteca para PLN de código abierto en Python que ofrece una interfaz sencilla y accesible para realizar tareas comunes como etiquetado de partes del discurso, extracción de frases nominales, análisis de sentimientos, clasificación, tokenización, corrección ortográfica y más (Loria, 2025a). Está construida sobre las robustas bibliotecas NLTK y Pattern (Tulkens, 2020), lo que le permite ofrecer resultados precisos sin requerir conocimientos avanzados en procesamiento lingüístico. Utiliza el algoritmo Naive Bayes de aprendizaje supervisado que se basa en el teorema de Bayes sobre la probabilidad condicional en la que se quiere calcular la probabilidad (Figura 12) de que ocurra un evento C, sabiendo que ha ocurrido X (Pauli, 2019).Toma cada clase como variable independiente y predice la probabilidad de que pertenezca a un grupo, sea este positivo, negativo o neutro.

Su diseño facilita a desarrolladores e investigadores trabajar con datos textuales mediante una API intuitiva que simplifica operaciones complejas, apoyándose en recursos como WordNet y permitiendo extensiones personalizadas (Loria, 2025b). Aunque es especialmente valorada por su facilidad de uso y versatilidad en tareas generales de PLN, su análisis de sentimientos se basa en un léxico predefinido, lo que la hace más adecuada para textos formales o generales que para lenguaje muy informal o jerga.

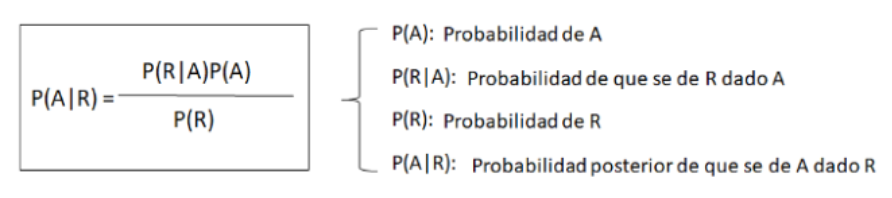


Figura 12 Probabilidad condicional (Roman, 2019)

* + - 1. Transformer

La librería transformers de Hugging Face es un proyecto de código abierto para Python que democratiza el acceso a miles de modelos preentrenados de última generación en PLN, visión por computadora, audio y aprendizaje multimodal (Wolf et al., 2020). Diseñada como una interfaz unificada para arquitecturas basadas en transformers como BERT, GPT, T5 y Llama, permite ejecutar inferencia, ajuste fino (fine-tuning) e implementación de modelos con apenas unas líneas de código. Su ecosistema incluye el Hugging Face Hub, un repositorio con más de 25 000 modelos preentrenados y 2 000 conjuntos de datos, que abarcan tareas como traducción automática, generación de texto, clasificación de imágenes, reconocimiento de voz y detección de objetos (Kelta, 2022). La biblioteca destaca por su soporte multiplataforma (PyTorch, TensorFlow, JAX) y su capacidad para manejar datos en más de 100 idiomas, incluyendo procesamiento de code-switching y lenguaje informal. Ofrece estructuras preconfiguradas para simplificar flujos de trabajo complejos desde la tokenización hasta la generación de resultados, encapsulando procesos como limpieza de texto, embebido y postprocesamiento (Wolf et al., 2020). Con herramientas como AutoModel y AutoTokenizer, los desarrolladores pueden cargar modelos específicos y adaptarlos a casos de uso personalizados sin requerir conocimientos profundos en redes neuronales (Gozlan, 2025). Su documentación detallada, integración con GPUs/TPUs y compatibilidad con formatos de despliegue la convierten en una solución industrialmente robusta para proyectos académicos y comerciales.

La Figura 13 muestra cómo el transformador actúa como núcleo de un sistema complejo que involucra modelos, datos, y métricas. Su capacidad para procesar información de manera efectiva y su flexibilidad hacen que los transformadores sean una herramienta indispensable en el campo del aprendizaje automático. La interconexión con tokenizadores y conjuntos de datos resalta la importancia de una preparación adecuada y un soporte técnico robusto para maximizar el potencial de estos modelos avanzados.

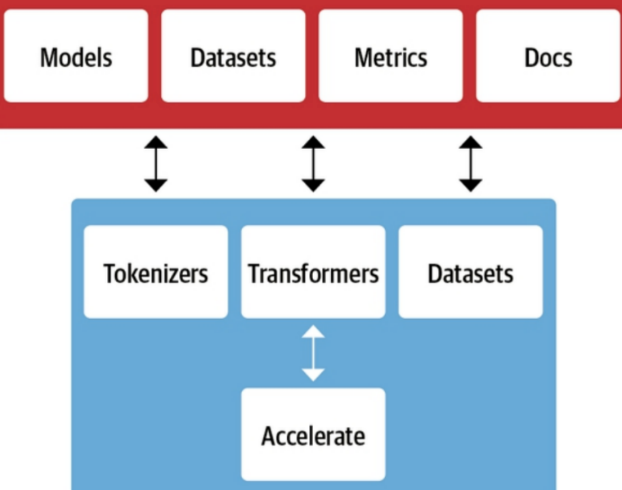


Figura 13 Interconexiones de herramientas en el workflow de HuggingFace

* + - 1. Pysentimiento

La librería pysentimiento se ha consolidado como una de las herramientas más versátiles para el análisis de sentimientos y tareas de PLN en contextos sociales. Desarrollada como un kit de herramientas Python de código abierto, combina modelos basados en transformers con técnicas especializadas de preprocesamiento para textos informales, particularmente eficaz en el análisis de redes sociales y contenido generado por usuarios. El núcleo de pysentimiento incluye un módulo de preprocesamiento específicamente diseñado para manejar las peculiaridades del lenguaje en redes sociales. Este sistema realiza transformaciones críticas como sustitución de menciones de usuario (@usuario) y URLs (url) por tokens especiales, normalización de risas repetidas ("jajajaja" a "jaja"), manejo de hashtags mediante separación de palabras compuestas (#UnaGenialidad a "una genialidad"), traducción de emojis a sus descripciones textuales (🎉 a "emoji party popper") (Pérez et al., 2024b).

Este preprocesamiento inteligente permite a los modelos transformer integrados en pysentimiento trabajar con texto social limpio y estandarizado, crucial para obtener resultados consistentes en análisis de sentimientos y detección de discurso de odio. La librería se apoya en arquitecturas transformer adaptadas al español, donde RoBERTuito actúa como base para tareas específicas (análisis de sentimientos, detección de emociones, etc.), demostrando su eficacia en el procesamiento de lenguaje coloquial y redes sociales (Pérez et al., 2024b). La integración con RoBERTuito representa un recurso clave para manejar la informalidad y peculiaridades del lenguaje en plataformas sociales como Twitter, está adaptado específicamente para:

1. Análisis de Sentimientos:
   1. Clasificación binaria (positivo/negativo) y multiclase.
2. Detección de Odio: Detecta tres categorías principales:
   1. Odio (hateful): Mensajes que promueven violencia o discriminación.
   2. Agresividad (aggressive): Lenguaje hostil o confrontativo.
   3. Dirigido (targeted): Contenido que ataca a individuos o grupos específico.
3. Detección de Emociones: Identificación de categorías como alegría, tristeza o enojo.
4. Reconocimiento de Entidades Nombradas: Extracción de personas, organizaciones y ubicaciones.
5. Etiquetado Gramatical: Identificar sustantivos, verbos, adjetivos y otros componentes morfosintácticos.

Frente a otras librerías de PLN, pysentimiento se distingue por la optimización para español, lo que incluye modelos preentrenados con grandes corpus hispanohablantes, preprocesamiento específico para jerga web y redes sociales manejando el lenguaje coloquial. El balance velocidad-precisión, que es una implementación eficiente que permite análisis en tiempo real y una interfaz unificada que cuenta de una API consistente para diferentes tareas de PLN. PySentimiento ya ofrece modelos optimizados que pueden integrarse directamente con APIs o Hugging Face, ver Figura 14.

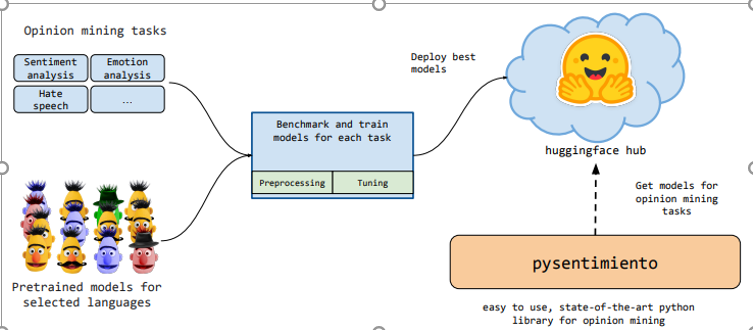


Figura 14 Flujo de trabajo para análisis de opiniones con Pysentimiento

Pysentimiento soporta varios idiomas para diferentes tareas de análisis y de PLN. Para tareas como reconocimiento de entidades nombradas (NER) y etiquetado gramatical (Hernandez Ucan, 2018), el soporte se limita a español e inglés, sin embargo, para las tareas principales como análisis de sentimiento, detección de discurso de odio, detección de ironía y análisis de emociones, pysentimiento soporta 4 idiomas; inglés (en), español (es), italiano (it) y portugués (pt) en los cuales muestra resultados muy buenos incluso al compararlo con otros modelos de PLN de gran popularidad en distintas bases de casos (Pérez et al., 2024a) como se puede observar en la Figura 15, y en la 16 con la detección de discurso de odio usando HateCheck de (Röttger et al., 2021).

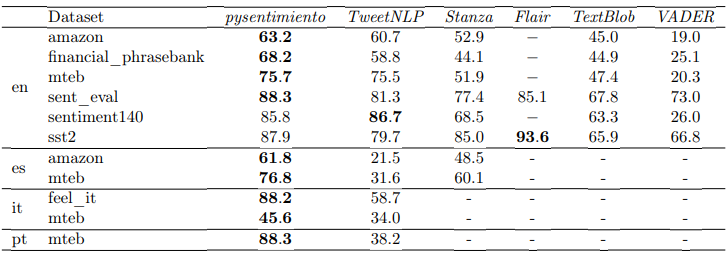


Figura 15 Comparación de pysentimiento con otras herramientas para el análisis de sentimientos. Los resultados se expresan mediante puntuaciones F1 Macro.

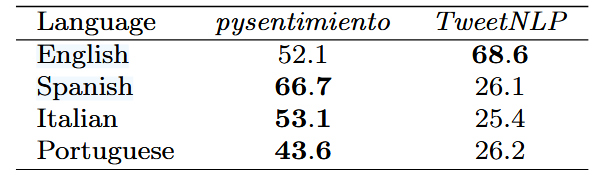


Figura 16 Camparación de pysentimiento con TweetNLP. Los resultados se expresan mediante puntuaciones F1 Macro.

* + 1. Análisis comparativo de herramientas

La selección de las herramientas computacionales apropiadas juega un papel crucial en este proceso. Atlas.ti además de ser un software de pago y la versión de prueba (5 días en un período de 20 días) tiene restricciones de tamaño para documentos y elementos de análisis, lo que afecta proyectos extensos (Friese, 2021). Brand24 permite rastrear y analizar en tiempo real las conversaciones en internet pero tiene ciertas limitaciones, el período de pruebas es solo de 14 días y el plan mensual más barato de la plataforma en estos momentos es de $199 USD y permite rastrear solo 2000 menciones, 3 palabras clave y solo un usuario (“Precios y Planes | Brand24,” 2025) y, aunque ofrece análisis de sentimientos basados en IA, algunos usuarios reportan necesidad de mejoras en la precisión, además de que tiene un acceso histórico limitado.

Los modelos de PLN ofrecen la capacidad de clasificar texto en diversas modalidades dependiendo de la tarea dada y muchos son de código abierto, libres de pago y con comunidades grandes y activas como la de Hugging Face, usarlos para la tarea de la detección de comportamiento transgresivo en redes sociales parece ser un enfoque correcto, TextBlob es muy popular y tiene buenos resultados en datasets ingleses pero al estar entrenado con texto enteramente en inglés no admite el análisis de texto en otros idiomas, NLTK es la librería más popular de Python para el procesamiento de lenguaje natural, tiene recursos e implementaciones muy útiles, como su propio módulo que permite cargar una lista de palabras vacías (Ganesan, 2023) y acceso completo a WordNet. También presenta a VADER para realizar análisis de sentimiento, pero, al igual que TextBlob, solo está entrenado para utilizarse en texto en inglés y aun así este no tiene muy buenos resultados clasificando texto en ese idioma, tampoco puede realizar otras tareas como la detección de odio. TweetNLP ofrece muy buenos resultados en la clasificación de texto en redes sociales en inglés, antes además proporcionaba una API capaz de sacar comentarios de Twitter de forma rápida y precisa, pero lastimosamente esta se quedó sin soporte desde la conversión de Twitter a X, además, en español, los resultados no son muy buenos y son superados con diferencia por los de pysentimiento, el cual implementa al modelo basado en Transformer RoBERTuito. Pysentimiento se alza sobre las demás librerías de PLN a la hora de procesar texto en español y específicamente el de las redes sociales, lo que la convierte en una herramienta muy útil para la detección de comportamiento transgresivo en línea mediante el uso de Transformer de Hugging Face.

* 1. Comportamiento Transgresivo en Línea

El comportamiento transgresivo en línea se refiere a las acciones realizadas en entornos digitales que violan normas sociales, éticas o legales establecidas. Este tipo de comportamiento puede incluir desde actos de desobediencia o desafío a la autoridad hasta actividades más graves como el acoso, la difusión de contenido inapropiado o ilegal, y la violación de derechos de privacidad. Estas conductas no son meras trasgresiones casuales, sino que reflejan dinámicas de poder, anonimato y desinhibición propias de los espacios digitales (Suler, 2004). Su impacto trasciende lo virtual, afectando la seguridad, la salud mental y los derechos fundamentales de individuos y comunidades.

* + 1. Tipos de comportamientos transgresivos en línea

Para detectar comportamientos transgresivos en línea en comentarios de redes sociales es importante saber clasificarlos para así ser capaces de tratarlos de manera correcta, según la gravedad del texto, asunto, personas involucradas, etc. Hay muchos indicadores que pueden proporcionar una visión amplia de los comportamientos transgresivos, como son los expuestos en los siguientes subepigrafes.

* + - 1. Violencia de género

En el universo interconectado de las plataformas digitales, la violencia de género adopta formas sutiles y complejas que trascienden las agresiones físicas. Bajo la apariencia de interacciones cotidianas, se esconden dinámicas de control, humillación y exclusión que refuerzan desigualdades estructurales. Desde comentarios aparentemente inofensivos hasta estrategias sistemáticas de desprestigio, estas manifestaciones operan en un espacio donde la identidad y el anonimato colisionan, dificultando su identificación y erradicación. Los discursos que naturalizan roles opresivos, la difusión de estereotipos dañinos y la invalidación de experiencias individuales son solo algunas de las caras de este fenómeno (Calderón Gómez et al., 2024; “Violencia de género,” 2025). En este contexto, para desarrollar estrategias efectivas de detección y prevención, y así avanzar hacia un entorno digital más seguro y equitativo es un paso crucial identificar y comprender las diversas formas en que se manifiesta la violencia de género en las redes sociales:

1. Acoso y control
   * Ejemplo: "¿Por qué no contestas mis mensajes? Te estoy viendo en línea. Dime con quién hablas ahora mismo."
2. Lenguaje ofensivo y denigrante:
   * Ejemplo: "Eres una #%$@ por subir esa foto. Deberías avergonzarte."
3. Amenazas y chantaje:
   * Ejemplo: "Si no me envías más fotos íntimas, publicaré las que ya tengo en todos lados."
4. Difamación y rumores:
   * Ejemplo: "¿Sabían que @usuario es la amante del profesor? Tengo pruebas."
5. Invasión de la privacidad:
   * Ejemplo: "Miren todos, aquí está la dirección y el número de teléfono de @usuario."
6. Presión para compartir contenido íntimo:
   * Ejemplo: "Si realmente me quieres, me enviarías una foto desnuda."
7. Censura digital:
   * Ejemplo: "Borra esa foto ahora mismo o terminaremos."
8. Suplantación de identidad:
   * Ejemplo: Crear un perfil falso con el nombre y foto de la víctima para publicar contenido dañino.
9. Normalización de conductas abusivas:
   * Ejemplo: "Todos los novios revisan el teléfono de sus parejas, es normal."
10. Acecho digital (stalking):
    * Ejemplo: "Sé que estuviste en el café ayer a las 3 PM. Vi tu ubicación."
11. Ciberacoso grupal y coordinado:
    * Ejemplo: Múltiples cuentas atacando simultáneamente: "Eres una vergüenza para la universidad, deberías irte."
12. Discurso de odio:
    * Ejemplo: "Las mujeres como tú no deberían tener derecho a opinar. Vuelve a la cocina."
13. Zoombombing:
    * Ejemplo: Interrumpir una videoconferencia académica con imágenes pornográficas no solicitadas.
14. Explotación sexual digital:
    * Ejemplo: "Si no haces un video erótico para mí, le diré a todos sobre tu orientación sexual."
15. Manipulación de algoritmos:
    * Ejemplo: Coordinar denuncias masivas falsas para que una cuenta sea bloqueada.
16. Cyberflashing:
    * Ejemplo: Enviar fotos de genitales no solicitadas en espacios públicos.
17. Doxing:
    * Ejemplo: "Aquí tienen la dirección de @usuario. Ya saben qué hacer."
18. Edición malintencionada de imágenes y videos:
    * Ejemplo: Crear y difundir un deepfake pornográfico de la víctima.
      + 1. Racismo

El racismo es la creencia en la existencia de razas superiores e inferiores de seres humanos. Puede formar parte de los prejuicios o convicciones personales de algunas personas o manifestarse en acciones políticas y sociales promovidas por partidos o gobiernos (Smedley, 2025). La base del racismo es la idea de que la humanidad se divide en razas. Esta idea es aceptada por algunos científicos, pero rechazada por otros, quienes prefieren hablar de grupos étnicos, pues consideran que los aspectos físicos no marcan diferencias significativas entre las personas, es decir, que no existen diferentes razas humanas, y que son las formas culturales las que realmente diferencian a los grupos humanos. En cualquier caso, el racismo supone una valoración de las razas o grupos étnicos y, por lo tanto, conlleva distintas formas de preferencia, segregación o exclusión basadas en el color de la piel, las facciones de la cara, la identidad étnica o la procedencia cultural (Gayubas, 2025).

El racismo suele conducir a prácticas discriminatorias, como la otorgación de privilegios (sociales, económicos, legales, políticos) a un grupo étnico sobre otros, la separación, marginación o ataques violentos contra determinadas poblaciones, o la negativa e incluso la prohibición de asociarse con personas provenientes de otras etnias (Tomás, 2021). Este fenómeno es conocido como discriminación racial y forma parte de los delitos y crímenes de odio tipificados en numerosas convenciones internacionales que defienden la igualdad entre las personas (“International Convention on the Elimination of All Forms of Racial Discrimination,” 2025).

Existen distintas formas o tipos de racismo:

1. Racismo cultural (“Discriminación racial y cultural: ¿Qué es?,” 2022): Consiste en el rechazo o la denigración de las tradiciones culturales de una etnia considerada “inferior” o “peligrosa”, por oposición a la cultura propia, que es identificada como “superior” o “correcta”. Si bien lo que se rechaza son las prácticas culturales, se suelen equiparar estas al origen étnico o “racial” de las personas que las practican.
2. Racismo biológico (Gayubas, 2025): Sostiene que existen razas biológicamente superiores e inferiores, y que las superiores están destinadas a dominar a las inferiores. Además de identificar a algunos pueblos como inferiores, suele definir al mestizaje como una “mezcla” de razas que genera impureza y atenta contra la “pureza racial”.
3. Racismo institucional (Buraschi and Idáñez, 2021): Es la aplicación, por parte de las instituciones del Estado, de medidas que se basan en la discriminación racial. Esto se da, por ejemplo, cuando la justicia actúa de manera distinta según cuál es el color de piel del ciudadano, o cuando las prácticas racistas se normalizan dentro de la policía, como ocurre en ciertos estados de Estados Unidos. Una forma particular de racismo institucional fue el apartheid en Sudáfrica, un sistema legal de segregación racial que favorecía a la “minoría blanca”.
4. Discriminación racial positiva (Arroyo, 2007): Consiste en promover políticas que favorezcan a individuos que pertenecen a una etnia frecuentemente discriminada, como compensación por el hecho de ser víctimas de discriminación o por otros motivos. El objetivo es facilitar su acceso a empleos, estudios o cargos públicos. Por ejemplo, cuando se establecen cupos para miembros de una determinada etnia en una universidad.
5. Racismo aversivo (Grande Yeves, 2022): Se trata de una forma de racismo y de xenofobia sutil. Suele manifestarse en la distancia o frialdad asumida frente a personas de otras etnias, o a la formación de prejuicios sin que estos motiven acciones violentas o agresivas.
6. Racismo oculto (Panchi et al., 2021): Es una forma de discriminación no explícita, que legitima el racismo de manera indirecta al no denunciar las prácticas discriminatorias, o que excluye de determinados ámbitos de sociabilidad, empleo o educación a personas pertenecientes a determinadas minorías étnicas sin hacerlo explícito.
   * + 1. Ciberacoso

El ciberacoso es el acoso que tiene lugar en dispositivos digitales, como teléfonos celulares, computadoras y tabletas. Puede ocurrir mediante mensajes de texto, textos y aplicaciones, o bien por Internet en las redes sociales, foros o juegos donde las personas pueden ver, participar o compartir contenido (“ciberacoso | Definición | Diccionario de la lengua española | RAE - ASALE,” 2024). Incluye enviar, publicar o compartir contenido negativo, perjudicial, falso, o cruel sobre otra persona, así como su información personal o privada, provocándole humillación o vergüenza. Algunos acosos por Internet pasan a ser un comportamiento ilegal o criminal. Los lugares más comunes donde ocurre el ciberacoso son redes sociales, como Facebook, Instagram, Snapchat y TikTok, mensajes de texto, aplicaciones de mensajería en dispositivos móviles y tabletas, mensajería instantánea, mensajes directos y chats por Internet, foros en Internet, salas de chat y tableros de mensajes, como Reddit, comunidades de juegos en Internet, etc (Vigderman, 2024).

El Suplemento sobre Delitos Escolares (Irwin et al., 2023) de la Encuesta Nacional de Víctimas del Delito (Centro Nacional de Estadísticas de Educación y la Oficina de Estadísticas Judiciales) indica que, a nivel nacional, entre los estudiantes de 12 a 18 años en los grados 6 a 12 que manifestaron haber sido víctimas de acoso ​​en la escuela el 19.2% de todos los estudiantes, aproximadamente el 21.6 % manifestó haber sido víctima de acoso en línea o por mensaje de texto, que incluye casi el doble de niñas (27.7 %) que de niños (14.1 %). Por otro lado, el (CDC, 2025) de los Centros para el Control y la Prevención de Enfermedades indica que aproximadamente el 16 % de los estudiantes de secundaria padecieron ciberacoso en los 12 meses previos a la encuesta.

En este sentido, resulta crucial analizar cómo las distintas formas de acoso se interrelacionan con otras problemáticas sociales, como el racismo o la desigualdad de género, para comprender su complejidad y diseñar estrategias integrales que promuevan entornos más seguros e inclusivos.

Los 5 tipos de ciberacoso más habituales son los siguientes:

1. Happy slapping (“El Happy Slapping – Asociación REA,” 2021): Traducido al español significa bofetada feliz y consiste en grabar una agresión física, verbal o sexual con la intención de difundirlo en internet. El objetivo del agresor se centra, en la mayoría de los casos, en ganar popularidad a costa de la humillación de su víctima.
2. Flameo o flaming (Arisanty and Wiradharma, 2022): Se centra en enviar mensajes ofensivos a una persona de manera masiva por una o varias personas. Puede originarse por una discrepancia de opiniones o cualquier otro pretexto y expone a la víctima a escarnio público.
3. Ciberacecho o stalking (Stevens et al., 2021): Consiste en espiar a una persona y realizar un seguimiento constante de sus acciones y movimientos en redes sociales o cualquier medio digital. Lo que se pretende es conseguir el mayor tipo de información de la persona objeto del espionaje con intención aviesa.
4. Sexting a menores (Colmenares-Guillen et al., 2024): Este ciberacoso es una extorsión a víctimas que han compartido previamente conversaciones, fotografías o vídeos de carácter íntimo, erótico y/o sexual. El acosador obliga a su víctima a enviar más contenido bajo la amenaza de difundir el material en redes para hacerlo público. Puede darse el caso de que la finalidad última sea hacer llegar ese contenido a terceras personas sin que el menor esté al corriente. En este contexto estaríamos hablando de sexting a menores sin consentimiento.
   * + 1. Discurso de odio

Los discursos de odio, como manifestaciones verbales o escritas de intolerancia y hostilidad, representan una amenaza insidiosa para la cohesión social y la dignidad humana. Estas expresiones, que a menudo se basan en prejuicios raciales, religiosos, de género o identidad, buscan deshumanizar y marginar a grupos específicos, alimentando un clima de miedo y división (“Discurso de odio,” 2025). A través de plataformas digitales, medios de comunicación y espacios públicos, los discursos de odio pueden difundirse rápidamente, incitando a la violencia y erosionando los valores de respeto e inclusión que sustentan las sociedades democráticas.

En el contenido de sus mensajes existe el uso de lenguaje ofensivo, insultos o epítetos raciales, religiosos, etc., promociona estereotipos negativos o deshumanizantes, llamados directos o indirectos a la violencia o exclusión de un grupo. En la sociedad actual, existen diferentes tipos de discursos de odio que pueden manifestarse de diversas formas (Ya, 2025). A continuación, se describen algunos de los más comunes:

1. Discriminación racial (García Cabezas, 2023): Este tipo de discurso se basa en el color de piel, origen étnico o nacionalidad de las personas. Puede incluir comentarios despectivos, estereotipos o exaltación de una supuesta superioridad racial.
2. Discriminación religiosa (Mosqueira, 2022): Se define como el trato desfavorable hacia personas o grupos por sus creencias, prácticas o afiliaciones religiosas, incluso en sociedades que garantizan constitucionalmente la libertad de culto
3. Discriminación por orientación sexual (Molano, 2023): Este tipo de discurso se dirige a personas LGBTQ+ y puede incluir insultos homofóbicos, transfóbicos o cualquier forma de desprecio hacia la diversidad sexual.
4. Xenofobia (Lazzarini and Leite, 2022): Se manifiesta con actitudes hostiles hacia personas extranjeras o inmigrantes. Se promueve el rechazo hacia aquellos que son percibidos como diferentes por su nacionalidad o cultura.
5. Discurso sexista (Grassi, 2018): Aquí se menosprecia a personas con base en su género. Se pueden propagar estereotipos dañinos sobre las capacidades y roles asignados a hombres y mujeres.
   1. Conclusiones parciales

El procesamiento del lenguaje natural es indispensable para la detección de los diferentes tipos de conductas transgresivas en línea. Su eficacia en la detección y análisis de estos comportamientos en redes sociales depende significativamente del modelo lingüístico empleado. La elección de un modelo adecuado como RoBERTuito, optimizado para el idioma español y entrenado con datos relevantes, permite una mayor precisión en la identificación de matices lingüísticos, contextos específicos y expresiones sutiles que caracterizan estos comportamientos.

Es fundamental destacar la diversidad de comportamientos transgresivos que pueden ser analizados mediante el PLN, la identificación y clasificación de estos comportamientos es esencial para su tratamiento adecuado y para el desarrollo de estrategias de prevención efectivas. La selección de la herramienta computacional Pysentimiento con el uso de Transformer de Hugging Face juega un papel crucial para el procesamiento del lenguaje natural en la detección del comportamiento transgresivo en linea.

**REFERENCIAS BIBLIOGRÁFICAS**

Amalou, I., Mouhni, N., Abdali, A., 2022. Multivariate time series prediction by RNN architectures for energy consumption forecasting. Energy Rep. 8, 1084–1091. https://doi.org/10.1016/j.egyr.2022.07.139

Amnistía Internacional [WWW Document], 2024. . Amnistía Int. URL https://www.amnesty.org/es/ (accessed 4.3.25).

Anders, V., 2025. TRANSGRESIÓN, radicación [WWW Document]. Etimologías Chile - Dicc. Que Explica El Origen Las Palabras. URL https://etimologias.dechile.net/?transgresio.n (accessed 4.2.25).

Anishnama, 2023. Understanding Gated Recurrent Unit (GRU) in Deep Learning. Medium. URL https://medium.com/@anishnama20/understanding-gated-recurrent-unit-gru-in-deep-learning-2e54923f3e2 (accessed 4.15.25).

Arisanty, M., Wiradharma, G., 2022. The motivation of flaming perpetrators as cyberbullying behavior in social media. J. Kaji. Komun. 10, 215. https://doi.org/10.24198/jkk.v10i2.39876

Arras, L., Montavon, G., Müller, K.-R., Samek, W., 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. https://doi.org/10.48550/arXiv.1706.07206

Arroyo, J.C.V., 2007. Discriminación positiva, diversidad cultural y justicia.

Aryan, A., 2023. Introduction to Large Language Models [WWW Document]. URL https://abiaryan.com/posts/intro-llms/ (accessed 4.18.25).

AWS, 2024. ¿Qué son los modelos fundacionales? [WWW Document]. Amaz. Web Serv. Inc. URL https://aws.amazon.com/es/what-is/foundation-models/ (accessed 4.6.25).

Barnard, J., 2023. What Are Word Embeddings? | IBM [WWW Document]. URL https://www.ibm.com/think/topics/word-embeddings (accessed 4.19.25).

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M., 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, in: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M. (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation. Presented at the SemEval 2019, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 54–63. https://doi.org/10.18653/v1/S19-2007

Bergmann, D., 2024. ¿Qué es el aprendizaje autosupervisado? | IBM [WWW Document]. URL https://www.ibm.com/es-es/topics/self-supervised-learning (accessed 4.7.25).

Bird, S., Klein, E., Loper, E., 2009a. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.

Bird, S., Klein, E., Loper, E., 2009b. NLTK Book [WWW Document]. URL https://www.nltk.org/book/ (accessed 4.20.25).

Buendía-Rodríguez, E., Vargas-Pérez, E., Leyva-Ovalle, Á., Terrazas-Domínguez, S., 2002. APLICACIÓN DE REDES NEURONALES ARTIFICIALES Y TÉCNICAS SIG PARA LA PREDICCIÓN DE COBERTURAS FORESTALES. Rev. Chapingo Ser. Cienc. For. Ambiente.

Buraschi, D., Idáñez, M.J.A., 2021. Racismo institucional: ¿De qué estamos hablando? [WWW Document]. The Conversation. URL http://theconversation.com/racismo-institucional-de-que-estamos-hablando-157152 (accessed 4.7.25).

Calderón Gómez, D., Puente Bienvenido, H., García Mingo, E., 2024. Generación expuesta: jóvenes frente a la violencia sexual digital.

Camacho Costales, A., 2021. El impacto de una nueva Internet en la esfera pública cubana | Cuba Capacity Building Project [WWW Document]. URL https://horizontecubano.law.columbia.edu/news/el-impacto-de-una-nueva (accessed 4.7.25).

Camacho-collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., Boisson, J., Espinosa Anke, L., Liu, F., Martínez Cámara, E., 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media, in: Che, W., Shutova, E. (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Abu Dhabi, UAE, pp. 38–49. https://doi.org/10.18653/v1/2022.emnlp-demos.5

Cañete, J., 2025. dccuchile/beto.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., Pérez, J., 2023. Spanish Pre-trained BERT Model and Evaluation Data. https://doi.org/10.48550/arXiv.2308.02976

Canorea, E., 2022. ¿Qué es la tokenización? | Claves para entenderlo [WWW Document]. Plain Concepts. URL https://www.plainconcepts.com/es/tokenizacion/ (accessed 4.6.25).

Carranza Coello, J.J., Carpio Ramírez, C.A., 2019. TRANSGRESIÓN DE NORMAS SOCIALES: EFECTOS DE LA HISTORIA DE VALORACIÓN. Rev. Electrónica Psicol. Iztacala 22.

CDC, 2025. YRBS Data Summary & Trends Report. Youth Risk Behav. Surveill. Syst. YRBSS.

ciberacoso | Definición | Diccionario de la lengua española | RAE - ASALE [WWW Document], 2024. URL https://dle.rae.es/ciberacoso (accessed 4.14.25).

Código de Familia [WWW Document], 2022. URL https://www.parlamentocubano.gob.cu/sites/default/files/documento/2022-09/goc-2022-o99.pdf (accessed 4.3.25).

Código Penal, 1987.

Colmenares-Guillen, L.E., Barrera-Márquez, R.A., Cerón-Garnica, C., 2024. Análisis del sexting como factor de riesgo en la difusión de pornografía infantil. CienciAmérica 13, 103–118. https://doi.org/10.33210/ca.v13i1.456

Common Crawl, 2024. . Wikipedia Encicl. Libre.

Cómo aplicar el procesamiento del lenguaje natural a la ciberseguridad [WWW Document], 2023. . Espanol News. URL https://espanol.news/como-aplicar-el-procesamiento-del-lenguaje-natural-a-la-ciberseguridad/ (accessed 4.6.25).

Cota, S., 2023. Deep Learning Basics — Part 7 — Feed Forward Neural Networks (FFNN) | by Sasirekha Cota | Medium [WWW Document]. URL https://medium.com/@sasirekharameshkumar/deep-learning-basics-part-10-feed-forward-neural-networks-ffnn-93a708f84a31 (accessed 4.16.25).

Cuba 2019 Constitución [WWW Document], 2019. URL https://www.constituteproject.org/constitution/Cuba\_2019?lang=es (accessed 4.3.25).

Daniel Llamas, 2019. Morphological Analysis [WWW Document]. URL https://medium-com.translate.goog/lastbasic/morphological-analysis-67e9eb57aee0?\_x\_tr\_sl=en&\_x\_tr\_tl=es&\_x\_tr\_hl=es&\_x\_tr\_pto=tc (accessed 4.20.25).

de la Torre Lorente, J.L., 2024. Los transformers en la inteligencia artificial: una explicación sencilla [WWW Document]. delatorre.ai. URL https://delatorre.ai/los-transformers-en-la-inteligencia-artificial-una-explicacion-sencilla/ (accessed 4.16.25).

Delgado, J.A., 2024. La importancia del Natural Language Processing en la IA [WWW Document]. KSchool Tu Cent. Estud. Espec. En Internet. URL https://kschool.com/blog/big-data/importancia-del-natural-language-processing-en-la-ia/ (accessed 4.6.25).

Devarakonda, K., A van Breen, J., Liem, M., 2020. The role of the Internet in transgressive behaviour: New tool or a new phenomenon? [WWW Document]. URL https://www.leidensecurityandglobalaffairs.nl/articles/the-role-of-the-internet-in-transgressive-behaviour-new-tool-or-a-new-phenomenon (accessed 4.3.25).

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Burstein, J., Doran, C., Solorio, T. (Eds.), . Presented at the NAACL-HLT 2019, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

Discriminación racial y cultural: ¿Qué es? [WWW Document], 2022. URL https://accioncontraelhambre.org/es/actualidad/discriminacion-racial-cultural-que-es (accessed 4.7.25).

Discurso de odio [WWW Document], 2025. URL https://es.wikipedia.org/wiki/Discurso\_de\_odio (accessed 4.7.25).

EDVAW Platform [WWW Document], 2022. . OHCHR. URL https://www.ohchr.org/en/edvaw (accessed 4.3.25).

El Happy Slapping – Asociación REA, 2021. URL https://www.asociacionrea.org/el-happy-slapping/ (accessed 4.14.25).

Erick, R.M., 2024. Procesamiento de Lenguaje Natural con Transformers: BERT y GPT [WWW Document]. URL https://lovtechnology.com/procesamiento-de-lenguaje-natural-con-transformers-bert-y-gpt/#google\_vignette (accessed 4.6.25).

Esuli, A., Sebastiani, F., 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining.

Ferreyra (comp), D., 2018. synset. Léxico de lingüística [WWW Document]. https://vocabularyserver.com/linguistica/. URL https://vocabularyserver.com/linguistica/?tema=1465 (accessed 4.20.25).

Friese, S., 2021. ATLAS.ti 9 Mac - Quick Tour.

Ganesan, K., 2023. What are Stop Words? [WWW Document]. URL https://kavita-ganesan.com/what-are-stop-words/ (accessed 4.22.25).

García Cabezas, N., 2023. Discriminación racial: definición y causas - Ayuda en Acción [WWW Document]. URL https://ayudaenaccion.org/blog/derechos-humanos/causas-discriminacion-racial/ (accessed 4.16.25).

García Cumbreras, M.Á., Julio Gonzalo, Martínez Cámara, E., Martínez Unanue, R., Rosso, P., Carrillo-de-Albornoz, J., Soto Montalvo, Chiruzzo, L., Collovini, S., Guitiérrez, Y., Jiménez Zafra, S., Krallinger, M., Montes-y-Gómez, M., Ortega-Bueno, R., Rosá, A., 2019. Preface on the Iberian Languages Evaluation Forum (IberLEF 2019) [WWW Document]. URL https://ceur-ws.org/Vol-2421/preface.pdf (accessed 4.23.25).

García Vega, M., Díaz Galiano, M.C., García Cumbreras, M.Á., 2020a. Overview of TASS 2020: Introducing Emotion Detection.

García Vega, M., Díaz Galiano, M.C., García Cumbreras, M.Á., Plaza del Arco, F.M., Montejo Ráez, A., Jiménez Zafra, S.M., Martínez Cámara, E., Aguilar, C.A., Murillo Casasola, E., Sobrevilla Cabezudo, M.A., Chiruzzo, L., A. Moctezuma, D., 2020b. TASS 2020 Workshop on Semantic Analysis at SEPLN 2020. URL http://tass.sepln.org/2020/ (accessed 4.23.25).

Gayubas, A., 2025. Racismo - Concepto, tipos, discriminación y prejuicios. https://concepto.de/.

Geetha, L., 2023. Vader: A Comprehensive Guide to Sentiment Analysis in Python. Medium. URL https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e (accessed 4.20.25).

Google Bert, 2024. Bert Base Multilingual Cased [WWW Document]. Bert Base Multiling. Cased. URL https://dataloop.ai/library/model/google-bert\_bert-base-multilingual-cased/ (accessed 4.6.25).

Gozlan, Y., 2025. Using transformers at Hugging Face [WWW Document]. URL https://huggingface.co/docs/hub/en/transformers (accessed 4.20.25).

Grande Yeves, P., 2022. ¿Qué es el racismo aversivo? - La Mente es Maravillosa [WWW Document]. URL https://lamenteesmaravillosa.com/que-es-el-racismo-aversivo/ (accessed 4.13.25).

Grassi, M.P.L., 2018. Definición del concepto de “sexismo”: influencia en el lenguaje, la educación y la violencia de género.

Hamad, R., 2023. What is LSTM? Introduction to Long Short-Term Memory [WWW Document]. URL https://medium.com/@rebeen.jaff/what-is-lstm-introduction-to-long-short-term-memory-66bd3855b9ce (accessed 4.15.25).

Haponik, A., 2023. What is Fine Tuning in NLP? Addepto. URL https://addepto.com/blog/what-is-fine-tuning-in-nlp/ (accessed 4.6.25).

Haskel-Ittah, M., 2023. Explanatory black boxes and mechanistic reasoning. J. Res. Sci. Teach. 60, 915–933. https://doi.org/10.1002/tea.21817

Hernandez Ucan, R., 2018. Etiquetado gramatical [WWW Document]. URL https://medium.com/soldai/etiquetado-gramatical-a418278e115c (accessed 4.22.25).

Hugging Face, 2025. . Wikipedia.

Husnain, M., Missen, M.M.S., Akhtar, N., Coustaty, M., Mumtaz, S., Prasath, V.B.S., 2021. A systematic study on the role of SentiWordNet in opinion mining. Front. Comput. Sci. 15, 154614. https://doi.org/10.1007/s11704-019-9094-0

Iglesias Álvarez, I., 2024. Procesamiento del lenguaje natural como parapeto del abuso y la violencia | CIO [WWW Document]. URL https://www.cio.com/article/2513277/procesamiento-del-lenguaje-natural-como-parapeto-del-abuso-y-la-violencia.html (accessed 4.6.25).

International Convention on the Elimination of All Forms of Racial Discrimination, 2025. . Wikipedia.

Irwin, V., Wang, K., Cui, J., Thompson, A., 2023. Report on Indicators of School Crime and Safety: 2023.

Ivezic, M., Ivezic, L., 2022. How Dynamic Data Masking Reinforces Machine Learning Security [WWW Document]. Secur. - Mar. Ivezic. URL https://securing.ai/ai-security/dynamic-data-masking-ml/ (accessed 4.6.25).

Jimenez Melo, D., 2023. Transformers [WWW Document]. Medium. URL https://medium.com/@djm9826/transformers-ab2099ce116e (accessed 4.6.25).

Josep, 2023. PLN: todo lo que necesitas saber sobre procesamiento de lenguaje natural [WWW Document]. Conectando Ideas. URL https://conectandoideas.net/pln-procesamiento-de-lenguaje-natural/ (accessed 4.6.25).

Justicia, J.M., 2003. Análisis cualitativo de datos textuales con ATLAS/ti.

Kayid, A., Reimers, N., 2022. Cohere’s Multilingual Text Understanding Model is Now Available [WWW Document]. URL https://cohere.com/es/blog/multilingual (accessed 4.5.25).

Kelta, Z., 2022. An Introduction to Using Transformers and Hugging Face [WWW Document]. URL https://www.datacamp.com/tutorial/an-introduction-to-using-transformers-and-hugging-face (accessed 4.20.25).

Lazzarini, A.C., Leite, P., 2022. Comunicación sin xenofobia.

Leal, D.A., Restrepo, Porto, J.P., Viloria, Algarín, C.A., Robles, 2021. El camino a las redes neuronales artificiales. Editorial Unimagdalena.

Ley de Comunicación [WWW Document], 2023. URL https://www.parlamentocubano.gob.cu/sites/default/files/documento/2023-05/proyecto-ley-de-comunicacion-social\_mayo-2023\_0.pdf (accessed 4.3.25).

Li, C., 2020. OpenAI’s GPT-3 Language Model: A Technical Overview [WWW Document]. URL https://lambda.ai/blog/demystifying-gpt-3 (accessed 4.5.25).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692

Loria, S., 2025a. sloria/TextBlob.

Loria, S., 2025b. TextBlob: Simplified Text Processing — TextBlob 0.19.0 documentation [WWW Document]. URL https://textblob.readthedocs.io/en/dev/ (accessed 2.21.25).

Malingan, N., 2023. Next Sentence Prediction with BERT [WWW Document]. Scaler Top. URL https://www.scaler.com/topics/nlp/bert-next-sentence-prediction/ (accessed 4.6.25).

Metzger, S., 2022. What are Tokens, Vectors, and Embeddings & How do you create them? Medium. URL https://medium.com/@saschametzger/what-are-tokens-vectors-and-embeddings-how-do-you-create-them-e2a3e698e037 (accessed 4.6.25).

Modelo autorregresivo [WWW Document], 2024. URL https://es.wikipedia.org/wiki/Modelo\_autorregresivo (accessed 4.6.25).

Molano, C., 2023. ¿Qué es la discriminación de género? [WWW Document]. Aldeas Infant. Colomb. URL https://www.aldeasinfantiles.org.co/noticias/noticias-2023/discriminacion-de-genero (accessed 4.16.25).

Molero, J.M., Pérez-Martín, J., Rodrigo, A., Peñas, A., 2023. Offensive Language Detection in Spanish Social Media: Testing From Bag-of-Words to Transformers Models. IEEE Access 11, 95639–95652. https://doi.org/10.1109/ACCESS.2023.3310244

Molina, F.S., 2022. La Inteligencia Artificial y su incidencia en las comunicaciones.

Moré, J., 2019. Extracción de sentimientos y opiniones [WWW Document]. URL https://openaccess.uoc.edu/bitstream/10609/148645/2/Modulo2\_ExtraccionDeSentimientosYOpiniones.pdf (accessed 4.19.25).

Mosqueira, M., 2022. Editorial: Discriminación por motivos religiosos, una agenda de investigación necesaria. Cult. Religión 16, i–v. https://doi.org/10.4067/S0718-47272022000200001

Moviefone: find it. watch it. [WWW Document], 2025. . Moviefone. URL https://www.moviefone.com/ (accessed 4.6.25).

Multilingual detection of hate speech against immigrants and women in Twitter (hatEval) [WWW Document], 2019. URL https://competitions.codalab.org/competitions/19935 (accessed 4.7.25).

Murel Ph.D., J., Kavlakoglu, E., 2024. What are masked language models? | IBM [WWW Document]. URL https://www.ibm.com/think/topics/masked-language-model (accessed 4.5.25).

Narein T, A., 2021. Embeddings in BERT [WWW Document]. OpenGenus IQ Learn Algorithms DL Syst. Des. URL https://iq.opengenus.org/embeddings-in-bert/ (accessed 4.6.25).

Nosouhian, S., Nosouhian, F., Kazemi Khoshouei, A., 2021. A Review of Recurrent Neural Network Architecture for Sequence Learning: Comparison between LSTM and GRU. https://doi.org/10.20944/preprints202107.0252.v1

OHCHR [WWW Document], 2025. . OHCHR. URL https://www.ohchr.org/es/hrbodies/hrc/home (accessed 4.3.25).

OPUS Home - Open and Universal Science (OPUS) Project, 2022. URL https://opusproject.eu/ (accessed 4.6.25).

Ortega-Bueno, R., Rangel, F., Farıas, D.I.H., Rosso, P., Montes-y-Gomez, M., Medina-Pagola, J.E., 2019. Overview of the Task on Irony Detection in Spanish Variants.

Padnos, D., 2023. AI21 Labs acelera la adopción de modelos de IA generativa con Amazon SageMaker [WWW Document]. URL https://aws.amazon.com/es/solutions/case-studies/AI21Labs-case-study/ (accessed 4.5.25).

Panchi, N.G., Hoz, H.B.D.L., Olivero, D.E., 2021. Racismo ambiental | Revista Mundo Financiero.

Pauli, P.A., 2019. Análisis de sentimiento Comparación de algoritmos predictivos y métodos utilizando un lexicon español.

Pavesi, E., 2020. Comportamientos transgresivos de los jóvenes y la falta de templanza. URL https://www.fiamc.org/bioethics/comportamientos-transgresivos-de-los-jovenes-y-la-falta-de-templanza/ (accessed 4.3.25).

Pérez, J.M., Furman, D.A., Alemany, L.A., Luque, F., 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. https://doi.org/10.48550/arXiv.2111.09453

Pérez, J.M., Rajngewerc, M., Giudici, J.C., Furman, D.A., Luque, F., Alemany, L.A., Martínez, M.V., 2024a. pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks. https://doi.org/10.48550/arXiv.2106.09462

Pérez, J.M., Rajngewerc, M., Mariela Rajngewerc, Giudici, J.C., Furman, D.A., Damián A. Furman, Luque, F., Alemany, L.A., Martínez, M.V., 2024b. pysentimiento: A Transformer-based library for SocialNLP tasks.

Petriconi, L., 2022. How to Do Twitter Sentiment Analysis with a Pre-Trained Language Model [Python]. Medium. URL https://medium.com/@lucapetriconi/how-to-do-twitter-sentiment-analysis-with-a-pre-trained-language-model-87614635c40e (accessed 4.20.25).

Plaza-del-Arco, F.M., Casavantes, M., Jair Escalante, H., Martín-Valdivia, M.T., Montejo-Ráez, A., Montes-y-Gómez, M., Jarquín-Vásquez, H., Villaseñor-Pineda, L., 2021. Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants. Proces. Leng. Nat. 183–194. https://doi.org/10.26342/2021-67-16

Precios y Planes | Brand24 [WWW Document], 2025. URL https://brand24.com/es/precios/ (accessed 4.22.25).

Preetham, F., 2024. Math Behind Positional Embeddings in Transformer Models. Auton. Agents. URL https://medium.com/autonomous-agents/math-behind-positional-embeddings-in-transformer-models-921db18b0c28 (accessed 4.6.25).

Princeton University, 2010. Princeton University “About WordNet.” [WWW Document]. URL https://wordnet.princeton.edu/ (accessed 2.21.25).

Raman, 2021. Introduction to BERT and Segment Embeddings. Anal. Vidhya. URL https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-bert/ (accessed 4.6.25).

Ramanathan, T., 2025. Natural language processing (NLP) | Definition, History, & Facts | Britannica [WWW Document]. URL https://www.britannica.com/technology/natural-language-processing-computer-science (accessed 4.6.25).

Recurrent neural network [WWW Document], 2024. URL https://en.wikipedia.org/wiki/Recurrent\_neural\_network (accessed 4.5.25).

Rodrigo, R., 2025a. Análisis Semántico: Concepto, Características y Ejemplos | Estudyando [WWW Document]. URL https://estudyando.com/analisis-semantico-concepto-caracteristicas-y-ejemplos/ (accessed 4.6.25).

Rodrigo, R., 2025b. Generación de Lenguaje Natural: Características y Ejemplos. Estudyando. URL https://estudyando.com/generacion-de-lenguaje-natural-caracteristicas-y-ejemplos/ (accessed 4.6.25).

Roman, V., 2019. Algoritmos Naive Bayes: Fudamentos e Implementación. Cienc. Datos. URL https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f (accessed 4.20.25).

Rosa, J. de la, Ponferrada, E.G., Villegas, P., Salas, P.G. de P., Romero, M., Grandury, M., 2022. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. https://doi.org/10.48550/arXiv.2207.06814

Sanz, F., 2020. Modelos de secuencia. Aprende TODO (RNN, LSTM, GRU BERT) [WWW Document]. URL https://www.themachinelearners.com/modelos-secuencia/ (accessed 4.6.25).

SCILabs, 2025. La ciberseguridad y el uso de Inteligencia Artificial de Procesamiento de Lenguaje Natural (NLP AI). Resour. Scitum. URL https://resources.scitum.com.mx/la-ciberseguridad-y-el-uso-de-inteligencia-artificial-de-procesamiento-de-lenguaje-natural-nlp-ai/ (accessed 4.6.25).

Servín-Palestina, M., Salazar-Moreno, R., López-Cruz, I., Medina-García, G., Cid-Ríos, J.Á., 2022. Predicción de la producción y rendimiento de frijol, con modelos de redes neuronales artificiales y datos climáticos. Biotecnia 24, 104–111. https://doi.org/10.18633/biotecnia.v24i2.1664

Shahid, N., Rappon, T., Berta, W., 2019. Applications of artificial neural networks in health care organizational decision-making: A scoping review. PLoS ONE 14, e0212356. https://doi.org/10.1371/journal.pone.0212356

Smedley, A., 2025. Racism | Definition, History, Laws, & Facts | Britannica [WWW Document]. URL https://www.britannica.com/topic/racism (accessed 4.7.25).

Statistiek, C.B. voor de, 2020. Veiligheidsmonitor 2019 [WWW Document]. Cent. Bur. Voor Stat. https://doi.org/10/veiligheidsmonitor-2019

Stevens, F., Nurse, J.R.C., Arief, B., 2021. Cyber Stalking, Cyber Harassment, and Adult Mental Health: A Systematic Review. Cyberpsychology Behav. Soc. Netw. 24, 367–376. https://doi.org/10.1089/cyber.2020.0253

Stryker, C., 2021. ¿Qué es una red neuronal recurrente (RNN)? [WWW Document]. URL https://www.ibm.com/es-es/think/topics/recurrent-neural-networks (accessed 4.18.25).

Stryker, C., Holdsworth, J., 2024. ¿Qué es el PLN (procesamiento del lenguaje natural)? | IBM [WWW Document]. URL https://www.ibm.com/es-es/think/topics/natural-language-processing (accessed 4.6.25).

Suler, J., 2004. The online disinhibition effect.

Syntactic Analysis in Natural Language Processing [WWW Document], 2025. URL https://www.tutorialspoint.com/natural\_language\_processing/natural\_language\_processing\_syntactic\_analysis.htm (accessed 4.6.25).

Tackling Online Transgressive Behaviours- Strengthening Institutional Policies on Digital Safety in Cuban University Community, 2024.

Tirado Picado, V.R., 2024. Redes Neuronales Artificiales como Modelo de Prediccion de los Factores Climaticos en Nicaragua en el Periodo 2021-2022. Cienc. Lat. Rev. Científica Multidiscip. 8, 1458–1474. https://doi.org/10.37811/cl\_rcm.v8i1.9541

Tomás, A., 2021. Racismo, subalternización política y derechos negados en América Latina.

Transformador (modelo de aprendizaje automático), 2025. . Wikipedia Encicl. Libre.

Tulkens, S., 2020. clips/pattern.

UCLV, 2025. Universidad Central “Marta Abreu” de Las Villas [WWW Document]. URL https://www.uclv.edu.cu/ (accessed 4.7.25).

UNDP [WWW Document], 2025. . UNDP. URL https://www.undp.org/es/home (accessed 4.3.25).

Vasiliev, Y., 2020. Natural Language Processing with Python and spaCy: A Practical Introduction. No Starch Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. ukasz, Polosukhin, I., 2017. Attention is All you Need, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Vigderman, liza, 2024. Cyberbullying: Twenty Crucial Statistics for 2025 | Security.org [WWW Document]. URL https://www.security.org/resources/cyberbullying-facts-statistics/ (accessed 4.14.25).

Violencia de género, 2025. . Wikipedia Encicl. Libre.

Wang, C., Cho, K., Gu, J., 2019. Neural Machine Translation with Byte-Level Subwords. https://doi.org/10.48550/arXiv.1909.03341

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

Wydmanski, W., 2022. Autoatención vs. atención en Transformers [WWW Document]. URL https://medium.com/@wwydmanski/whats-the-difference-between-self-attention-and-attention-in-transformer-architecture-3780404382f3 (accessed 4.15.25).

Ya, V., 2025. Discursos de odio: significado, clasificación y estrategias para su erradicación. VenceYa.

Yadav, A., 2024. Word2Vec vs GloVe: Which Word Embedding Model is Right for You? [WWW Document]. URL https://medium.com/biased-algorithms/word2vec-vs-glove-which-word-embedding-model-is-right-for-you-4dfc161c3f0c (accessed 4.19.25).

Yeolekar, A., 2024. Whole Word Masking [WWW Document]. URL https://www.atharvyeoelekar.blog/post/whole-word-masking (accessed 4.6.25).