

¿OFENSIVO O NO? LA DETECCIÓN AUTOMÁTICA DE MISOGINIA EN TWITTER MEDIANTE UN MODELO NEURONAL MULTILINGÜE

UN ESTUDIO COMPARATIVO ENTRE EL NEERLANDÉS, EL
INGLÉS Y EL ESPAÑOL

Palabras: 14983

Jirky Dumon

Número estudiantil: 02208146

Directora de tesis: Prof. Dr. Els Lefever

Codirector de tesis: Prof. Dr. Patrick Goethals

Trabajo de fin de máster presentado para obtener el título de Máster en Interpretación

Año académico: 2025 – 2026

Agradecimientos

En primer lugar, deseo expresar mi más sincero agradecimiento a la profesora Lefever, al profesor Goethals y al Dr. Maladry por su orientación, sus comentarios y sus palabras de ánimo a lo largo de este proceso. Sus consejos han sido de gran ayuda para llevar a cabo este trabajo.

Asimismo, quiero dar las gracias a mi familia, a mis amigos y a mi novio por estar siempre a mi lado, tanto en los buenos como en los malos momentos, y por ofrecerme su apoyo incondicional durante todo este periodo. Os quiero mucho a todos. Sin vosotros no lo habría conseguido.

Resumen

Considerando que la cantidad de mensajes publicados diariamente en línea excede la capacidad de revisión manual, resulta imprescindible desarrollar métodos automatizados para la detección eficiente de dichos contenidos mediante técnicas de procesamiento del lenguaje natural (PLN). Sin embargo, estos sistemas continúan enfrentándose a una serie de desafíos, entre los cuales destacan la escasa disponibilidad de datos de entrenamiento en lenguas distintas del inglés, la dificultad para reconocer el lenguaje implícito y la limitada capacidad de los investigadores para identificar el origen de los errores, debido a la falta de transparencia en el proceso de razonamiento de los sistemas. Con el fin de responder, al menos parcialmente, a estas limitaciones, en el presente estudio se ajusta el modelo multilingüe de tipo encoder mmBert para la tarea de detección automática del discurso de odio misógino en datos procedentes de Twitter en tres lenguas. El objetivo es responder a dos preguntas de investigación: ¿en qué medida permite un modelo lingüístico multilingüe detectar el discurso de odio misógino en tuits redactados en neerlandés, inglés y español?, y ¿se observan diferencias entre estas tres lenguas en los puntos fuertes y débiles de la detección automática al considerar categorías discursivas como la implicidad, la agresividad, los estereotipos y el humor? Los resultados muestran que el sistema obtiene el mejor rendimiento en inglés, seguido muy de cerca por el español y, finalmente, por el neerlandés, que presenta una puntuación F1 considerablemente más baja. En cuanto a las categorías discursivas analizadas, estas parecen plantear un grado de dificultad relativamente comparable en las tres lenguas. Las principales excepciones se observan en los tuits implícitos en español y en los tuits humorísticos en neerlandés, categorías en las que el sistema muestra mayores dificultades. Todo lo anterior nos lleva a afirmar que resulta necesario proseguir con la investigación a fin de desarrollar un sistema óptimo.

< 304 palabras >

Tabla de contenidos

1	Introducción	8
2	Detección automática del discurso de odio	10
2.1.	Definición de la noción ‘discurso de odio’	10
2.2.	Estudio comparativo de los métodos utilizados	14
2.2.1.	Modelos basados en reglas.....	15
2.2.2.	Modelos basados en lexicones	16
2.2.3.	Modelos de aprendizaje automático.....	16
2.3.	Detección automática del discurso de odio misógino	21
2.3.1.	Enfoques clásicos de aprendizaje automático	21
2.3.2.	Enfoques de aprendizaje profundo.....	22
2.4.	Problemas a los que se enfrenta la investigación	23
2.4.1.	Escasez de datos no ingleses.....	23
2.4.2.	Carácter implícito de discurso de odio.....	24
2.4.3.	Falta de sistemas transparentes	25
2.4.4.	dificultad de distinguir entre palabrotas y discurso de odio.....	26
3.	Datos	27
3.1.	Recopilación de datos.....	27
3.2.	Anotación de datos	29
4.	Configuración experimental.....	36
5.	Resultados	38
6.	Discusión.....	41
6.1.	Análisis de errores	42
6.1.1.	Análisis de errores por categorías	42
6.1.2.	Patrones generales de error	46
6.2.	Interpretación de los resultados	49
7.	Conclusión	50
	Bibliografía	52

< Lista de abreviaturas >

1. PLN = Procesamiento del Lenguaje Natural (NLP; Natural Language Processing)
2. IA = Inteligencia Artificial
3. SVM = Support Vector Machine
4. LLM = Large Language Model
5. RBF = Radial Basis Function
6. MLP = Multilayer Perceptron
7. MNP = Masked Numeral Prediction
8. CNN = Convolutional Neural Networks
9. GRU = Gated Recurrent Unit
10. CRF = Conditional Random Field
11. API = Application Programming Interface

< Lista de tablas y figuras

Figuras

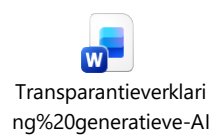
1.	pp. 18
2.	pp. 30

Tablas

1.	pp. 28
2.	pp. 38
3.	pp. 39
4.	pp. 39
5.	pp. 40
6.	pp. 40
7.	pp. 45

Imágenes

1.	pp. 35
2.	pp. 35
3.	pp. 36



1 INTRODUCCIÓN

El primer tuit fue publicado por el fundador de la plataforma, Jack Dorsey, el 21 de marzo de 2006. Desde entonces, transcurrieron nada menos que tres años hasta la publicación del milmillonésimo tuit. Desde el inicio del año en que se alcanzó este hito, en 2009, el número de tuits emitidos diariamente ha aumentado de 2,5 a 500 millones (*Twitter Usage Statistics - Internet Live Stats*, s. f.), lo que implica que, en la actualidad, solo se requieren dos días para alcanzar esa misma cantidad de mensajes publicados. En opinión de Tontodimamma et al. (2020), esto se debe a la forma en que el rápido crecimiento de redes sociales, como la mencionada anteriormente, ha cambiado nuestra manera de comunicarnos en las últimas décadas. Mientras que antes las ideas debían comunicarse oralmente a un pequeño grupo de personas, actualmente se puede elegir entre una amplia variedad de redes para difundirlas a un público extenso, ya sea por escrito. Ese fenómeno ha generado beneficios sustanciales en términos de interconectividad, al permitir una difusión masiva de ideas y facilitar que los negocios prosperen a nivel global con mayor facilidad que en cualquier otra época (Cignarella et al., 2025). Además, los usuarios experimentan una mayor libertad en la manera de expresarse al escribir, ya que no tienen que identificarse necesariamente para publicar mensajes, lo cual les permite distanciarse de su verdadera identidad. Desafortunadamente, no todos los usuarios emplean esta libertad de expresión para redactar mensajes positivos; por el contrario, con creciente frecuencia, se observa la proliferación de tuits ofensivos o que incitan al odio. Las redes sociales también están al tanto de esta problemática y han implementado estrategias para protegerse tanto a sí mismas como a sus usuarios, moderando, en la medida de lo posible, la comunicación que promueve el discurso de odio. El presente trabajo se inscribe en este marco general, pero se centra específicamente en el discurso de odio misógino.

En los comienzos, recurrían a anotadores humanos para llevar a cabo esta tarea, quienes debían revisar manualmente todo lo que se publicaba. No obstante, el volumen de contenido generado en la red es tan vasto que resulta imposible controlarlo todo, lo que llevó a que una parte considerable del lenguaje ofensivo pasara desapercibida. Además, este proceso requiere una inversión significativa de tiempo y de recursos económicos, dado que requiere la contratación de un gran número de empleados para asumir la moderación de contenido. Como consecuencia de estas limitaciones, se ha vuelto imprescindible desarrollar métodos automatizados que permitan la detección eficiente del discurso de odio en redes sociales mediante técnicas de

procesamiento del lenguaje natural (PLN). Aunque la mayoría de los estudios sobre la detección automática del discurso de odio misógino en redes sociales se han realizado en lengua inglesa, en el presente trabajo se ha optado por ampliar el análisis más allá del inglés e incluir también el español y el neerlandés. Esta decisión responde a la escasez de investigaciones y recursos disponibles en estas lenguas, especialmente en el caso del neerlandés, como se abordará más adelante.

La intención de esta tesina es abordar las siguientes preguntas de investigación:

1. ¿En qué medida permite un modelo lingüístico multilingüe detectar el discurso de odio misógino en tuits redactados en neerlandés, inglés y español?
2. ¿Se observan diferencias entre el neerlandés, el inglés y el español en los puntos fuertes y débiles de la detección automática del discurso de odio misógino al considerar categorías discursivas como la implicitud, la agresividad, los estereotipos y el humor?

En el segundo capítulo se expone el marco teórico de esta investigación, proporcionando una revisión y síntesis de una selección de los estudios más relevantes respecto a la detección automática del discurso de odio misógino en las redes sociales. Se definen algunas nociones básicas relacionadas con el PLN, se exponen los métodos más relevantes para la detección automática y se matizan los estudios, insistiendo en los problemas a los que a menudo se enfrentan los sistemas investigados.

El tercer capítulo se centra en el proceso de recopilación y anotación de los datos. Se describen tanto la construcción del conjunto de datos en neerlandés como las distintas subcategorías incluidas en el esquema de anotación aplicado a las tres lenguas. Asimismo, se presentan las directrices seguidas durante este proceso, acompañadas de ejemplos concretos en neerlandés, inglés y español, con el fin de delimitar con claridad los criterios aplicados.

En el cuarto capítulo se expone la configuración experimental adoptada en el estudio. Se presenta el modelo seleccionado y se justifica su elección en función de los objetivos de la presente investigación, antes de describir las tres configuraciones de ajuste aplicadas.

El quinto capítulo presenta los resultados obtenidos por el modelo, primero desde una perspectiva general y, a continuación, en función de las categorías discursivas analizadas. En

el sexto capítulo, estos resultados se interpretan y contextualizan mediante un análisis de errores, tanto a partir de las categorías consideradas por separado como desde una perspectiva más general sobre los tipos de errores observados en las tres lenguas.

Finalmente, en la conclusión se ofrece una síntesis de las conclusiones más significativas y se proponen algunas posibles líneas de investigación futura.

Antes de seguir, conviene señalar que en la redacción de este trabajo de fin de máster, se ha hecho un uso responsable de herramientas de inteligencia artificial. En otras palabras, dado que el español no es la lengua materna de la autora, estas herramientas se han utilizado principalmente como apoyo lingüístico. En una primera instancia, su empleo se centró en incrementar la precisión lingüística del texto y, posteriormente, en adaptar la redacción a un registro más académico. Sin embargo, en todo momento, la redacción inicial fue realizada antes de incorporar el contenido a un modelo de IA. Asimismo, se ha utilizado IA para visualizar el esquema de anotación presentado en el apartado 3.2. Las herramientas empleadas con mayor frecuencia para este propósito han sido ChatGPT y Microsoft CoPilot.

2 DETECCIÓN AUTOMÁTICA DEL DISCURSO DE ODIOS

2.1. Definición de la noción ‘discurso de odio’

Antes de desarrollar el marco teórico de la investigación, resulta pertinente señalar que el discurso de odio no constituye un fenómeno abordado exclusivamente desde la perspectiva del procesamiento del lenguaje natural (PLN), sino que también ha constituido un objeto de estudio en disciplinas como la sociología, entre otras. Este enfoque multidisciplinar ha propiciado la formulación de diversas denominaciones a lo largo de la evolución de su estudio, antes de consolidarse en el término “discurso de odio”.

Como señalan Schmidt y Wiegandt en su obra (2017), en las primeras investigaciones sobre este fenómeno, autores como Spertus (1997) defendieron el uso del término ‘discurso ofensivo’, mientras que, en años más recientes, otros investigadores (Xu et al., 2012; Hosseinmardi et al., 2015; Zhong et al., 2016; Van Hee et al., 2015; Dadvar et al., 2013; Dinakar et al., 2012) han tendido a emplear la noción de ‘ciberacoso’ para referirse a una

categoría amplia que engloba no solo el discurso de odio, sino también otras formas de agresión o comportamiento perjudicial en línea. No obstante, en el ámbito del PLN, se ha optado de manera predominante por la expresión ‘discurso de odio’, no solo por su capacidad para englobar distintos fenómenos interrelacionados bajo un solo término paraguas, sino también por su reconocimiento como categoría jurídica en numerosos países a nivel global.

Así en la ley española, por ejemplo, se puede leer:

Serán castigados con una pena de prisión de uno a cuatro años y multa de seis a doce meses:

- a) Quienes públicamente fomenten, promuevan o inciten directa o indirectamente al odio, hostilidad, discriminación o violencia contra un grupo, una parte del mismo o contra una persona determinada por razón de su pertenencia a aquel, por motivos racistas, antisemitas, antigitanos u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, aporofobia, enfermedad o discapacidad (*Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal.*, 1995).

Si bien muchos países, como España, regulan este fenómeno a través de distintos marcos jurídicos, no existe una normativa universalmente uniforme. Esta ausencia de una definición comúnmente aceptado, tanto al nivel jurídico que al nivel científico, constituye uno de los problemas fundamentales al cual se enfrenta el ámbito del PLN en la actualidad. Según Strossen (2016), esta falta de precisión conceptual se debe a la naturaleza subjetiva de la evaluación del lenguaje, lo que dificulta establecer con exactitud si una expresión es ofensiva o incita al odio.

Por ejemplo, el Oxford Learner’s Dictionary define el discurso de odio como “speech or writing that attacks or threatens a particular group of people, especially on the basis of race, religion or sexual orientation” (*hate speech noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner’s Dictionary at OxfordLearnersDictionaries.com*, s. f.).

No obstante, dado que los juicios sobre el lenguaje se encuentran mediados por las experiencias individuales, esta definición permite un margen de interpretación. Así, por ejemplo, la falta de experiencia directa en determinadas situaciones puede dificultar el reconocimiento de las injusticias que afectan a otros, lo que a menudo conduce a una menor sensibilidad frente a dichas problemáticas. Por esta razón, resulta fundamental que las personas encargadas de

etiquetar los datos presenten perfiles diversos, a fin de garantizar la mayor precisión posible en los resultados.

Con el propósito de reducir aún más esta subjetividad en la clasificación de un tuit como ‘discurso de odio’, los autores Waseem & Hovy (2016) proponen la utilización de una lista de criterios específicos con los que se comprobarán los mensajes, evitando así que los casos problemáticos pasen desapercibidos debido a discrepancias entre los anotadores.

Según los autores, un tuit puede considerarse ofensivo si:

1. Utiliza un insulto sexista o racista;
2. Ataca a una minoría;
3. Busca a silenciar una minoría;
4. Critica a una minoría (sin tener un argumento bien fundamentado);
5. Promueve, pero no utiliza directamente, el discurso de odio o crimen violento;
6. Critica a una minoría y utiliza argumentos de hombre de paja;
7. Tergiversa descaradamente la verdad o intenta distorsionar las opiniones sobre una minoría con afirmaciones infundadas;
8. Muestra apoyo a hashtags problemáticos, como #BanIslam, #whoriental o #whitegenocide;
9. Estereotipa negativamente a una minoría;
10. Defiende la xenofobia o el sexismo;
11. Contiene un nombre de usuario ofensivo, según los criterios anteriores, el tuit es ambiguo (en el mejor de los casos), y el tuit trata un tema que cumple alguno de los criterios anteriores.

La evolución en la definición del discurso de odio, pasando de la ambigua expresión ‘lenguaje que incita al odio’ a un enfoque basado en estas características específicas, representa un avance significativo, en los ámbitos jurídico y científico. Este cambio no solo permite una detección más precisa y eficiente, tanto por parte de anotadores humanos como de sistemas automatizados, sino que también mejora la transparencia del proceso. La mayor claridad en los criterios de identificación contribuye a reducir el riesgo de sesgo en los resultados y en las conclusiones, lo que, a su vez, fortalece la fiabilidad de los estudios científicos en este ámbito. Este cambio de enfoque ha sido ampliamente reconocido en el ámbito del PLN, donde un número creciente de investigadores ha profundizado en el análisis de los diversos factores

asociados al discurso de odio con el objetivo de minimizar el margen de interpretación y garantizar una mayor precisión en su definición y detección. Como resultado, la tendencia ha sido avanzar hacia una definición que refleje con mayor exactitud la complejidad de este fenómeno.

En la actualidad, la definición propuesta por John Nockleby es ampliamente utilizada como referencia para este propósito en el ámbito del PLN (Warner & Hirschberg, 2012). Según su propuesta, el discurso de odio se puede definir como “any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion or other characteristic” (citado en Schmidt & Wiegandt, 2017, p. 1). Contrariamente a lo que cabría esperar, esta definición no es reciente, sino que fue formulada en el año 2000, lo que evidencia que la búsqueda de una definición unificada en este ámbito ha sido un tema de estudio recurrente en la comunidad científica durante años. En una fase posterior, esta definición se aplicará específicamente al discurso dirigido contra las mujeres y se utilizará como criterio operativo para determinar qué casos pueden clasificarse como discurso de odio misógino en el marco de la presente investigación.

Antes de proceder con la sección comparativa de este estudio, resulta pertinente reflexionar brevemente sobre otros posibles problemas asociados a la definición del discurso de odio. De hecho, no es cierto que el problema expuesto en las páginas anteriores se resolvería de manera inmediata si existiera una definición inequívoca y universalmente aceptada. Esto se debe, principalmente, a que la incitación al odio está estrechamente vinculada al concepto de libertad de expresión, un tema que, por sí mismo, plantea cuestiones complejas y delicadas. Al igual que ocurre con el concepto de incitación al odio, la libertad de expresión también se configura como un concepto jurídico, sobre el cual la legislación española establece lo siguiente: “La libertad de expresión y el derecho a la difusión de informaciones por medio de impresos gráficos o sonoros, no tendrá más limitaciones que las establecidas en el ordenamiento jurídico con carácter general” (*BOE-A-1977-9008 Real Decreto-ley 24/1977, de 1 de abril, sobre libertad de expresión.*, s. f.). Al comparar este texto legal con el relativo a la incitación al odio, se observa una zona ambigua que se sitúa entre ambos conceptos y que es interpretada de manera diversa según el contexto y la persona. De este modo, lo que para algunos constituye una expresión legítima de la libertad de expresión, para otros puede considerarse un exceso que entra en el ámbito de la incitación al odio. Esta ambigüedad se ve aún más complejizada por el hecho de que la incitación al odio se manifiesta predominantemente en el entorno digital, lo

que dificulta la determinación del marco jurídico específico bajo el cual debe evaluarse. Con frecuencia, un usuario de un país emite críticas hacia otro de un país diferente, mientras que la sede de la plataforma donde ocurre el incidente se encuentra en un país distinto a los dos previamente mencionados. Casos como este ilustran claramente los desafíos prácticos involucrados, lo que complica significativamente el enjuiciamiento de individuos por este tipo de conducta. A pesar de la complejidad de la problemática, su importancia se ve reforzada por su dimensión jurídica, dado que el incumplimiento de las normativas que regulan este fenómeno puede acarrear consecuencias legales considerables, como hemos mencionado anteriormente. Por tanto, si bien es cierto que este ámbito se enfrenta a múltiples problemas, el más fundamental sigue siendo la ausencia de una definición de trabajo comúnmente aceptada del término.

2.2. Estudio comparativo de los métodos utilizados

En la última década, el campo del PLN ha experimentado un notable impulso por avances tecnológicos, una mayor disponibilidad de datos a gran escala y el desarrollo de algoritmos y modelos más sofisticados (Van Hee & Hoste, 2024). A partir de estos avances recientes, se han desarrollado numerosas investigaciones (Gitari et al., 2015; Waseem & Hovy, 2016; Schmidt & Wiegand, 2017; Röttger et al., 2022) orientadas a la detección automática de discursos de odio en contenidos textuales. Antes de examinar los principales métodos que han sido propuestos a lo largo del tiempo con este propósito, resulta pertinente detenerse brevemente en la definición del procesamiento del lenguaje natural.

Según se expone en el sitio web de IBM, el PLN se define como “un subcampo de la informática y la inteligencia artificial (IA) que utiliza el machine learning para permitir que los ordenadores entiendan y se comuniquen con el lenguaje humano” (*¿Qué es el PLN (procesamiento del lenguaje natural)?*, 2021). En otras palabras, las técnicas de PLN permiten a los sistemas informáticos interpretar y ejecutar instrucciones expresadas en lenguaje natural. En los párrafos que siguen, se presenta una visión general de los sistemas desarrollados en este ámbito hasta la fecha.

2.2.1. Modelos basados en reglas

Los primeros sistemas de PLN eran basados en reglas para el análisis y procesamiento de textos (Van Hee & Hoste 2024). Estas reglas, elaboradas por expertos en lingüística, se utilizaban para llevar a cabo tareas específicas, como el análisis sintáctico, el etiquetado gramatical o la extracción de información en lenguaje natural.

Así, por ejemplo, Chikersal et al. (2015) propusieron las siguientes reglas para determinar si un tuit es positivo o negativo:

- Si un tuit contiene únicamente emoticonos positivos y ninguno negativo, se clasifica como positivo.
- Si un tuit contiene únicamente emoticonos negativos y ninguno positivo, se clasifica como negativo.

Aunque los modelos basados en reglas representaron un avance significativo en el momento de su aparición, con el tiempo se han puesto de manifiesto diversas limitaciones inherentes a este enfoque (Clarke et al., 2023). La elaboración de reglas puede volverse compleja y extensa, ya que estas deben contemplar numerosas excepciones, lo que requiere una considerable inversión de tiempo y esfuerzo por parte de expertos humanos. Además, dichas reglas deben actualizarse periódicamente, ya que no incorporan de manera automática los cambios y la evolución propios del lenguaje. Asimismo, presentan una limitada capacidad para gestionar la ambigüedad, lo que los hace especialmente propensos a errores en contextos lingüísticos complejos. No obstante, la principal fortaleza de los enfoques basados en reglas radica en su transparencia y en la facilidad con que pueden ser interpretados por los investigadores, en contraste con los sistemas de aprendizaje profundo que predominan en la actualidad (Clarke et al., 2023). Esta característica ha motivado que algunos autores, como Gitari et al. (2015), propongan enfoques híbridos que combinan la precisión y el control de los sistemas basados en reglas con la flexibilidad y la capacidad de generalización de los modelos contemporáneos para la tarea de detección automática de discurso de odio. Como continuación de esta línea de trabajo, en una etapa posterior se optó por integrar también enfoques basados en lexicones, concebidos como un complemento a los sistemas basados en reglas.

2.2.2. Modelos basados en lexicones

Los enfoques basados en lexicones comenzaron a desarrollarse y a adquirir relevancia a finales del siglo XX, como lo demuestran las numerosas investigaciones realizadas en esa época al respecto (Hatzivassiloglou & Mckeown, 1997; Turney, 2002). En aquel momento, sin embargo, la detección del discurso de odio aún no constituía un área de estudio consolidada, por lo que estos enfoques se aplicaban principalmente al análisis de sentimientos.

El análisis de sentimientos basado en lexicones se sustenta en un lexicón de polaridad emocional previamente definido, con el objetivo de determinar si el tono afectivo de un texto es positivo, negativo o neutro (Voll & Taboada, 2007). Dicho lexicón consiste, por lo general, en un conjunto de unidades léxicas que han sido etiquetadas manualmente en función de su orientación semántica. Aunque estos sistemas presentan ciertas limitaciones, como lo exponen Van Hee y Hoste (2024), continúan siendo empleados en la actualidad para la tarea del análisis de sentimientos, particularmente en contextos de marketing, donde resulta fundamental para los vendedores conocer la percepción general que los consumidores tienen acerca de sus productos o servicios. Dado el estrecho vínculo entre el análisis de sentimientos y la detección de discurso de odio, diversos investigadores (Gitari et al., 2015; MacAveney et al., 2019; Tanev, 2024) han optado por aplicar enfoques basados en lexicones también en este ámbito. Según lo planteado por MacAveney et al. (2019), estos enfoques se caracterizan por una alta precisión, pero presentan una capacidad de recuperación limitada, debido a las dificultades para resolver la ambigüedad léxica y para interpretar adecuadamente el lenguaje figurado. Como consecuencia, estos sistemas pueden pasar por alto expresiones de discurso de odio que no contienen términos explícitamente ofensivos. En contraste, el enfoque léxico propuesto por Gitari et al. (2015) ofrece una perspectiva más matizada, al evidenciar niveles relativamente altos tanto de precisión como de recuperación, lo cual también se observa en la investigación de Tanev (2024).

2.2.3. Modelos de aprendizaje automático

La principal diferencia con respecto a los dos sistemas previamente descritos radica en que los que se presentan a continuación incorporan técnicas de aprendizaje automático (*machine learning*). En términos concretos, esto implica que se centran en “entrenar a computadoras y máquinas para imitar el modo en que aprenden los humanos, realizar tareas de forma autónoma

y mejorar su rendimiento y precisión a través de la experiencia y la exposición a más datos” (*¿Qué es el aprendizaje automático (ML)?*, 2021). Para lograrlo, estos sistemas pueden ser entrenadas utilizando conjuntos de datos previamente etiquetados, a partir de los cuales adquieren la capacidad de generalizar el conocimiento aprendido y de realizar predicciones sobre datos no observados previamente (Van Hee y Hoste, 2024). El campo de investigación se divide, por lo general, en dos grandes vertientes: los enfoques basados en características, que constituyeron el método predominante hasta el año 2015, y los enfoques de aprendizaje profundo, dentro de los cuales se encuentran los transformadores, que siguen siendo en la actualidad la tecnología más avanzada para la detección automática del discurso de odio. En los párrafos que siguen se profundizará en ambos vertientes.

2.2.3.1. Modelos basados en características (*feature-based*)

El ejemplo por excelencia de un sistema de este tipo es el modelo basado en características. Antes de abordar su funcionamiento, es preciso señalar que la simple incorporación de datos al sistema no resulta suficiente, ya que, antes de ser utilizados por los algoritmos de aprendizaje automático, estos deben ser adecuadamente procesados (K. Pitsilis et al., 2018). Esta fase inicial, conocida como preprocesamiento, consiste en la limpieza, transformación y estructuración de los datos, con el objetivo de asegurar que sean apropiados y eficaces para su posterior análisis por parte del modelo (Kowsari et al., 2020). A continuación, tiene lugar la fase de extracción de características (*features*), un proceso mediante el cual los datos originales son transformados matemáticamente en un nuevo conjunto de variables que, como lo explican Van Hee y Hoste (2024), capturan la información más relevante para la tarea específica que debe abordar el modelo. La importancia de esta etapa radica en su capacidad para revelar estructuras subyacentes que podrían pasar desapercibidas en los datos en su forma original. Un ejemplo concreto lo proporcionan Ibrohim & Budi (2019), quienes aplican un conjunto de características ampliamente reconocidas en el ámbito de la detección automática del discurso de odio, entre las cuales destacan la frecuencia de términos mediante n-gramas de palabras y de caracteres, así como distintos indicadores ortográficos, como el uso de signos de exclamación y de letras mayúsculas.

Antes de que el sistema pueda ser utilizado realmente para la clasificación de tuits como discurso de odio o no, es necesario que atraviere una fase de entrenamiento, en la cual se le presentan datos previamente etiquetados por anotadores humanos a un algoritmo de

aprendizaje automático (Van Hee & Hoste, 2024). El objetivo fundamental de esta etapa es permitir que el modelo identifique y aprenda patrones subyacentes en dichos datos etiquetados. A partir de este proceso, el sistema es capaz de extraer de forma automática reglas o regularidades estadísticas, que, posteriormente, podrá aplicar a textos no vistos previamente, con el fin de realizar predicciones precisas sobre su carácter potencialmente ofensivo u odioso. Para llevar a cabo esta tarea, existen numerosos algoritmos, cada uno de los cuales adopta enfoques distintos para el proceso de aprendizaje. En el presente trabajo se abordarán dos de los más ampliamente utilizados para la tarea de detección automática de discurso de odio, a saber las máquinas de vectores de soporte y los clasificadores Naive Bayes.

Las máquinas de vectores soporte (Support Vector Machines, SVM) ya fueron desarrolladas en 1995 por Vladimir N. Vapnik y sus colaboradores, quienes presentaron su propuesta en el artículo titulado Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. Este tipo de modelo distingue entre dos clases mediante la identificación del hiperplano óptimo que maximiza el margen entre los puntos de datos más cercanos pertenecientes a clases opuestas. En el ámbito del aprendizaje automático, las SVM han sido ampliamente adoptadas, ya que pueden aplicarse tanto a tareas de clasificación lineal como no lineal. En este último caso, uno de los elementos fundamentales es el uso de funciones kernel, las cuales se emplean cuando los datos no son linealmente separables, ya que permiten transformarlos a un espacio de mayor dimensión en el que se vuelve posible establecer una separación lineal entre las clases.

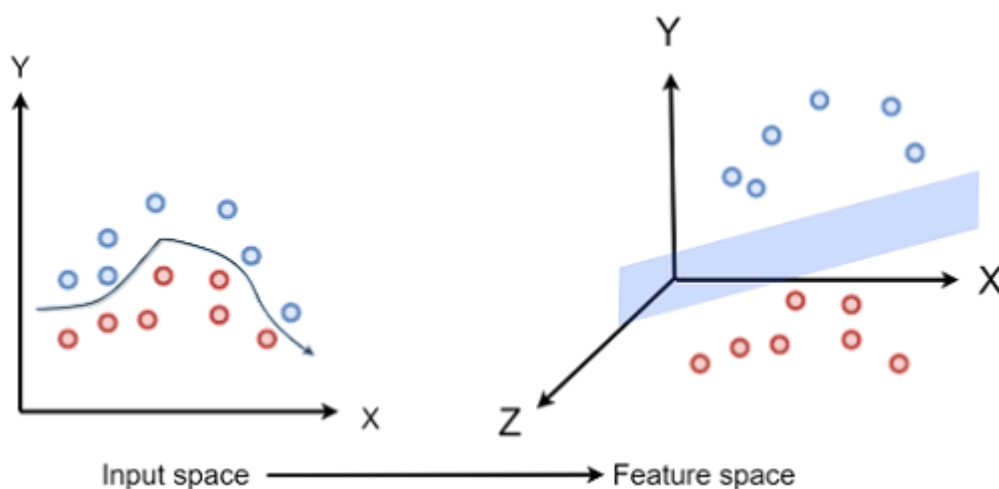


Figura 1: SVM con RBF | Sidharth GN, 2022, QuarkML

Un ejemplo concreto es el estudio de Chen et al. (2012), quienes emplearon máquinas de vectores de soporte (SVM) en su investigación sobre la detección automática del discurso de odio, y sostienen que este enfoque resulta más eficaz para esta tarea que el modelo de Naive Bayes, el cual se aborda en el párrafo que sigue.

El clasificador Naive Bayes, también conocido bajo el nombre de clasificador probabilístico, se basa en el teorema del matemático inglés Thomas Bayes. Según lo que se expone en el sitio web de IBM (*¿Qué son los clasificadores Naive Bayes?*, s. f.), pertenece a la familia de algoritmos de aprendizaje generativo, lo cual implica que su objetivo es modelar la distribución de las características de entrada asociadas a una clase o categoría específica. Para ello, el algoritmo calcula la probabilidad asociada a cada clase y selecciona aquella con la mayor probabilidad estimada como la predicción final. Asogwa et al. (2022) emplearon tanto máquinas de vectores soporte como el modelo Naive Bayes en su estudio, y este demostró, al igual que el de Chen et al. (2012) mencionado más arriba, que los SVM resultan más eficaces para la tarea de clasificación automática de discursos de odio. Antes de avanzar en el desarrollo del marco teórico, conviene señalar que existen otros algoritmos que también abordan esta tarea, pero que han resultado menos relevantes para los objetivos de la presente investigación.

2.2.4. Modelos de aprendizaje profundo

A partir del año 2013, ha cobrado un interés creciente un subdominio del aprendizaje automático, conocido como aprendizaje profundo. La diferencia fundamental entre ambos conceptos radica en la arquitectura subyacente: mientras que el aprendizaje automático tradicional se basa en algoritmos tales como árboles de decisión, regresión logística o SVM, el aprendizaje profundo emplea redes neuronales considerablemente más complejas, compuestas por múltiples capas y capaces de extraer automáticamente características a partir de grandes cantidades de datos con un alto grado de complejidad (*¿Qué es el aprendizaje profundo?* | IBM, 2024). Dicha arquitectura confiere a estos sistemas una capacidad sobresaliente para el reconocimiento de patrones en datos (semi)estructurados, entre los cuales el lenguaje humano constituye un ejemplo representativo. Si bien existen múltiples enfoques dentro del aprendizaje profundo, el presente estudio se centra exclusivamente en los modelos de transformadores, por ser los más relevantes en lo que respecta a la detección del discurso de odio.

Esta arquitectura fue presentada por primera vez en el artículo titulado *Attention is All You Need* (Vaswani et al., 2017). Ese nombre no fue elegido de manera arbitraria, sino que alude al hecho de que los transformadores emplean el mecanismo de autoatención (*self-attention mechanism*), el cual permite identificar qué palabras poseen mayor relevancia dentro de una frase, de modo que pueda asignarles una atención proporcionalmente mayor y, con ello, lograr una mejor comprensión del enunciado (Van Hee & Hoste, 2024). Además, este tipo de modelo es capaz de procesar grandes volúmenes de datos simultáneamente, los cuales se introducen en el sistema en forma de incrustaciones de palabras. Estas incrustaciones constituyen una forma particular de vectores de palabras en la que tanto el significado de los términos como las relaciones entre ellos se codifican mediante un conjunto limitado de valores numéricos (Van Hee & Hoste, 2024). Al estar compuestas por cifras, dichas representaciones pueden presentarse en espacios multidimensionales, lo que permite su manipulación matemática por parte de los programas informáticos, permitiendo, por ejemplo, predecir la palabra más probable que seguirá en una secuencia (Vaswani et al., 2017). Dichos cálculos sustentan el funcionamiento de los grandes modelos lingüísticos (*Large Language Models; LLM*), los cuales se basan en redes neuronales profundas para aprender a reconocer y generar patrones en contextos lingüísticos complejos. Un ejemplo representativo es BERT, el LLM desarrollado por Google. Estos modelos pueden, además, ajustarse a tareas específicas, como ha ocurrido con modelos como HateBert, que es un modelo BERT reentrenado para la detección de lenguaje abusivo en inglés. Fue entrenado utilizando RAL-E, un corpus a gran escala conformado por comentarios en inglés extraídos de comunidades de Reddit que fueron vetadas por contener contenido ofensivo, abusivo o cargado de odio (Caselli et al., 2021). Tal como señalan Caselli et al. (2021), el preentrenamiento adicional constituye una estrategia eficaz para adaptar modelos lingüísticos a tareas específicas, como la detección de discurso de odio, lo que ha propiciado una amplia cantidad de investigaciones basadas en dicho modelo (Nirmal et al., 2024; Guillaume et al., 2022; Kim et al., 2022). En la actualidad, también han cobrado importancia los modelos generativos, que han mostrado un rendimiento prometedor en tareas multilingües, lo que los convierte en herramientas potencialmente útiles para la detección automática del discurso de odio (Ghorbanpour et al., 2025). Sin embargo, dado que su ajuste requiere una cantidad considerable de recursos computacionales y que no constituyen el objeto central del presente estudio, se ha decidido no abordarlos en detalle en esta síntesis.

2.3. Detección automática del discurso de odio misógino

Dado que el campo de la detección automática del discurso de odio misógino en las redes sociales es relativamente reciente y solo empezó a desarrollarse hace aproximadamente diez años (Hewitt et al., 2016), no existe aún un marco teórico ampliamente consolidado. Por ello, la literatura existente se basa en gran medida en el ámbito más general de la detección automática del discurso de odio y en descripciones de tareas compartidas como AMI@Evalita, AMI@IberEval, HatEval, OffenseEval y SemEval. Al igual que en el resto de este campo de investigación, pueden identificarse dos enfoques principales en la subtask de detección de misoginia: las técnicas clásicas de aprendizaje automático y el uso de redes neuronales. Este último enfoque ha adquirido un papel predominante, no solo por su elevado rendimiento, sino también porque la consolidación del campo de la detección automática del discurso de odio misógino coincidió con una etapa en la que los modelos neuronales comenzaban a ganar relevancia en el ámbito del PLN. En los párrafos que siguen, se presenta un panorama general de los métodos empleados hasta la fecha, ofreciendo una visión representativa de los modelos más relevantes.

2.3.1. Enfoques clásicos de aprendizaje automático

Aunque Basile et al. (2019) señalan que la mayoría de los cinco equipos con mejor rendimiento en la *SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter* emplearon redes neuronales, resulta destacable que el valor más alto de macro-promedio para la lengua inglesa en la subtask A (0.65) fuera obtenido por el equipo Fermi mediante un modelo SVM con kernel RBF entrenado sobre los datos proporcionados. Asimismo, para la misma tarea en lengua española, los mejores resultados se alcanzaron mediante sistemas SVM, aunque seguidos muy de cerca por un sistema que adapta modelos BERT preentrenados a datos de Twitter, utilizando un corpus de tuits recopilados durante el mismo período temporal que el conjunto de datos de entrenamiento de HatEval (Basile et al., 2019). En la misma línea, Nina-Alcocer (2018) emplea una combinación de enfoques de aprendizaje automático como SVM, perceptrón multicapa (MLP) y Naive Bayes Multinomial (MNB), dado el alto rendimiento de estos modelos al utilizar un gran número de características. También en la subtask B de AMI@EVALITA 2018, centrada en la clasificación de los tuits según tipos más específicos de misoginia y según el blanco del mensaje, los mejores resultados fueron obtenidos mediante técnicas clásicas de aprendizaje

automático. Concretamente, estos fueron alcanzados por un conjunto de modelos del equipo Himani, un sistema basado en SVM del equipo CrotoneMilano y un modelo de regresión logística del equipo Hateminers (Shushkevich y Cardiff, 2019).

2.3.2. Enfoques de aprendizaje profundo

Aunque los enfoques clásicos de aprendizaje automático ofrecen resultados sólidos en la detección y clasificación del discurso de odio misógino, Shushkevich y Cardiff (2019) subrayan que las redes neuronales representan una alternativa prometedora, capaz de alcanzar resultados igualmente competitivos en este ámbito. Una de sus principales ventajas es que, al tratarse de modelos capaces de aprender automáticamente a partir de datos etiquetados, no requieren un proceso manual de ingeniería de características, lo que facilita su aplicación a nuevos conjuntos de datos. Esta característica justifica la elección de un enfoque neuronal en el presente estudio, aun cuando los modelos clásicos también han mostrado resultados comparables en este ámbito.

En esta línea, Goenaga et al. (2018) proponen una arquitectura neuronal basada en una red Bi-LSTM con CRF para AMI-IberEval 2018, con la que obtienen una precisión de 78.9 en los tuits en inglés y de 76.8 en los tuits en español. Los autores justifican la elección de esta arquitectura por su capacidad para procesar el texto de manera secuencial y captar información contextual, algo que resulta más difícil para técnicas clásicas como SVM. Como señalan Shushkevich y Cardiff (2019), Zhang y Luo (2018) también emplean una arquitectura neuronal, concretamente una red neuronal convolucional (*convolutional neural network, CNN*) con tres capas convolucionales y una capa GRU añadida. Según los autores, esta estructura más simple favorece el entrenamiento y la generalización en conjuntos de datos de menor tamaño. Los resultados obtenidos respaldan esta elección, ya que el modelo alcanza un rendimiento elevado en la clasificación de datos de Twitter en tres grupos: mensajes sexistas, mensajes racistas y tuits que no pertenecen a ninguna de estas dos categorías.

Por su parte, Rodríguez-Sánchez et al. (2020) comparan tanto modelos clásicos de aprendizaje automático como modelos neuronales. De su estudio se desprende que las configuraciones con mejores resultados son mBERT combinado con características textuales y Bi-LSTM con embeddings preentrenados WESB, lo que confirma el buen rendimiento de los enfoques neuronales en este tipo de tareas. En el caso de mBERT, este rendimiento se atribuye a su

capacidad para captar información sintáctica y contextual, en consonancia con la justificación planteada por Goenaga et al. (2018). Por otro lado, la eficacia de Bi-LSTM con WESB puede explicarse por el hecho de que estos embeddings fueron entrenados sobre tuits y, por tanto, se ajustan mejor a las particularidades lingüísticas de Twitter. Dentro de los modelos neuronales aplicados en este estudio, los resultados obtenidos son, además, muy comparables.

Aunque estos modelos muestran resultados prometedores y avances considerables, los análisis de errores muestran que la detección automática del discurso de odio misógino sigue planteando dificultades. Estos problemas se relacionan, entre otros aspectos, con fenómenos lingüísticos y discursivos que los sistemas automáticos aún no logran captar de manera plenamente fiable. Por ello, futuras investigaciones deberán ofrecer soluciones que permitan reducir estas limitaciones y avanzar hacia el desarrollo de modelos más robustos.

2.4. Problemas a los que se enfrenta la investigación

Así pues, como se ha expuesto en los párrafos anteriores, se han logrado avances significativos en la detección automática del discurso de odio en un corto periodo de tiempo. De este modo, en menos de cincuenta años, se ha pasado de sistemas basados en lexicones completamente manuales a modelos automáticos fundamentados en inteligencia artificial generativa. No obstante, la investigación sigue enfrentándose a diversas limitaciones que restringen su impacto en el desarrollo de sistemas de detección aplicables en contextos reales.

2.4.1. Escasez de datos no ingleses

Un primer desafío importante al que se enfrenta el ámbito de la detección automática de discurso de odio es que, como lo exponen Röttger et al. (2022), si bien el discurso de odio constituye un problema a nivel global, la investigación se ha centrado casi exclusivamente en datos en lengua inglesa. Esta tendencia se debe, en parte, al hecho de que el inglés suele funcionar como idioma predominante en la comunicación dentro de las redes sociales y, además, a la disponibilidad relativamente amplia de corpus anotados en inglés en comparación con otros idiomas. Así, por ejemplo, Abercrombie et al. (2023) constatan que el 41,3% de los estudios incluidos en su revisión continúan fundamentándose en conjuntos de datos en lengua inglesa. No obstante, los autores subrayan que también se han desarrollado contribuciones significativas en quince lenguas adicionales, entre las cuales destacan el español (12,7%), el

árabe (12,7%) y el francés (7,9%). En la misma línea, Basile et al. (2019) señalan que se están desarrollando cada vez más colecciones de datos en lenguas distintas del inglés, como el español y el italiano, impulsadas en gran medida por el aumento de tareas compartidas multilingües. A pesar de estos avances, resulta revelador que, aunque el chino constituye la segunda lengua más hablada a nivel global, apenas se registre un único estudio centrado en esta lengua (Jiang et al., 2022). Esto pone de manifiesto el profundo desequilibrio lingüístico que aún caracteriza a la disciplina, si bien pueden observarse avances incipientes hacia una mayor equidad. Dado que, hasta donde alcanza nuestro conocimiento, no se han identificado estudios previos centrados en la lengua neerlandesa, el objetivo de la presente investigación consiste en contribuir a subsanar esta laguna en el ámbito de la detección automática de odio misógino.

2.4.2. Carácter implícito de discurso de odio

Como lo señalan Vidgen et al. (2019), la investigación sobre la detección automática del discurso de odio resulta fundamental no solo para el ámbito del PLN, sino también para la gestión eficaz de las redes sociales. Estas, conscientes de la imposibilidad de filtrar manualmente el ingente volumen de mensajes publicados diariamente, dependen cada vez más de métodos automáticos. Una segunda gran dificultad radica en el hecho de que, como advierten MacAveney et al. (2019), los propagadores del discurso de odio, conscientes del aumento del escrutinio, recurren con frecuencia al lenguaje implícito para difundir sus mensajes de manera encubierta.

Actualmente, este fenómeno continúa representando un problema considerable, dado que los sistemas más utilizados en la actualidad aún no son capaces de detectar adecuadamente estas formas de incitación al odio. Así, por ejemplo, el estudio de MacAveney et al. (2019) evidenció que el uso del sarcasmo y de las metáforas constituía la principal causa de la clasificación errónea de tuits, seguido de la escasez de contexto y de la presencia de mensajes agresivos que no cumplían estrictamente con la definición de discurso de odio.

Para ilustrar esta problemática, Schmidt y Wiegand (2017) presentan el siguiente ejemplo:

- 1) *Kermit the frog called and he wants his voice back.*

Como seres humanos, somos capaces de interpretar dicho mensaje como una manifestación de burla dirigida a la voz de la mujer aludida y, en consecuencia, como una forma de discurso de

odio implícito. No obstante, los sistemas automáticos de detección carecen, en muchas ocasiones, del conocimiento extratextual necesario para realizar dicha inferencia (Van Hee et al., 2018). De hecho, la ironía se basa a menudo en una contradicción que puede manifestarse en dos componentes literales, o bien en un único componente que entra en conflicto con el conocimiento que poseemos del contexto o con el sentido común (Maladry et al., 2024). Al no disponer por sí solo de esta información, al sistema le resulta imposible inferir el significado subyacente a partir de una expresión lingüística determinada, que, en el caso del discurso de odio, adopta con frecuencia una forma irónica. Como resultado, una proporción significativa de mensajes de odio sigue circulando en Internet, lo que puede desembocar en la comisión de actos de odio (Cignarella et al., 2025).

Para abordar esta problemática, autores como Gao y Huang (2017) proponen entrenar los sistemas de detección automática del discurso de odio con corpus anotados que incluyan información contextual adicional sobre el texto original y la persona que lo escribió, de manera que los modelos puedan integrar dicha información en sus procesos de clasificación. Este último estudio (Gao & Huang, 2017) demostró que, en promedio, los sistemas obtenían una puntuación F entre un 3 % y un 4 % más alta cuando se les proporcionaba contexto que cuando se evaluaban sin él. No obstante, como lo señalan Markov & Daelemans en su artículo (2022), Pavlopoulos et al. (2020) constataron que, si bien la incorporación del contexto modifica de manera significativa el proceso de anotación por parte de los anotadores, no existen pruebas concluyentes de que dicho contexto mejore necesariamente el rendimiento de los sistemas automáticos en la clasificación de mensajes. También el estudio de Menini et al. (2021) reveló que la inclusión de información contextual no se traduce en una mejora del rendimiento en la tarea de detección. Así pues, persiste la falta de consenso en torno a si la inclusión del contexto conduce de manera sistemática a una mejora en el rendimiento de los sistemas de PLN, una cuestión que únicamente podrá ser esclarecida mediante investigaciones futuras.

2.4.3. Falta de sistemas transparentes

No obstante, al intentar identificar en qué aspectos falla la clasificación de dicho lenguaje implícito con el propósito de encontrar posibles soluciones, nos enfrentamos de inmediato a lo que Vidgen et al. (2019) denominan el problema de la caja negra (*black box*). Como lo comentan Nirmal et al. (2024), este concepto hace referencia a la opacidad inherente de los sistemas más recientes basados en aprendizaje profundo, los cuales no revelan de manera

transparente su proceso de razonamiento. En consecuencia, resulta sumamente complejo introducir mejoras, dado que la falta de comprensión sobre el origen de los errores obstaculiza su corrección. Por consiguiente, un número creciente de autores (Vaughan & Wallach, 2021; Toktarova et al., 2023) sostiene que los sistemas y modelos actuales deben procurar una mayor transparencia respecto a las razones subyacentes que motivan sus predicciones, con el fin de que sus resultados sean interpretables. Esta exigencia reviste especial relevancia no solo porque las nociones de incitación al odio y de libertad de expresión están estrechamente vinculadas, lo que implica que los usuarios deben poder comprender por qué un determinado mensaje es calificado como ofensivo, sino también porque permite evidenciar los posibles sesgos (*bias*) presentes tanto en los datos de entrenamiento como en los propios sistemas (Nirmal et al., 2024). A este respecto, aunque estudios como el de Nirmal et al. (2024) han abordado esta problemática y sus propuestas para incrementar la interpretabilidad resultan prometedoras, los resultados obtenidos evidencian que el rendimiento de los métodos propuestos sigue dependiendo en gran medida tanto de la elección del modelo lingüístico subyacente como de la calidad de los datos de entrenamiento. Esto pone de manifiesto que la investigación en este ámbito aún se encuentra en una fase incipiente, y que será necesario un esfuerzo sostenido para desarrollar sistemas de detección del discurso de odio que sean plenamente transparentes e interpretables.

2.4.4. dificultad de distinguir entre palabrotas y discurso de odio

Aunque ya se ha señalado en los párrafos anteriores que el reconocimiento de actos lingüísticos implícitos resulta complejo, esto no implica que la clasificación de enunciados más explícitos no lo sea. Según muestran los resultados de Malmasi y Zampieri (2017), el principal obstáculo al clasificar tuits radica en la distinción entre el uso de palabrotas y el discurso de odio. Dado que ambos fenómenos suelen aparecer estrechamente vinculados, la investigación en este ámbito se ha basado en gran medida en la presencia de palabrotas como característica para el desarrollo de sistemas de clasificación (Malmasi & Zampieri, 2017). Así, por ejemplo, Kwok y Wang (2013) señalan que el 86% de los tuits de su corpus fue caracterizado como racista precisamente por contener lenguaje ofensivo. Sin embargo, en ocasiones un usuario puede publicar un comentario que contiene este tipo de expresiones sin intención de herir a la persona implicada. Para ilustrar esta problemática, Velankar et al. (2022) emplean el siguiente ejemplo:

2) *He is fucking genius.*

Los autores (Velankar et al., 2022) explican que, debido a la presencia de la palabra *'fucking'*, este tuit inofensivo puede ser caracterizado erróneamente como discurso de odio por el sistema. El mismo problema se observa en tuits que reproducen títulos o letras de canciones populares, como *'Bitch Better Have My Money'* de Rihanna (Hewitt et al., 2016). En la misma línea, Bauwelinck et al. (2019) señalan que su clasificador presenta dificultades a la hora de diferenciar entre insultos genuinos y bromas entre amigos, aunque constatan que el rendimiento de la clasificación mejora cuando los tuits contienen palabras con connotación positiva, como *'love'* y *'friend'*. Todas estas dificultades ponen de manifiesto que la identificación del discurso de odio sigue siendo un problema no resuelto, lo que justifica la necesidad de continuar investigando en este ámbito.

3. DATOS

Para la recopilación de datos, Twitter fue hasta 2023 la plataforma más accesible por excelencia, debido a que proporcionaba una API y a su relativa flexibilidad en cuanto la recopilación y diseminación de datos (Abercrombie et al., 2023). Por este motivo, la mayoría de los estudios que abordan esta temática recurrieron a la plataforma. No obstante, desde que Twitter pasó a denominarse X y se discontinuó el acceso gratuito a la API, esto ha cambiado de manera considerable. Aun así, dado que gran parte de los corpus existentes proceden de Twitter, precisamente por su accesibilidad anterior, también en la presente investigación se ha decidido utilizar esta plataforma como base. Sin embargo, con el fin de añadir otra dimensión al campo de investigación, se ha optado por no limitar el análisis al inglés, que sigue siendo la lengua más representada en este ámbito, sino que incorpora también el español y el neerlandés. De este modo, se presta atención a lenguas que han recibido una cobertura más limitada, en particular el neerlandés, para el cual, hasta donde se ha podido comprobar, no existen estudios previos sobre la detección automática del discurso de odio misógino.

3.1. Recopilación de datos

Como el presente artículo es el primero que se centra en la detección automática de discurso de odio misógino en lengua neerlandesa, se ha comenzado por construir un nuevo conjunto de

datos que sirva como punto de partida. Este se ha elaborado a partir de dos listas que contienen palabrotas disponibles en el sitio web Github (bronvermelding), ampliadas con los conocimientos de la autora sobre las injurias que suelen dirigirse a las mujeres en las redes sociales. Como resultado, se ha obtenido un conjunto de 27 términos y locuciones ofensivos, que se ha sometido a una primera búsqueda exploratoria en Twitter para comprobar si generaban los tuits deseados. Por estos se entienden tuits claramente ofensivos alternados con tuits no ofensivos que eventualmente contenían palabrotas, con el fin de poder entrenar en una fase posterior un modelo capaz de distinguir adecuadamente entre ambas categorías. Basándose en esta exploración inicial, se han realizado algunas modificaciones en la lista original para obtener finalmente una lista definitiva que contiene un total de 24 palabras o locuciones clave (véase tabla 1), las cuales se han utilizado para llevar a cabo la búsqueda final de tuits.

Alle vrouwen zijn	Je bent een hoer	Slet
Allemaal hoeren	Je bent een slet	Sletterig
Allemaal sletten	Kankerhoer	Stoephoer
Bitch	Kankerslet	Teef
Griet	Kankerwif	Trut
Hoer	Kuthoer	Tyfushoer
Hoerig	Kutwif	Vrouwen moeten
Hysterisch	Rotwif	Wif

Tabla 1: Palabras clave utilizadas para la recopilación de tuits neerlandeses

Estas palabras clave se han buscado en una colección de tuits neerlandeses recopilada antes de 2023 mediante la API, de modo que el resultado fue una lista de 393 566 tuits que contenían al menos una de la palabrotas. A estos se les ha añadido la condición de que los resultados debían corresponder a palabras completas, con el fin de evitar que términos neutros como “versleten” generaran un falso positivo al contener la secuencia “slet”. De estos resultados se ha extraído una muestra arbitraria de como máximo cien ejemplos por palabra clave, generando finalmente la colección de datos definitiva, compuesta por 1920 tuits, que se ha anotado en una etapa posterior.

En lo que respecta a los datos en lengua inglesa y española, se ha optado por no elaborar un conjunto de datos completamente nuevo, sino por anotar con mayor detalle uno ya existente, ya que, como se ha expuesto anteriormente, la investigación se ha centrado hasta ahora principalmente en estas dos lenguas. Con este propósito, se ha seleccionado el corpus HatEval, desarrollado para la tarea compartida “SemEval-2019 Task 5”, puesto que constituye uno de los corpus de mayor tamaño que incluyen discurso de odio dirigido específicamente a mujeres y, además, integra ambas lenguas de interés para la presente investigación. La tarea compartida se centra tanto en el discurso de odio contra mujeres como en el discurso de odio contra inmigrantes. Si bien el corpus completo es más amplio y está compuesto por un total de 19 600 tuits, se han utilizado exclusivamente los ejemplos considerados pertinentes en el contexto del presente estudio, concretamente los datos de entrenamiento cuyo blanco son mujeres. De esta selección se ha extraído una muestra arbitraria de 1 200 tuits por lengua, teniendo en cuenta que existe un equilibrio entre los datos que contienen discurso de odio y los que no lo contienen, de manera que en ambas lenguas se disponen de 600 ejemplos por categoría

3.2. Anotación de datos

En lo que sigue, se describe el proceso de anotación en las tres lenguas, ilustrándolo mediante ejemplos concretos con el fin de facilitar su comprensión.

Como la presente investigación se desarrolla en el contexto de un trabajo de fin de máster, el conjunto de datos ha sido anotado únicamente por la autora. No obstante, con el fin de garantizar la mayor consistencia posible en las anotaciones, se han adoptado las siguientes medidas metodológicas: la realización de una segunda ronda de revisión para verificar las etiquetas asignadas, la presentación de los casos más ambiguos a investigadores con experiencia en el ámbito de la detección automática de discurso de odio misógino y la elaboración de directrices de anotación que han funcionado como referencia en situaciones de duda.

En cuanto al procedimiento de anotación, se ha seguido en gran medida el esquema utilizado en el proyecto *RAINBOW*, tal como se presenta en Cignarella et al. (2025), aunque se ha añadido la categoría de humor con el fin de adaptarlo mejor a los objetivos del presente estudio. Esta categoría resulta particularmente relevante, dado que la ironía y el lenguaje figurado son

están estrechamente vinculados con el discurso implícito. Dado que este fenómeno sigue planteando dificultades en la detección automática, su anotación puede resultar útil para el análisis posterior de errores.

El esquema final se presenta a continuación:

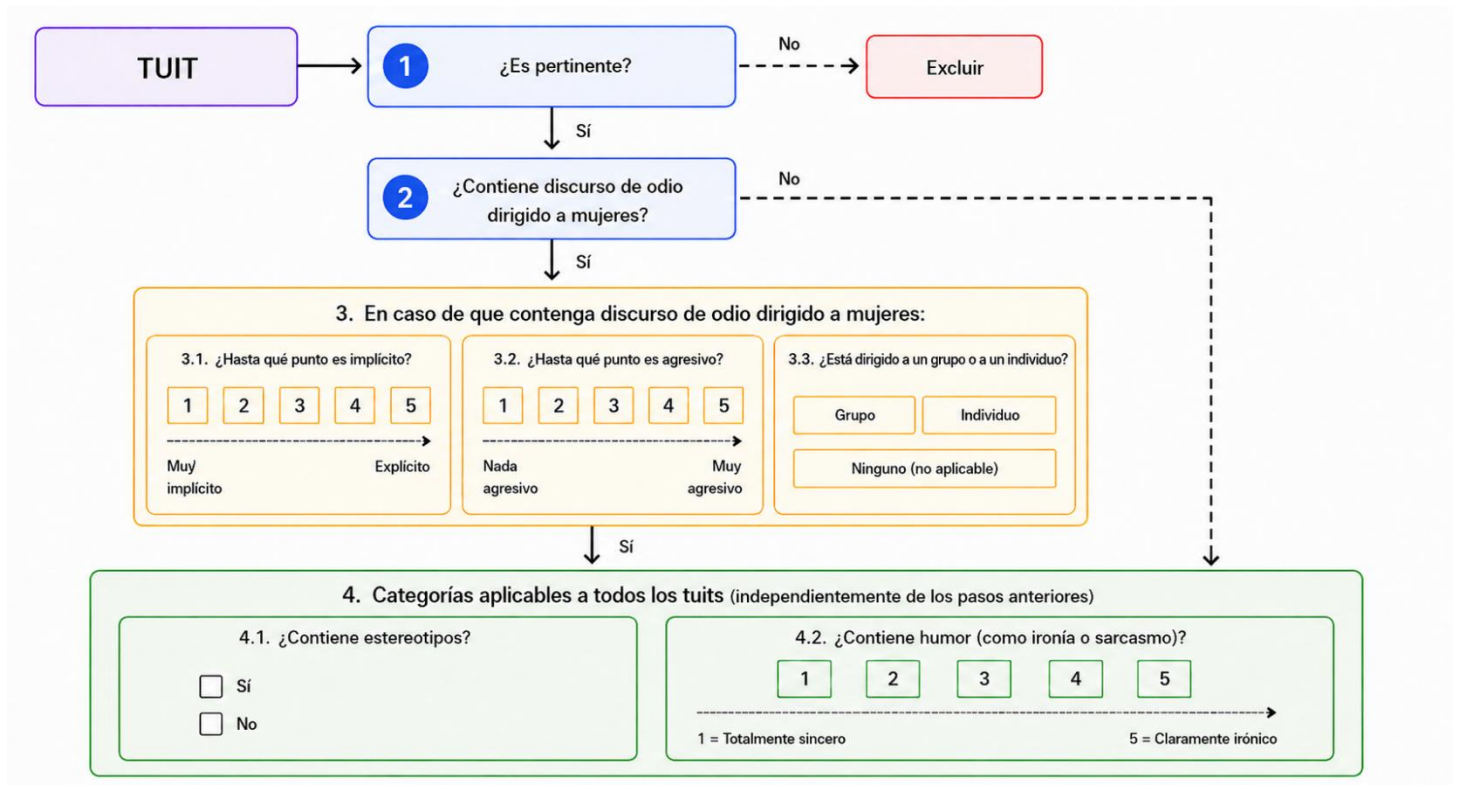


Figura 2: visualización del esquema de anotación

El procedimiento de anotación resultó prácticamente idéntico en las tres lenguas, con la única excepción de que los datos empleados procedentes de *HatEval* ya contaban con anotaciones previas para las categorías de discurso de odio, blanco meta y agresividad. Sin embargo, dado que dichas anotaciones no siempre coincidían con los criterios establecidos para el presente estudio, los tuits correspondientes fueron marcados como no pertinentes en aquellos casos en los que se consideró necesario. Esta divergencia puede explicarse, al menos en parte, por el escaso consenso existente en torno a lo que constituye exactamente el discurso de odio y, en consecuencia, por la limitada fiabilidad de este tipo de anotaciones (Bauwelinck et al., 2019). Por ello, se ha considerado pertinente explicar el procedimiento de anotación adoptado en el presente trabajo mediante ejemplos en las tres lenguas.

Tal como ocurrió en el proyecto *RAINBOW* (Cignarella et al., 2025), los tuits se consideran pertinentes cuando se dirigen a mujeres (ejemplo 3), cuando contienen opiniones, menciones, reacciones o actitudes en contra de las mujeres (ejemplo 4) o cuando incluyen palabras que suelen utilizarse para insultar a mujeres (ejemplo 5), independientemente de que sean ofensivos o no. Asimismo, cuando no puede determinarse a partir del tuit si este se dirige a una mujer o a un hombre, pero contiene una palabrota utilizada predominantemente para insultar a mujeres, el tuit también se considera pertinente.

- 3) *Tfoe je bent echt sletterig!*
- 4) *ajsjsjdk mi mamá me vino a enseñar su delineador y yo que perra que perra que perra mi amiga y se ENOJÓ wtf gabriela es buen plan*
- 5) *you guys nailed it. I've had 3 seperate fits of hysterical laughter and I didn't even make it to the school yet.*

Por el contrario, quedan excluidos de esta categoría aquellos tuits que están escritos en una lengua distinta de la del resto del conjunto (ejemplo 6), que son de carácter racista (ejemplo 7), que contienen imágenes o textos pornográficos (ejemplo 8) o que resultan incomprensibles (ejemplo 9).

- 6) @user 🧒👩👉🤔🤔 | Het ñ oproep gehad van ñ teef in Costa Rica wat sê eks haar kind se Pa en moet dadelik my pligte aanvaar of sy maak ñ saak. | Sy stuur toe ñ foto van die kind. Ek gaan dit nie eers beveg of toetse doen nie! | Pappa is erg lief vir jou my kleinding!!! 🤔 <http://someurl>
- 7) @vikidonda Cállate negra de mierda. Tu jefe político se cago en el los jubilados cuando fue jefe de gabinete. Flor de hija de puta Sos
- 8) *Can't fall asleep someone come choke me till I pass out you can do whatever you want to my body after I'm passed out #cumslut #slut #horny #wet #pussy #ass #anal #fuck #fuckme #naked #nudes #whore #bitch #naughty #cum #cumtribute #jailbait #blowjob #girls #sex #submissive*
- 9) *haha met @user ; stoepslet , stoephoer , bitch , kutwif , ditdat x*

También se excluyen los tuits dirigidos a hombres (ejemplo 10), incluso cuando incluyen simultáneamente discurso de odio dirigido a una mujer (ejemplo 11), dado que no forman parte del público meta del presente estudio. Como se ha señalado anteriormente, en el caso de los datos procedentes de *HatEval*, aquellos ejemplos cuyas anotaciones originales no se ajustaban a los criterios establecidos en estas directrices (ejemplo 12) fueron clasificados como no pertinentes, con el propósito de preservar la coherencia metodológica entre los conjuntos de datos en las tres lenguas.

- 10) *@SergioMassa Callate hijo de puta! Mi madre murió esperando que le pagues el juicio que ganó cuando vos estabas en el ANSES. No tenés moral. Caradura!!!!*
- 11) *Heard a boy joking about date rape drugs and calling women “bitches.” First of all, you look like a peeled banana. Second of all, stop*
- 12) *But the plant gets devoured by the animal, and the animal devoured by us, because nature, rather like I spoke of women before, is a cunt.*

De los tuits considerados pertinentes, se clasifican como ofensivos aquellos que, de un modo u otro, menosprecian a las mujeres, de conformidad con la definición de Nockleby (2000) presentada en la revisión bibliográfica (véase capítulo 2.1). Aunque a primera vista pueda parecer una distinción sencilla, a menudo resulta más difícil establecerla de lo que cabría esperar. Así, por ejemplo, no se consideran ofensivos los casos de discurso de odio reportado en los que el autor se distancia del comentario original mediante la burla, la crítica o un contraataque (ejemplo 13), así como aquellos en los que otra persona interviene en defensa de la víctima (ejemplo 14). Sin embargo, los tuits que contienen discurso de odio y que se limitan a reportar el hecho sin expresar una postura al respecto (ejemplo 15), o que incluso intensifican la burla hacia la víctima (ejemplo 16), sí se consideran ofensivos.

- 13) *Heel veel mensen zeggen opeens, kuthoer, jemoeder, wegslet, hoiiii & stoefhoer. Mooi dit xD*
- 14) *Las mujeres tb son libres de ligarse a los hombres. El asunto es cómo se hace. La violacion ,el acoso ,el abuso no es la forma. <https://t.co/4YA52foBA0>*
- 15) *i was called some type of whore or slut bc i was more developed. like since i was like 11 man http*
- 16) *Lo mejor del día, escuchar como mi sobrino le dijo CALLATE PUTA a mi hermana*



La categoría de tuits ofensivos abarca tanto los casos en los que el menosprecio se manifiesta de manera explícita (ejemplo 17) como aquellos en los que adopta formas más implícitas (ejemplo 18). No obstante, la delimitación entre ambas dimensiones no siempre resulta evidente, especialmente debido a que los tuits combinan con frecuencia formas de menosprecio más sutiles con insultos explícitos. Esta coexistencia dificulta su asignación a una única categoría. Por este motivo, se ha optado por emplear una escala de cinco niveles, que permite representar con mayor precisión los casos intermedios. En aquellos casos en los que el tuit presentaba rasgos tanto explícitos como implícitos (ejemplo 19), se anotó en una posición intermedia de la escala.

17) @nahimachoura @SanCadilla *Ya cállate pinche perra.*

18) *What's the difference between a blonde and a washing machine? A washing machine won't follow you around all day after you drop a load in it*

19) @user *Roelof is terecht boos. De onzin die de zgn coach uitkraamt is zo beschadigend dat 'akelig rotwijf' een vrij adequate benaming is.*

A diferencia de las categorías anteriores, determinar si un tuit es agresivo o no resulta considerablemente más sencillo, dado que los mensajes agresivos suelen manifestarse de manera explícita. También para esta categoría se ha optado por utilizar una escala de cinco niveles, con el fin de permitir una mayor matización. Así, los mensajes que contenían amenazas claras se anotaron como agresivos (ejemplo 20), aquellos que incluían una orden dirigida a una mujer se situaron en una posición intermedia (ejemplo 21), y los mensajes que se limitaban a insultar a una mujer sin amenazarla se clasificaron como no agresivos (ejemplo 22).

20) @manuroque21 @marionaquadrada *Eres una marrana asquerosa mala puta, malparida, no mereces haber nacido, mal follada, tú sí que no te entereas de nada, que vives en las cavernas, no sabes, no conoces no vives en tí, vives en un sin vivir en tí mismo, pobre*

21) *wat is dat toch een heks oprotten met dat kutwijf <http://someurl>*

22) @Tuckedofftae *No bitch this girl annoying*

Para la determinación del blanco del comentario, se ha considerado que un tuit se dirige a un grupo cuando el contenido está claramente orientado hacia las mujeres en general (ejemplo

23). Por el contrario, se ha anotado como dirigido a un individuo cuando del propio tuit se desprende claramente que el mensaje está dirigido a una persona concreta (ejemplo 24). Esta última categoría también se ha aplicado en aquellos casos en los que el tuit se dirige a varias mujeres concretas (ejemplo 25), siempre que del contenido del mensaje resulte evidente que no se trata de una referencia a las mujeres como grupo, sino a destinatarias específicas.

23) @user Paar maanden geleden... meid van 22 heeft een kat in de fik gestoken hier in de buurt. Kat is uiteindelijk na gruwelijke pijn overleden 🙄 alle vrouwen zijn kut...

24) so my Chloe don't want to see me win she unfollowed because I made a post i HIT 100 FOLLOWERS and then she unfollowed me bitch fuck u whore

25) @azucenau @navarreteprida @MarcelaGuerraNL @mileniotv ¿ Esas mujeres son las que se quejan del acoso...? ¿Sí habrá quien las acose...?

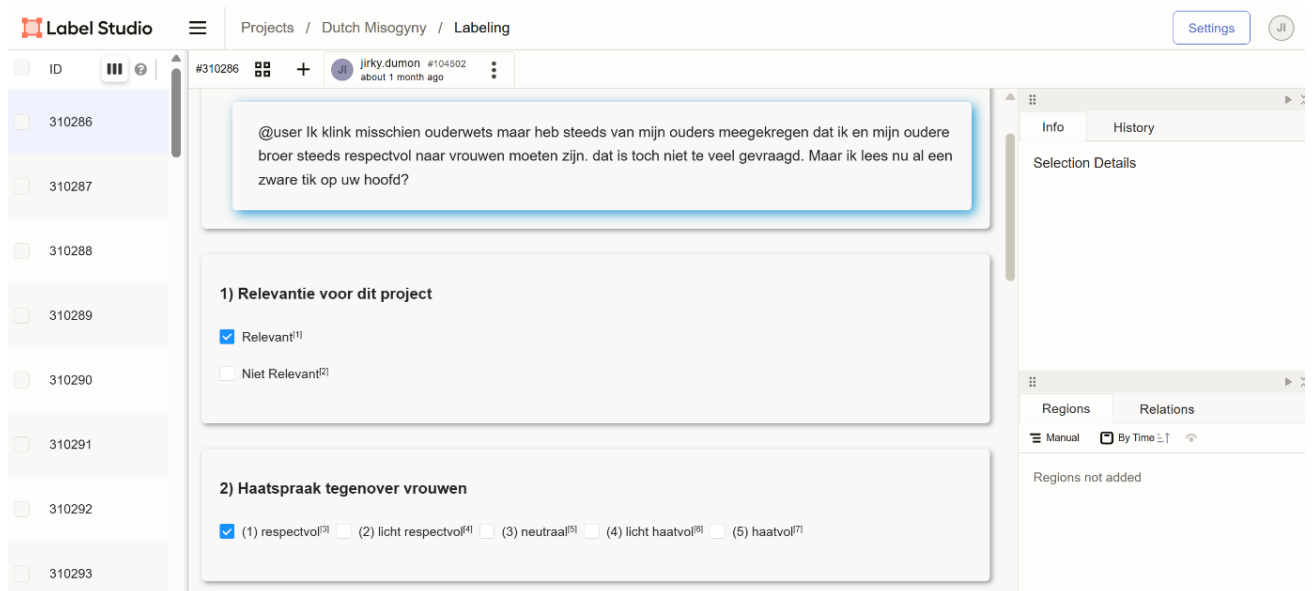
Dentro del esquema de anotación, también se ha considerado la presencia de estereotipos como una dimensión relevante, dado su papel en la construcción del discurso de odio misógino. Como señalan Cignarella et al. (2025), un aspecto a menudo pasado por alto en esta línea de investigación es el hecho de que los estereotipos suelen estar en el origen de la incitación al odio. En dicho estudio, los estereotipos se definen como “widely held, simplified, and often generalized beliefs about a particular group of people” (Cignarella et al., 2025, p. 4). Al igual que la incitación al odio, estos suelen manifestarse de forma implícita, por lo que resulta particularmente relevante incorporarlos al presente estudio. Siguiendo esta definición, se ha anotado un tuit como portador de un estereotipo sobre las mujeres cuando contiene una generalización explícita sobre ellas, como en formulaciones del tipo “vrouwen zijn” (ejemplo 26). Esta categoría también se ha aplicado a los casos en los que el tuit se dirige a una mujer concreta, pero recurre a palabras u órdenes que remiten expectativas, roles o comportamientos asociados específicamente con las mujeres (ejemplo 27).

26) @user En alle vrouwen zijn Mentaal beperkt

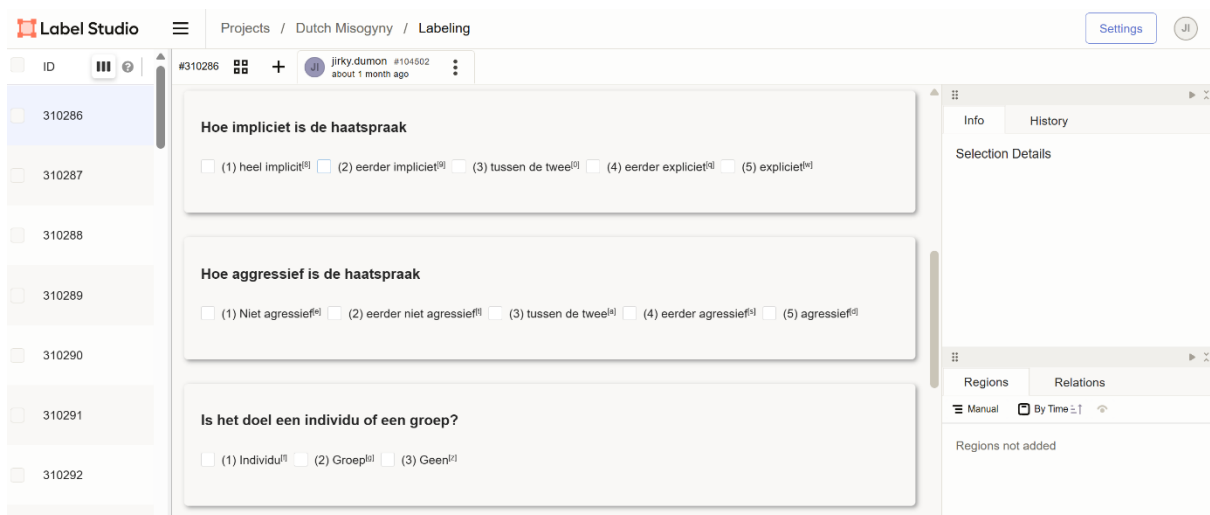
27) @Laura99CM Ahora que me has faltado el respeto te lo puedo decir por fin: ¡A fregar puta!

Finalmente, se ha anotado si los tuits del conjunto contienen algún tipo de humor, como por ejemplo sarcasmo o ironía, independientemente de si contienen discurso ofensivo y sin tomar en consideración la posible intención del autor del mensaje.

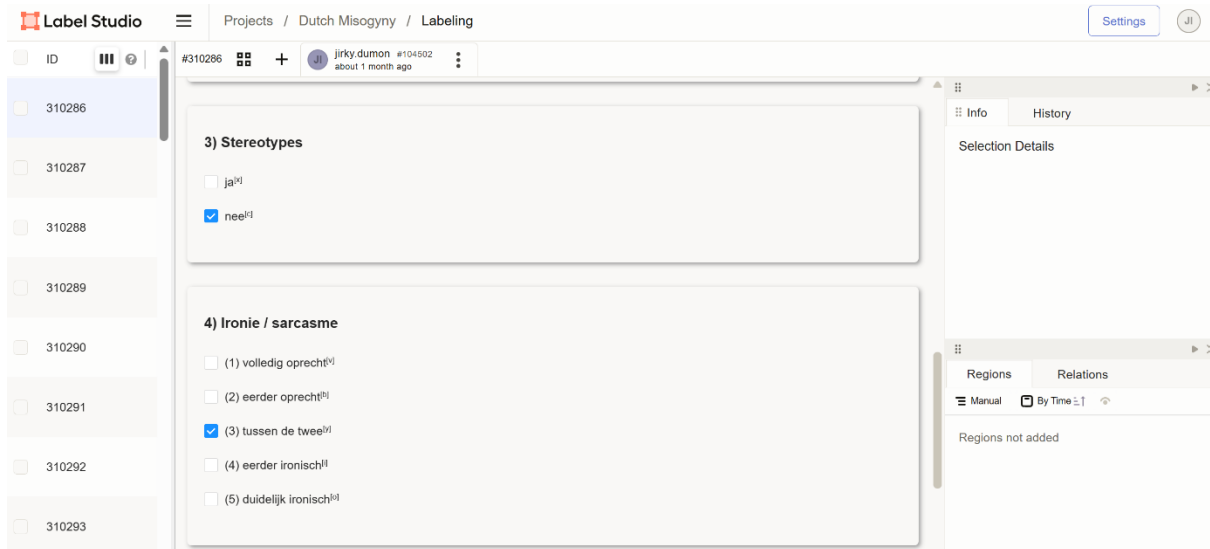
La anotación se ha llevado a cabo mediante la plataforma *Label Studio*, una herramienta de código abierto empleada para la etiquetación y anotación de datos en el ámbito del PLN. Aplicando el esquema presentado en la figura anterior, la interfaz utilizada para los datos en neerlandés presenta el aspecto siguiente:



Imágen 1: Captura De Pantalla de la Interfaz Utilizada Para Anotar Los Datos



Imágen 2: Captura De Pantalla de la Interfaz Utilizada Para Anotar Los Datos



Imágen 3: Captura De Pantalla de la Interfaz Utilizada Para Anotar Los Datos

Los conjuntos de datos y los criterios de anotación descritos en este apartado constituyen el punto de partida para la siguiente fase del estudio. A continuación, se presenta la configuración experimental adoptada para el entrenamiento y la evaluación del modelo.

4. CONFIGURACIÓN EXPERIMENTAL

Para llevar a cabo la parte experimental del presente trabajo, se ha optado por utilizar el modelo multilingüe de tipo encoder *mmBERT*. Esta elección se justifica, en primer lugar, por el hecho de que muchos estudios anteriores se basan en modelos menos recientes, mientras que *mmBERT* constituye una alternativa más actual dentro de los modelos multilingües. Además, dado que existen pocos estudios sobre la detección automática del discurso de odio en neerlandés, resulta pertinente evaluar hasta qué punto un modelo de estas características puede aplicarse a un conjunto de datos que incluye fenómenos como la implicitud, la agresividad, los estereotipos y el humor, los cuales pueden plantear dificultades específicas.

El modelo ha sido preentrenado con más de un billón de tokens en más de 1800 lenguas, lo que le permite procesar datos en las tres lenguas consideradas en este estudio dentro de una única arquitectura compartida. Por este motivo, resulta especialmente adecuado para una

investigación de carácter multilingüe, ya que permite analizar su rendimiento tanto de manera individual para cada lengua como de forma conjunta en la totalidad de los datos.

El modelo se ha afinado para la tarea de detección de discurso de odio misógino a partir de los tuits anotados, tal como se ha descrito en el apartado anterior. Solo se han empleado los datos considerados pertinentes de acuerdo con los criterios de anotación, dado que los casos no pertinentes a menudo pueden ser ofensivos, pero no estar dirigidos al grupo meta del presente estudio. Por ello, no resultaría adecuado clasificarlos simplemente como datos sin discurso de odio, ya que esto podría introducir ambigüedad en el proceso de aprendizaje del modelo. Para evaluar el rendimiento del modelo, se han aplicado tres configuraciones experimentales a los datos de cada lengua.

En la primera configuración, se ha aplicado un procedimiento de validación cruzada de cinco pliegues (*five-fold cross-validation*). Para ello, cada uno de los tres corpus se ha dividido en cinco subconjuntos de tamaño comparable. En cada iteración, el sistema se ha entrenado con cuatro de estos subconjuntos y se ha probado en el restante, de modo que cada tuit se ha utilizado una vez como dato de evaluación. La aplicación de esta estrategia se ha considerado especialmente pertinente dado el tamaño relativamente limitado de los corpus empleados en esta investigación en términos de PLN.

En la segunda configuración, el modelo se ha afinado a partir de todos los datos de *HatEval* relativos a la misoginia en inglés y en español. Cabe señalar que estos no incluyen los datos seleccionados y reanotados para el presente estudio, con el propósito de evitar que un mismo tuit aparezca tanto en la fase de entrenamiento como en la fase de evaluación.

En la tercera configuración, el modelo se ha afinado mediante una combinación de las dos configuraciones anteriores. Concretamente, se han utilizado los datos de *HatEval* relativos a la misoginia en inglés y en español, junto con la validación cruzada de cinco pliegues aplicada a los datos anotados para la lengua correspondiente.

Solo se han empleado los datos considerados pertinentes de acuerdo con los criterios de anotación, dado que los casos no pertinentes a menudo pueden ser ofensivos, pero no estar dirigidos al grupo meta del presente estudio. Por ello, no resultaría adecuado clasificarlos simplemente como datos sin discurso de odio, ya que esto podría introducir ambigüedad en el

proceso de aprendizaje del modelo. En el apartado siguiente se presentan los resultados obtenidos.

5. RESULTADOS

En la tabla 2 se presentan los resultados de la tarea binaria de detección automática de discurso de odio misógino en los tres conjuntos de datos anotados para la presente investigación, expresados mediante las métricas puntuación F1, precisión, exactitud y recuperación. Además, en la primera columna se recoge el número total de tuits mal clasificados por lengua.

	n	F1	Precisión	Exactitud	Recuperación
Neerlandés	926	0.696	78.64	75.92	88.66
Inglés	483	0.865	88.19	86.94	89.65
Español	544	0.845	90.35	86.44	89.52

***Tabla 2:** Resultados De La Clasificación Binaria Del Discurso De Odio Misógino Por Lengua*

A primera vista, se observa que el modelo presenta un rendimiento inferior en neerlandés en comparación con el inglés y el español. Esta diferencia se refleja especialmente en la puntuación F1, que es considerablemente más baja para el neerlandés. Sin embargo, los valores de recuperación se mantienen en un nivel similar en las tres lenguas, lo que indica que el modelo consigue detectar una proporción elevada de los tuits misóginos. La principal dificultad parece situarse, por tanto, en la precisión y la exactitud: en neerlandés, el sistema clasifica con mayor frecuencia como discurso de odio tuits que, según las anotaciones manuales, no pertenecen a esta categoría.

En las tablas 3 – 6 se resume, para cada categoría, tanto su frecuencia en el corpus como el número y el porcentaje de clasificaciones erróneas, lo que permite identificar con mayor claridad en qué tipos de casos falla el modelo.

	# tuits implícitos	# tuits mal clasificados	% tuits mal clasificados
Neerlandés	178	43	0.24
Inglés	149	30	0.20
Español	62	19	0.31

Tabla 3: Resultados De La Clasificación De Los Tuits Implícitos

Cabe destacar que, en esta tabla, se han incluido como tuits implícitos aquellos anotados como (1) claramente implícito, (2) más bien implícito y (3) entre los dos. Los datos muestran que este fenómeno genera un número considerable de clasificaciones erróneas en las tres lenguas. No obstante, la proporción más elevada de errores se registra en español, a pesar de que este conjunto contiene un número considerablemente menor de tuits implícitos que los otros dos. Esto indica que el alto número de tuits mal clasificados en neerlandés no puede atribuirse únicamente a este factor.

	# tuits agresivos	# tuits mal clasificados	% tuits mal clasificados
Neerlandés	112	7	0.06
Inglés	131	9	0.07
Español	475	37	0.08

Tabla 4: Resultados De La Clasificación De Los Tuits Agresivos

En cuanto a la agresividad, se han incluido únicamente los tuits anotados como (4) más bien agresivo y (5) agresivo. Los resultados muestran que esta categoría presenta su mayor

frecuencia en español, donde el número de casos es aproximadamente tres veces superior al observado en neerlandés e inglés. Esto vuelve a llamar la atención sobre el carácter más explícito y agresivo del conjunto español en comparación con los otros dos conjuntos de datos. No obstante, esta diferencia en términos de frecuencia no se traduce en un aumento significativo de errores, ya que las clasificaciones erróneas se mantienen en niveles reducidos y relativamente similares en las tres lenguas. Cabe señalar, además, que se trata de la única subcategoría analizada en la que el neerlandés obtiene el mejor resultado, por encima del inglés y del español.

	# tuits con estereotipos	# tuits mal clasificados	% tuits mal clasificados
Neerlandés	220	52	0.24
Inglés	170	20	0.12
Español	83	22	0.27

Tabla 5: Resultados De La Clasificación De Los Tuits Con Estereotipos

A diferencia de la categoría anterior, los resultados relativos a los estereotipos presentan una variación más marcada entre lenguas. El sistema alcanza el porcentaje de error más bajo en inglés, mientras que en neerlandés y español la proporción de tuits mal clasificados es considerablemente más elevada. El porcentaje más alto se registra en español, donde el 27% de los tuits que contienen estereotipos fueron clasificados erróneamente.

	# tuits humorísticos	# tuits mal clasificados	% tuits mal clasificados
Neerlandés	251	87	0.35
Inglés	127	26	0.20
Español	95	23	0.24

Tabla 6: Resultados De La Clasificación De Los Tuits Humorísticos

Tal como se hizo en el caso de los tuits implícitos, en la categoría de los tuits humorísticos se han incluido tanto los casos anotados como (3) intermedios como aquellos clasificados como (4) más bien humorísticos o (5) claramente humorísticos. El neerlandés presenta una presencia notablemente mayor de humor, con casi el doble de casos en comparación con las otras dos lenguas. Al mismo tiempo, esta lengua obtiene los resultados más bajos en dicha categoría, con una diferencia considerable respecto del inglés y el español, así como en comparación con las puntuaciones obtenidas en las demás subcategorías. Conviene subrayar, además, que esta es la única categoría en la que el rendimiento del neerlandés es inferior al del español.

6. DISCUSIÓN

Considerando que la cantidad de mensajes publicados diariamente en línea excede la capacidad de revisión manual, resulta imprescindible desarrollar métodos automatizados que permitan detectar estos contenidos de manera eficiente mediante técnicas de procesamiento del lenguaje natural. Sin embargo, estos sistemas continúan enfrentándose a una serie de problemas, entre los cuales destacan la escasa disponibilidad de datos de entrenamiento en lenguas distintas del inglés, la dificultad para reconocer el lenguaje implícito y la limitada capacidad de los investigadores para identificar el origen de los errores, debido a la falta de transparencia en el proceso de razonamiento de los sistemas.

A partir de estas limitaciones, el presente estudio interpreta los resultados obtenidos con el fin de formular una respuesta a las dos siguientes preguntas de investigación:

1. ¿En qué medida permite un modelo lingüístico multilingüe detectar el discurso de odio misógino en tuits redactados en neerlandés, inglés y español?
2. ¿Se observan diferencias entre el neerlandés, el inglés y el español en los puntos fuertes y débiles de la detección automática del discurso de odio misógino al considerar categorías discursivas como la implicitud, la agresividad, los estereotipos y el humor?

De los resultados se desprende que los datos en inglés y, sobre todo, en neerlandés contienen una proporción considerablemente mayor de tuits implícitos, estereotipados y humorísticos, mientras que la presencia de tuits agresivos es mucho más reducida. Sin embargo, estas cuatro

subcategorías parecen plantear un grado de dificultad relativamente comparable en las tres lenguas. Las principales excepciones se observan en los tuits implícitos en español y en los tuits humorísticos en neerlandés, categorías en las que el sistema muestra mayores dificultades.

En los párrafos que siguen, se analizan e interpretan de manera cualitativa los resultados descritos en el apartado anterior. Para ello, se examinan primero los patrones observados en las distintas categorías anotadas y, posteriormente, los errores de clasificación de manera general, con el fin de formular una respuesta fundamentada a las preguntas de investigación planteadas.

6.1. Análisis de errores

En muchos de los tuits mal clasificados resulta difícil determinar con precisión por qué se produjo el error, dado que suelen contener al menos un palabrota o alguna manifestación de discurso ofensivo. A ello se suma el problema de la caja negra (Vidgen et al., 2019), que impide establecer con certeza qué elementos concretos han influido en cada clasificación errónea. A continuación, se presentan algunos ejemplos de tuits clasificados erróneamente:

28) @user *Cyan magenta yellow. Wat een tyfushoer is dat zeg. Ik snap nog steeds niet precies wat ik daar nou aan moet doen. Ga je Project L echt proberen te spelen of ga je mensen cheesen totdat het niet meer werkt en dan opgeven?*

29) @user *First of all I'm not transgender but what if I was? Stupid petty ass bitch* 🍊👤

30) @user *PERRITA TU NOVIO NO TE MERECE EL ES TAN LINDO Y HERMOSO ES COMO UN OSITO DE PELUCHE MEDIO PENDEJO..PERO TU ERES UNA PUTA*

No obstante, aunque no siempre sea posible determinar con exactitud la causa de cada error, una revisión cualitativa de los tuits mal clasificados permite identificar algunos patrones recurrentes que podrían haber contribuido a las predicciones incorrectas.

6.1.1. Análisis de errores por categorías

Como de los resultados se desprende que el modelo obtiene un rendimiento más bajo en neerlandés, resulta pertinente no solo realizar un análisis general de los errores, sino también examinar las categorías individuales. Este análisis permite comprobar si alguna de ellas

contribuye de manera particular a los resultados menos favorables observados en esta lengua, en consonancia con los objetivos de la presente investigación.

6.1.1.1.Implicitud

En cuanto al carácter implícito del discurso, se han tomado en consideración los tuits anotados como (1) claramente implícito, (2) más bien implícito y (3) entre los dos en la sección de resultados, dado que se considera que las dos categorías restantes son las únicas que pueden calificarse realmente como explícitas. No obstante, resulta relevante no solo incluir los casos anotados como intermedios, sino también examinarlos por separado, ya que esta etiqueta refleja precisamente la dificultad de establecer una delimitación clara entre lo implícito y lo explícito. De este modo, se puede comprobar si el sistema clasifica estos casos de manera similar a los claramente implícitos o si, por el contrario, presentan un patrón de errores distinto.

	# tuits intermedios	# tuits mal clasificados	% tuits mal clasificados
Neerlandés	99	20	0.20
Inglés	74	11	0.15
Español	22	7	0.32

Tabla 7: Resultados De Clasificación De Los Tuits Anotados Como Intermedios (Implicitud)

Los resultados indican que el orden general entre las lenguas no cambia: los mejores resultados se siguen obteniendo en inglés, seguido del neerlandés y, finalmente, del español. Al analizar por separado los tuits anotados como intermedios, se observa que estos se clasifican con mayor acierto que los casos claramente implícitos. Una posible explicación es que este tipo de tuits suele combinar elementos implícitos y explícitos en un mismo mensaje, lo que podría facilitar parcialmente su identificación por parte del sistema. Esta tendencia se observa en neerlandés e inglés, pero no en español, y no parece haber una explicación evidente para esta diferencia. Por tanto, al menos en neerlandés e inglés, los datos no sugieren que las categorías intermedias sean las más problemáticas para el sistema. Más bien, los errores parecen aumentar a medida que el contenido se vuelve más implícito. En este sentido, puede señalarse que el

reconocimiento del discurso implícito plantea dificultades especialmente en español, aunque también se registran numerosos errores de este tipo en las otras lenguas.

6.1.1.2. Agresividad

En la categoría de los tuits agresivos, se han incluido únicamente aquellos anotados como (4) más bien agresivo y (5) agresivo, dejando fuera los casos intermedios, a diferencia de lo que se hizo en la categoría anterior. Esta decisión se debe a que se anotaron como intermedios principalmente aquellos casos que contenían una orden, como *cállate*, los cuales, a juicio de la autora, pueden considerarse más bien no agresivos en una clasificación binaria, dado que no incitan a causar daño físico a otra persona.

Los resultados muestran con claridad que esta categoría es la más fácil de distinguir en las tres lenguas, ya que es la única en la que el margen porcentual de error se mantiene por debajo del 10 %. Otro aspecto destacable es que, en este caso, el mejor resultado no se obtiene en inglés, sino en neerlandés. Una posible explicación es que los tuits agresivos contienen, en la mayoría de los casos, amenazas o insultos explícitos, lo que facilita su detección (Malmasi & Zampieri, 2017). Dado que el corpus neerlandés fue recopilado mediante búsquedas basadas en palabrotas, procedimiento que puede favorecer la presencia de tuits agresivos, este factor podría contribuir al elevado rendimiento observado en esta categoría. Sin embargo, esta explicación no resulta suficiente, ya que el español contiene más del triple de tuits agresivos que las otras lenguas sin que ello se traduzca en un rendimiento claramente superior.

6.1.1.3. Estereotipos

Se observa que los resultados difieren considerablemente entre las lenguas, lo que puede deberse, al menos en parte, a que la formulación lingüística de los estereotipos varía de una lengua a otra (Schmeisser-Nieto et al., 2024). En los conjuntos de datos analizados en la presente investigación, los estereotipos en neerlandés aparecen con frecuencia vinculados a valoraciones sobre el comportamiento o la apariencia de las mujeres, a menudo expresadas mediante términos como *hoerig* o *sletterig*. En inglés, por otro lado, suelen adoptar con mayor frecuencia la forma de bromas relacionadas, por ejemplo, con la inteligencia. En español, se observa una presencia relativamente marcada de estereotipos asociados a los roles domésticos,

como ocurre en expresiones que remiten a cocinar, limpiar o servir al hombre. Estas diferencias pueden contribuir a explicar por qué el grado de dificultad varía entre lenguas. No obstante, esta interpretación debe matizarse, dado que el humor también constituye un fenómeno difícil de reconocer para los sistemas automáticos (véase más adelante), por lo que no permite explicar plenamente el mejor rendimiento observado en inglés.

Aunque en inglés esta categoría no parece constituir una fuente de error tan importante, en neerlandés y español se registra una proporción considerablemente más elevada de clasificaciones erróneas. Este puede deberse a que los estereotipos no siempre adoptan formas abiertamente agresivas, sino que tienden a expresarse mediante formulaciones implícitas y normalizadas que perpetúan visiones jerárquicas y sesgadas sobre las mujeres (Cignarella et al., 2025). Por ello, su detección exige que el sistema no se limite a identificar insultos o palabrotas, sino que también sea capaz de reconocer formas menos explícitas de menosprecio hacia las mujeres. Además, se ha observado que algunos tuits incorporan estereotipos no para reproducirlos, sino para cuestionarlos o criticarlos. Esta postura crítica suele manifestarse mediante preguntas, emoticonos o repeticiones irónicas, lo que introduce una ambigüedad adicional en la clasificación. En consecuencia, el sistema debe determinar si el estereotipo se reproduce de manera afirmativa o si se emplea con una intención crítica.

6.1.1.4. Humor

Tal como se hizo con los tuits implícitos, también en la categoría de humor se han tomado en consideración los casos anotados como (3) entre los dos, (4) más bien humorístico y (5) claramente humorístico. Asimismo, los tuits intermedios se han examinado por separado, con el fin de comprobar si esta posición ambigua dentro de la escala da lugar a un comportamiento distinto en la clasificación del sistema.

	# tuits intermedios	# tuits mal clasificados	% tuits mal clasificados
Neerlandés	107	30	0.28
Inglés	76	13	0.17
Español	70	16	0.23

Tabla 7: *Resultados De Clasificación De Los Tuits Anotados Como Intermedios (Humor)*

A diferencia de lo observado en la categoría de los tuits implícitos, en el caso del humor el orden entre lenguas sí se modifica al considerar únicamente los casos intermedios. El inglés sigue obteniendo los mejores resultados, pero el español ocupa ahora la segunda posición, mientras que el neerlandés presenta el rendimiento más bajo. Asimismo, los resultados para los tuits intermedios son ligeramente inferiores a los obtenidos para la categoría de humor en su conjunto, aunque la diferencia se sitúa en torno al 5%. Esto indica que, en el presente estudio, la ambigüedad de estos casos también se refleja en una mayor dificultad para el sistema de clasificación, aun no muy grande. seguido del español en vez del neerlandés, que en este caso presenta el peor rendimiento.

Tanto en estos resultados como en los descritos anteriormente para la categoría de humor en general se observa un número considerable de errores en las tres lenguas, un fenómeno que también se ha señalado en investigaciones anteriores. Así, por ejemplo, el estudio de MacAveney et al. (2019) evidenció que el uso del sarcasmo y de las metáforas constituía la principal causa de la clasificación errónea de tuits. Maladry et al. (2024) explican que esto se debe al hecho de que la ironía se basa a menudo en una contradicción que puede manifestarse en dos componentes literales, o bien en un único componente que entra en conflicto con el conocimiento que poseemos del contexto o con el sentido común. Al no disponer por sí solo de esta información, al sistema le resulta difícil inferir el significado subyacente a partir de una expresión lingüística determinada, que, en el caso del discurso de odio, adopta con frecuencia una forma irónica. Como resultado, este tipo de mensajes puede no ser detectado correctamente por el modelo.

6.1.2. Patrones generales de error

En términos generales, el sistema parece otorgar un peso considerable a la presencia de determinadas palabrotas o insultos, una tendencia que también se ha señalado en otros estudios del ámbito y que puede explicarse por el estrecho vínculo entre el lenguaje ofensivo y el discurso de odio (Malmasi y Zampieri, 2017). Esta sensibilidad léxica puede favorecer que los tuits que contienen este tipo de términos sean clasificados como discurso de odio, incluso cuando no cumplen plenamente los criterios de anotación establecidos. Como ilustran los ejemplos siguientes, este fenómeno se manifiesta con especial claridad en relación con ciertos

términos dirigidos a mujeres, como *griet* o *wijf* en neerlandés, *hysterical* o *whore* en inglés y *puta* en español. Este patrón resulta relevante, además, porque permite contextualizar varias de las subcategorías de falsos positivos que se examinan a continuación.

31) *Yeahhhhh. Superbe #griet. This is #hockey #Goducsgo*

32) *@user You must listen to my dad wrote a porno - it will make your shifts hysterical*

33) *Ni perra , ni puta, ni zorra nada de esa mierda LA MUJER ES MUJER Y SE RESPETA ASI TENGA MIL NOVIO Y SE DIVIERTA CON ELLOS COMO SE LE DE LA HPTA GANA :) (Y)*

La primera subcategoría se relaciona con las limitaciones del sistema para considerar plenamente el contexto pragmático en el que se produce el mensaje, lo que dificulta la distinción entre insultos empleados con intención ofensiva e insultos utilizados de manera jocosa en interacciones entre amigos. Esta dificultad queda reflejada en los ejemplos incluidos a continuación. En Bauwelinck et al. (2019) se observa que el modelo tiende a clasificar correctamente estos casos cuando el tuit, aun conteniendo lenguaje ofensivo, incluye también palabras con connotación positiva, como *love* o *friend*, o cuando aparece en el contexto de una felicitación de cumpleaños. No obstante, esta tendencia no parece reproducirse en la presente investigación.

34) *RT @user Ik vind @user echt een rotwijf, begint altijd te praten als ik nét ga slapen • Hahahaahah laaaat mij rotzak*

35) *Happy halloween! Hope everyone lives up to their full hoe potential tonight! I believe in you, you bust ass super skanks* ❤️ 🍑

36) *@user Me olvidé un millón de cosas, pero bueno, tu novia es muy doris. TE AMO MUCHO, ZORRA. ME ENCANTAS DEMASIADO.*

Cabe mencionar, sin embargo, que en algunos casos la intención del tuit resulta tan ambigua que incluso para anotadores humanos puede ser difícil clasificarlo de manera inequívoca. Por ello, resulta comprensible que el modelo cometa errores en este tipo de casos.

De forma similar, el sistema clasifica como discurso de odio algunos enunciados ofensivos dirigidos al propio autor del mensaje. Sin embargo, de acuerdo con los criterios de anotación

establecidos para este estudio, estos casos no se consideran ofensivos, dado que no parecen estar orientados a causar daño a otra persona.

37) *Ik ben een gemeen rotwijf, een intens slechte moeder... Maar ik kan hier om gieren*

  <http://someurl>

38) *@user Um excuse you I am a twank, a twink and a skank*

39) *Fui buena con un montón de personas que no se lo merecian, a partir de hoy voy a ser mucho más hija de puta con todos*

Otra categoría de errores corresponde a los tuits que contienen discurso reportado. Dado que su clasificación depende de si el mensaje se limita a reproducir el contenido original, si defiende a la víctima o si intensifica el acoso hacia ella, el sistema debe ser capaz de distinguir entre matices muy sutiles. Esto suele resultar difícil porque el modelo no siempre capta la postura del autor frente al contenido citado o reproducido, por lo que puede confundir una denuncia o una crítica del discurso de odio con una reproducción ofensiva del mismo. A continuación, se presentan algunos ejemplos de este tipo de casos.

40) *@user Af en toe uit impulsiviteit iets ongelukkigs eruit flappen is 1, maar op z'n consistente wijze de ene na de andere leugen eruit persen, mensen uitschelden voor o.a. kankerslet en dat maar scharen onder reageren uit emotie is 2 t/m 10 zögezegd #mankementenshow*

41) *@user Telemarketer did same to me called me a whore and said he would rape me. Nice.*

42) *Lo mejor del día, escuchar como mi sobrino le dijo CALLATE PUTA a mi hermana*

Una consecuencia directa de que el sistema atribuya un peso considerable a la presencia de palabrotas es que su ausencia puede dar lugar a falsos negativos, una dificultad también señalada por Dinakar et al. (2011). Así, por ejemplo, los mensajes que contienen estereotipos claramente denigrantes, pero no insultos explícitos, suelen pasar desapercibidos para el modelo, lo que puede apreciarse en los casos presentados a continuación.

- 43) *Maar effe serieus, vrouwen moeten natuurlijk gewoon het huishouden doen, daar zijn ze voor gemaakt. Beetje zorgen en stofferen enzo. En dat moet ze blijven doen voor de man die na een leven lang hard werken aan zijn welverdiende rust toe is.* <http://someurl>
- 44) *@user rape is no joke and something we do not support, the kitchen needs to be cleaned by the way.*
- 45) *@user Movimiento SEMEN! A fregar todas!* <http>

6.2. Interpretación de los resultados

Los mejores resultados del modelo se obtienen claramente en inglés, tanto en la evaluación general como en la mayoría de las subcategorías, donde esta lengua presenta casi siempre el rendimiento más elevado. Le siguen el español y, finalmente, el neerlandés, aunque en el caso de estas dos últimas lenguas la diferencia no resulta tan clara como con respecto al inglés. En términos generales, el español obtiene mejores resultados que el neerlandés, sobre todo debido al elevado número de falsos positivos observado en los datos neerlandeses. Sin embargo, dado que el modelo sí logra reconocer una parte considerable de los casos de discurso de odio misógino en neerlandés, esta misma tendencia no se reproduce necesariamente en el análisis por subcategorías. En efecto, en todas las subcategorías, salvo en la detección automática de humor, el neerlandés presenta mejores resultados que el español.

Esto conduce a una respuesta matizada a las preguntas de investigación planteadas, a saber: ¿en qué medida permite un modelo lingüístico multilingüe detectar el discurso de odio misógino en tuits redactados en neerlandés, inglés y español? y ¿se observan diferencias entre el neerlandés, el inglés y el español en los puntos fuertes y débiles de la detección automática del discurso de odio misógino al considerar categorías discursivas como la implicitud, la agresividad, los estereotipos y el humor? A partir del análisis cuantitativo, puede constatar que los puntos fuertes y débiles del modelo son, en términos generales, similares en las tres lenguas, aunque se manifiestan con distinta intensidad. En otras palabras, las categorías que resultan más o menos problemáticas tienden a coincidir entre los tres conjuntos de datos, pero el porcentaje de errores asociado a cada una de ellas varía, especialmente entre el neerlandés y el español.

Una posible explicación de esta tendencia es que el modelo lingüístico utilizado, es decir, mmBERT, haya sido preentrenado con una cantidad considerablemente mayor de datos en inglés y en español que en neerlandés. Por ello, es posible que reconozca mejor determinados patrones lingüísticos en estas dos lenguas, independientemente de su capacidad específica para detectar discurso de odio.

7. CONCLUSIÓN

El presente estudio ha evaluado la detección automática del discurso de odio misógino en tuits neerlandeses, ingleses y españoles mediante un modelo lingüístico multilingüe de tipo encoder, mmBert. Como parte de esta evaluación, se ha examinado el papel de distintas subcategorías discursivas en los errores de clasificación. Los resultados muestran que el sistema obtiene el mejor rendimiento en inglés, seguido muy de cerca por el español y, finalmente, por el neerlandés, que presenta una puntuación F1 considerablemente más baja. El análisis cuantitativo indica que los puntos fuertes y débiles del modelo presentan una tendencia general similar en los tres conjuntos de datos. Las subcategorías que resultan más o menos problemáticas tienden a coincidir, aunque la proporción de errores asociada a cada una de ellas varía según la lengua, especialmente al comparar el neerlandés y el español.

Aunque los resultados obtenidos para el inglés y el español son relativamente altos, sigue siendo pertinente continuar investigando sistemas más eficaces, con el fin de reducir al mínimo la circulación de tuits ofensivos.

Para avanzar hacia este objetivo, una posible línea de investigación futura consiste en trabajar con un mayor número de anotadores. La anotación de datos mediante subcategorías como la ironía es una tarea inherentemente subjetiva (Maladry et al., 2024), por lo que resulta relevante comprobar el grado de acuerdo entre anotadores. De este modo, se puede garantizar, en la medida de lo posible, que las categorías estén suficientemente delimitadas para los seres humanos antes de trasladar la tarea al sistema.

Además, sería pertinente trabajar con conjuntos de datos de mayor tamaño. Cuanto más amplio sea el conjunto de datos, mayores serán las posibilidades de que el sistema generalice adecuadamente y reconozca patrones relevantes, lo que podría traducirse en mejores resultados.

Esto resulta especialmente importante para las subcategorías analizadas, ya que un corpus más amplio permitiría contar con un mayor número de instancias por categoría. En el presente estudio, algunas de ellas están representadas por menos de cien casos, lo que puede dificultar que el sistema aprenda sus rasgos característicos de manera fiable.

Finalmente, también resulta relevante explorar formas de construir conjuntos de datos que no dependan exclusivamente de la búsqueda mediante palabras ofensivas, por ejemplo, recopilando tuits a partir de perfiles de usuarias conocidos por publicar contenido ofensivo en la red. Aunque la búsqueda de palabrotas sigue siendo una práctica habitual, puede dar lugar a una recopilación desproporcionada de tuits ofensivos. Como consecuencia, el sistema puede aprender a asociar en exceso la presencia de palabrotas con el discurso de odio, lo que puede favorecer la sobregeneralización y generar un mayor número de falsos positivos.

Cabe mencionar que en la presente investigación también se intentó recopilar datos a partir de temas más generales susceptibles de generar controversia, como el aborto, con el fin de incluir casos menos explícitos. Sin embargo, resultó difícil encontrar términos de búsqueda que fueran lo suficientemente implícitos como para captar formas menos explícitas de misoginia y, al mismo tiempo, lo suficientemente precisos como para recuperar un número suficiente de casos de discurso de odio. Por esta razón, finalmente se decidió continuar trabajando con palabras clave más o menos ofensivas. Aunque esta tarea resulta compleja, sigue siendo fundamental continuar desarrollando métodos que contribuyan a reducir al mínimo el contenido misógino en línea. Solo así podrá avanzarse hacia espacios digitales más seguros para las mujeres.

BIBLIOGRAFÍA

- Abercrombie, G., Jiang, A., Gerrard-abbott, P., Konstas, I., & Rieser, V. (2023). Resources for Automated Identification of Online Gender-Based Violence: A Systematic Review. En Y.-L. Chung, P. Röttger, D. Nozza, Z. Talat, & A. Mostafazadeh Davani (Eds.), *The 7th Workshop on Online Abuse and Harms (WOAH)* (pp. 170-186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.woah-1.17>
- Ali, S., Blackburn, J., & Stringhini, G. (2025). *Evolving Hate Speech Online: An Adaptive Framework for Detection and Mitigation* (arXiv:2502.10921). arXiv. <https://doi.org/10.48550/arXiv.2502.10921>
- Athoillah, M., & Putri, R. K. (2024). Utilizing Support Vector Machines to Detect Hate Speech on Social Media. *Science, Engineering and Technology*, 4(2), 53-60. <https://doi.org/10.54327/set2024/v4.i2.129>
- 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. En J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 54-63). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S19-2007>
- Bauwelinck, N., Jacobs, G., Hoste, V., & Lefever, E. (2019). LT3 at SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval). En J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 436-440). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S19-2077>
- BOE-A-1977-9008 *Real Decreto-ley 24/1977, de 1 de abril, sobre libertad de expresión.* (s. f.). Recuperado 12 de marzo de 2025, de <https://www.boe.es/buscar/doc.php?id=BOE-A-1977-9008>
- BOE-A-1995-25444 *Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal.* (s. f.). Recuperado 12 de marzo de 2025, de <https://www.boe.es/buscar/act.php?id=BOE-A-1995-25444>
- Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for Abusive Language Detection in English. En A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 17-25). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.woah-1.3>
- sensor-text. (s.f.). *Profanity-list/list/nl.txt at main · sensor-text/profanity-list*. GitHub. Recuperado 26 de mayo de 2026, de <https://github.com/sensor-text/profanity-list/blob/main/list/nl.txt>
- Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, 71-80. <https://doi.org/10.1109/socialcom-passat.2012.55>

- Chikersal, P., Poria, S., & Cambria, E. (2015). SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado. <https://doi.org/10.18653/v1/s15-2108>
- Cignarella, A. T., Giachanou, A., & Lefever, E. (2026). A survey on stereotype detection in natural language processing. *ACM COMPUTING SURVEYS*, 58(5). <https://doi.org/10.1145/3770754>
- Clarke, C., Hall, M., Mittal, G., Yu, Y., Sajeev, S., Mars, J., & Chen, M. (s. f.). *Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection*. <https://doi.org/10.48550/arXiv.2307.12935>
- Dinakar, K., Reichart, R., & Lieberman, H. (2011, 21 de julio). Modeling the detection of textual cyberbullying [Ponencia de conferencia]. Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Cataluña, España. <https://roireichart.com/wp-content/uploads/2023/03/3841-16937-1-pb.pdf>
- Gao, L., y Huang, R. (2017). Detecting Online Hate Speech Using Context Aware Models. En R. Mitkov y G. Angelova (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)* (pp. 260-266). Incoma Ltd. https://doi.org/10.26615/978-954-452-049-6_036
- Ghorbanpour, F., Dementieva, D., & Fraser, A. (2025). Can prompting LLMs unlock hate speech detection across languages? A zero-shot and few-shot study. En *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)* (pp. 413 – 425). Association for Computational Linguistics. <https://aclanthology.org/2025.woah-1.39/>
- Gitari, N. D., Zhang, Z., Damien, H., & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4), 215-230. <https://doi.org/10.14257/ijmue.2015.10.4.21>
- Goenaga, I., Atutxa, A., Gojenola, K., Casillas, A., de Ilarraza, A. D., Oronoz, M., Perez, A., & de Vinaspre, O. P. (2018). *Automatic Misogyny Identification Using Neural Networks*. https://ceur-ws.org/Vol-2150/AMI_paper1.pdf
- Guillaume, P., Duchêne, C., y Dehak, R. (2022). Hate Speech and Toxic Comment Detection Using Transformers. En *Conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC)* (2022). https://www.researchgate.net/publication/360085544_Hate_Speech_and_Toxic_Comment_Detection_using_Transformers
- hate speech* noun—Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com. (s. f.). Recuperado 12 de marzo de 2025, de <https://www.oxfordlearnersdictionaries.com/definition/english/hate-speech?q=hate+speech>
- Hatzivassiloglou, V., & McKeown, K. R. (s. f.). *Predicting the Semantic Orientation of Adjectives*. Ibrohim, M. O., & Budi, I. (2019). Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. *Proceedings of the Third Workshop on Abusive Language Online*.

- Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy.
<https://doi.org/10.18653/v1/w19-3506>
- Hewitt, S., Tiropanis, T., & Bokhove, C. (2016). The problem of identifying misogynist language on Twitter (and other online social spaces). En *Websci '16: Proceedings of the 8th Conference on Web Science* (pp. 333-335). Association for Computing Machinery.
<https://doi.org/10.1145/2908131.2908183>
- Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232.
<https://doi.org/10.1016/j.neucom.2023.126232>
- Jiang, A., Yang, X., Liu, Y., & Zubiaga, A. (2021). *SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection* (arXiv:2108.03070). arXiv. <https://doi.org/10.48550/arXiv.2108.03070>
- Kim, Y., Park, S., & Han, Y.-S. (2022). Generalizable Implicit Hate Speech Detection Using Contrastive Learning. En N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6667-6679). International Committee on Computational Linguistics. <https://aclanthology.org/2022.coling-1.579/>
- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*, 10(4), 150.
<https://doi.org/10.3390/info10040150>
- Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1), 1621-1622.
<https://doi.org/10.1609/aaai.v27i1.8539>
- MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8), e0221152.
<https://doi.org/10.1371/journal.pone.0221152>
- Maladry, A., Cignarella, A. T., & Lefever, E., Van Hee, C., y Hoste, V. (2024). Human and System Perspectives on the Expression of Irony: An Analysis of Likelihood Labels and Rationales. En N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, y N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 8372-8382). ELRA y ICCL.
<https://aclanthology.org/2024.lrec-main.734/>
- Malmasi, S., & Zampieri, M. (2017). Detecting Hate Speech in Social Media. En R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017* (pp. 467-472). INCOMA Ltd.
https://doi.org/10.26615/978-954-452-049-6_062
- Markov, I., y Daelemans, W. (2022). The Role of Context in Detecting the Target of Hate Speech. En R. Kumar, A. K. Ojha, M. Zampieri, S. Malmasi, y D. Kadar (Eds.), *Proceedings of the Third Workshop on Threat, Agression and Cyberbullying (TRAC 2022)* (pp. 37-42). Association for Computational Linguistics. <https://aclanthology.org/2022.trac-1.5/>

- Menini, S., Apro시오, A. P., & Tonelli, S. (2021). *Abuse is Contextual, What about NLP? The Role of Context in Abusive Language Annotation and Detection* (arXiv:2103.14916). arXiv. <https://doi.org/10.48550/arXiv.2103.14916>
- Dinakar, K., Reichart, R., & Lieberman, H. (2011, 21 de julio). Modeling the detection of textual cyberbullying [Ponencia de conferencia]. Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Cataluña, España. <https://roireichart.com/wp-content/uploads/2023/03/3841-16937-1-pb.pdf>
- Nina-Alcocer, V. (2018). *AMI at IberEval2018 Automatic Misogyny Identification in Spanish and English Tweets*. https://ceur-ws.org/Vol-2150/AMI_paper8.pdf
- Nirmal, A., Bhattacharjee, A., Sheth, P., & Liu, H. (2024). *Towards Interpretable Hate Speech Detection using Large Language Model-extracted Rationales* (arXiv:2403.12403). arXiv. <https://doi.org/10.48550/arXiv.2403.12403>
- Not All Words Are Created Equal: Extracting Semantic Orientation as a Function of Adjective Relevance. (s. f.). En K. Voll & M. Taboada, *Lecture Notes in Computer Science* (pp. 337-346). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-76928-6_35
- Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). Toxicity Detection: Does Context Really Matter? En D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4296-4305). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.396>
- Pitsilis, G. K., Ramampiaro, H., y Langseth, H. (2018). Detecting Offensive Language in Tweets Using Deep Learning. https://www.researchgate.net/publication/322577160_Detecting_Offensive_Language_in_Tweets_Using_Deep_Learning
- ¿Qué es el aprendizaje automático (ML)? | IBM. (2021, septiembre 22). <https://www.ibm.com/mx-es/think/topics/machine-learning>
- ¿Qué es el aprendizaje profundo? | IBM. (s. f.). Recuperado 25 de mayo de 2025, de <https://www.ibm.com/mx-es/think/topics/deep-learning>
- ¿Qué es el PLN (procesamiento del lenguaje natural)? | IBM. (2021, septiembre 23). <https://www.ibm.com/es-es/think/topics/natural-language-processing>
- ¿Qué son los clasificadores Naive Bayes? | IBM. (2021, octubre 6). <https://www.ibm.com/es-es/think/topics/naive-bayes>
- Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., & Plaza, L. (2020). Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access*, 8, 219563-219576. <https://doi.org/10.1109/ACCESS.2020.3042604>
- Röttger, P., Nozza, D., Bianchi, F., & Hovy, D. (2022). Data-Efficient Strategies for Expanding Hate Speech Detection into Under-Resourced Languages. *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, 5674-5691.
<https://doi.org/10.18653/v1/2022.emnlp-main.383>

Schmeisser-Nieto, W. S., Cignarella, A. T., Bourgeade, T., Frenda, S., Ariza-Casabona, A., Laurent, M., Cicirelli, P. G., Marra, A., Corbelli, G., Benamara, F., Bosco, C., Moriceau, V., Paciello, M., Patti, V., Taulé, M., & D'Errico, F. (2025). StereoHoax: A multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*, 59(3), 2031-2069. <https://doi.org/10.1007/s10579-024-09791-3>

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain. <https://doi.org/10.18653/v1/w17-1101>

Shushkevich, E., & Cardiff, J. (2019). Automatic Misogyny Detection in Social Media: A Survey. *Computación y Sistemas*, 23(4), 1159-1164. <https://doi.org/10.13053/cys-23-4-3299>

Sidharth GN. (2022, octubre). The RBF kernel in SVM: A complete guide. Quark Machine Learning. <https://www.quarkml.com/2022/10/the-rbf-kernel-in-svm-complete-guide.html>

Strossen, N. (2016). Freedom of Speech and Equality: Do We Have to Choose? *Journal of Law and Policy*, 25(1), 185-225. <https://brooklynworks.brooklaw.edu/jlp/vol25/iss1/7>

Tanev, H. (2024). JRC at ClimateActivism 2024: Lexicon-based Detection of Hate Speech. En A. Hürriyetoglu, H. Tanev, S. Thapa, y G. Uludoğan (Eds.), *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)* (pp. 85-88). Association for Computational Linguistics. <https://aclanthology.org/2024.case-1.11/>

The RBF kernel in SVM: A Complete Guide. (2025, abril 6). *Quark Machine Learning*. <https://www.quarkml.com/2022/10/the-rbf-kernel-in-svm-complete-guide.html>

Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., Suieueva, N., & Kerimbekov, M. (2023). Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods. *International Journal of Advanced Computer Science and Applications*, 14(5). <https://doi.org/10.14569/ijacsa.2023.0140542>

tommasoc80. (s. f.). *DALC/GROF_LEX/groflex.tsv at master · tommasoc80/DALC*. GitHub. Recuperado 26 de mayo de 2026, de https://github.com/tommasoc80/DALC/blob/master/GROF_LEX/groflex.tsv

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, 126(1), 157-179. <https://doi.org/10.1007/s11192-020-03737-6>

Turney, P. D. (2001). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 417. <https://doi.org/10.3115/1073083.1073153>

- Twitter Usage Statistics—Internet Live Stats*. (s. f.). Recuperado 12 de marzo de 2025, de <https://www.internetlivestats.com/twitter-statistics/>
- Van Hee, C., & Hoste, V. (2024). *Taaltechnologie ontrafeld*. Pelckmans. <http://hdl.handle.net/1854/LU-01J0RGC1ABW7A5YGMSJAGDBZYR>
- Van Hee, C., Lefever, E., & Hoste, V. (2018). We Usually Don't Like Going to the Dentist: Using Common Sense to Detect Irony on Twitter. *Computational Linguistics*, 44(4), 793-832. https://doi.org/10.1162/coli_a_00337
- Vapnik, V., Golowich, S. E., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. *Proceedings of the 10th International Conference on Neural Information Processing Systems, NIPS'96*, 281-287. <https://dl.acm.org/doi/10.14778/3489496.3489503>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. <https://doi.org/10.48550/arXiv.1706.03762>
- Vaughan, J. W. y Wallach, H. (2021). A Human-Centered Agenda for Intelligible Machine Learning. En M. Pelillo y T. Scantamburlo (Eds.), *Machines We Trust: Perspectives on Dependable AI* (pp. 123-138). The MIT Press. <https://doi.org/10.7551/mitpress/12186.003.0014>
- Velankar, A., Patil, H., & Joshi, R. (2022). *A Review of Challenges in Machine Learning based Automated Hate Speech Detection*. arXiv.Org. <https://arxiv.org/abs/2209.05294v1>
- Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *Proceedings of the Third Workshop on Abusive Language Online*. Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy. <https://doi.org/10.18653/v1/w19-3509>
- Warner, W., y Hirschberg, J. (2012). Detecting Hate Speech on the World Wide Web. En S. O. Sood, M. Nagarajan, y M. Gamon (Eds.), *Proceedings of the Second Workshop on Language in Social Media* (pp. 19-26). Association for Computational Linguistics. <https://aclanthology.org/W12-2103/>
- Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*. Proceedings of the NAACL Student Research Workshop, San Diego, California. <https://doi.org/10.18653/v1/n16-2013>