Identificación automática de lenguaje ofensivo en redes sociales contra la comunidad LGBTIQ+

Automatic identification of offensive language on social media against the LGBTIQ+ community

Por Jesús Donovan Maldonado Mondragón, Asdrúbal López-Chau, Maricela Quintana López, Saturnino Job Morales Escobar

Resumen: Este artículo explora la importancia de desarrollar modelos de identificación de lenguaje ofensivo en español, cuyo objetivo es proteger a la comunidad LGBTIQ+ y garantizar que las plataformas digitales sean espacios de respeto y aceptación. El análisis que aquí presentamos se enfoca en identificar términos y expresiones que, en su contexto, resultan discriminatorias u hostiles, lo cual permite una detección más precisa de aquellas palabras o frases que pueden tener un impacto negativo en las personas afectadas.

Palabras clave: comunidad LGBT, aprendizaje supervisado, lenguaje ofensivo.

Abstract: This article explores the importance of developing models for identifying offensive language in Spanish, which aims to protect the LGBTIQ+ community and ensure that digital platforms are spaces of respect and acceptance. The analysis presented here focuses on identifying terms and expressions that, in context, are discriminatory or hostile, which allows a more accurate detection of those words or phrases that may have a negative impact on the people affected.

Keywords: LGBT community, supervised learning, offensive language.

En las redes sociales, el lenguaje puede ser una poderosa herramienta de alianza o de agresión. Para la comunidad LGBTIQ+, el lenguaje ofensivo en plataformas digitales representa una amenaza que va más allá de un simple comentario. Estas expresiones pueden perpetuar prejuicios, promover el odio y crear barreras que impiden la libre participación en entornos virtuales (Vikas *et al.* 2020). La identificación y mitigación de este tipo de lenguaje no es solo un reto técnico, sino una necesidad para construir comunidades virtuales inclusivas y seguras.

EL LENGUAJE OFENSIVO Y SU IMPACTO

Algunas palabras o frases aparentemente inofensivas pueden ser usadas de manera intencional para causar un efecto perjudicial en quienes las reciben. Esta discriminación virtual, aunque intangible, tiene efectos psicológicos reales que fomentan el aislamiento, reducen la autoestima y generan desconfianza hacia los entornos digitales. Por tanto, reducir el lenguaje ofensivo es fundamental para crear espacios de respeto e inclusión.

La identificación de lenguaje ofensivo es especialmente importante para aquellos que buscan interactuar y compartir sus experiencias en línea. A través de la implementación de sistemas de detección, se busca reducir la exposición de la comunidad LGBTIQ+ a contenido dañino, y con ello contribuir a un ambiente digital donde puedan expresarse sin temor de ataques personales o discriminación.

METODOLOGÍA PARA LA DETECCIÓN DE LENGUAJE OFENSIVO

La detección de lenguaje ofensivo en español se basa en técnicas de aprendizaje supervisado (Sengupta *et al.*, 2022) en las que los modelos de inteligencia artificial aprenden a clasificar texto mediante conjuntos de datos previamente etiquetados. Durante este proceso, los términos ofensivos se identifican y contextualizan, permitiendo que el modelo "aprenda" a distinguir entre lenguaje ofensivo, neutral y no ofensivo.

Para aumentar la precisión, se utilizan lexicones especializados que contienen términos y frases específicas que pueden resultar ofensivas en el idioma español (Bashar *et al.*, 2021). Estos lexicones incluyen variaciones de palabras, modismos y expresiones regionales que permiten una mayor sensibilidad a los distintos usos del español en diferentes contextos. Además, el análisis de sentimientos complementa la identificación al evaluar el tono general del mensaje y detectar la hostilidad oculta en frases aparentemente neutras.

ALGORITMOS DE CLASIFICACIÓN Y SU EFICIENCIA

Los algoritmos de clasificación, como *support vector machines* (SVM) y redes neuronales (NN, por sus siglas en inglés), son comunes en la identificación de lenguaje discriminatorio y violento. Las SVM resultan particularmente eficaces en la clasificación de datos, pues establecen límites claros entre categorías (Chih-Wei *et al.*, 2016), lo que permite distinguir con

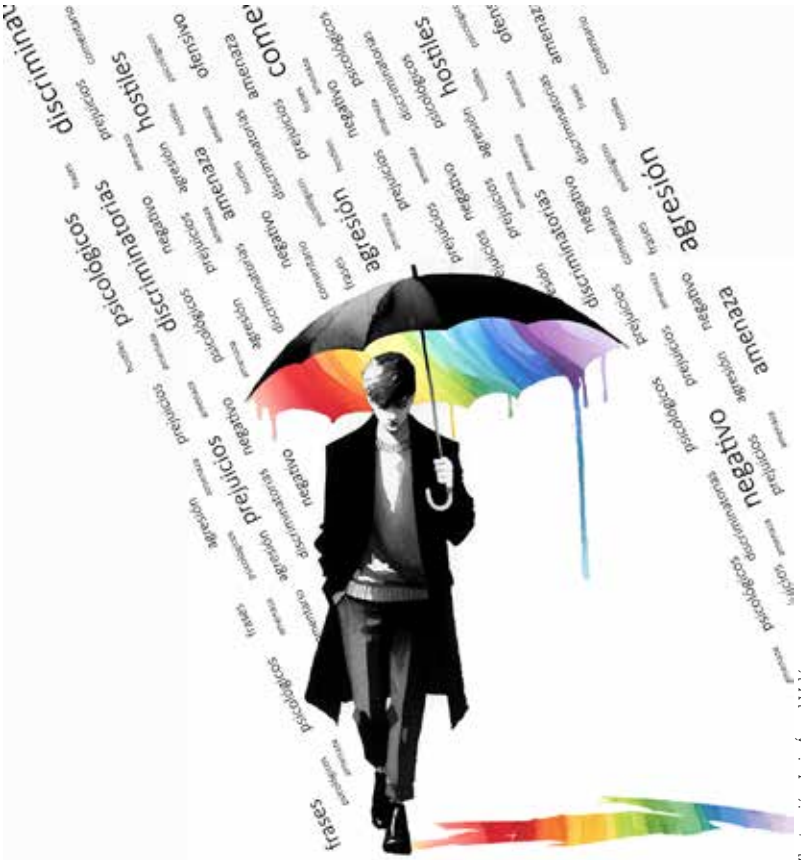

Ilustración: Luis Ángel Velázquez

exactitud las expresiones hostiles de las neutras. Las NN interpretan las palabras de manera matizada gracias al análisis de las relaciones complejas entre sí (Bashah y Kalita, 2020). También se utilizan otros métodos de clasificación para este propósito, como bosques aleatorios, clasificadores bayesianos y regresión logística.

Para fines de este estudio, se evaluó un lexicón especializado como complemento de varios algoritmos supervisados. Al realizar la ampliación del vocabulario disponible para el análisis, se logró una mejora de aproximadamente 15% a diferencia de la implementación exclusiva de algoritmos de aprendizaje supervisado.

Como resultado, el clasificador que obtuvo los mejores resultados fue el de bosques aleatorios, con valores entre 0.84 y 0.94 en términos de F1-Score. Esto demuestra que la implementación de un lexicón generado específicamente para este contexto mejora la precisión en la identificación de ofensas. Además, esta mejora está estrechamente relacionada con la cantidad de datos disponibles para su creación; cuantos más datos se tengan, mejor será la identificación y más amplio será el vocabulario utilizado.

Aunque es posible aplicar modelos más complejos, como los de la IA generativa (ChatGPT, Gemini, Copilot, Claude, etc.), estos son computacionalmente costosos y menos escalables para este problema, donde la cantidad de comentarios a procesar alcanza miles o millones por minuto.

IMPACTO POSITIVO Y POTENCIAL TRANSFORMADOR

Los modelos de identificación de lenguaje ofensivo en español son herramientas valiosas para fomentar una cultura de respeto en los espacios digitales. La detección de este tipo de lenguaje no solo protege a las personas directamente afectadas, sino que también envía un mensaje claro sobre los valores de respeto e inclusión en el ámbito digital. Para la comunidad LGBTIQ+, esto significa que pueden participar y expresarse libremente sin temor a ser atacados. Si bien existen desafíos, como la ambigüedad y el sesgo, el desarrollo continuo y la mejora de estos modelos pueden contribuir significativamente a la creación de entornos inclusivos. Con el tiempo, estos avances en inteligencia artificial tienen el potencial de transformar las redes en espacios donde la diversidad y el respeto sean los valores principales.

Referencias

Basemah Alshemali y Jugal Kalita (2020). "Improving the Reliability of Deep Neural Networks in NLP: A Review", en *Knowledge-Based Systems,* vol. *191*, 0950-7051.

Bashar, Md Abul, Richi Nahyak, Khanh Luong y Thirunavukarasu Balasubramaniam (2021). "Progressive domain adaptation for detecting hate speech on social media with small training set and its application to COVID-19 concerned posts", en *Social Network Analysis and Mining*, vol. 11, núm.1, 69.

Chih-Wei Hsu, Chang Chih-Chung y Lin Chih-Jen (2016). "A Practical Guide to Support Vector Classification", en *Department of Computer Science National Taiwan University*.

Sengupta, Ayan *et al.* (2022). "Does aggression lead to hate? Detecting and reasoning offensive traits in hinglish code-mixed texts", en *Neurocomputing,* vol. 488, 598-617.

Vikas Kumar Jha *et al.* (2020). "DHOT-Repository and Classification of Offensive Tweets in the Hindi Language", en *Procedia Computer Science*, vol. 171, 2324-2333.

Jesús Donovan Maldonado Mondragón es Profesor de Asignatura en el Centro Universitario UAEM Valle de México, de la Universidad Autónoma del Estado de México. Arquitecto de software especializado en aplicaciones empresariales y de comercio electrónico. Sus investigaciones actuales se centran en el procesamiento de lenguaje natural y la aplicación de modelos de inteligencia artificial para el análisis de sentimientos.

Asdrúbal López-Chau es profesor de tiempo completo en el Centro Universitario UAEM Zumpango, de la Universidad Autónoma del Estado de México. Integrante del Sistema Nacional de Investigadoras e Investigadores Nivel 1 desde 2015, y miembro del Research Laboratory in Engineering and Applied Sciences del CU UAEM Zumpango. Sus investigaciones actuales se centran en el diseño y aplicación de modelos de inteligencia artificial en problemas nacionales.

Maricela Quintana López es Profesora de Tiempo Completo y Coordinadora de Estudios Avanzados del Centro Universitario UAEM Valle de México, de la Universidad Autónoma del Estado de México. Sus líneas de investigación son Minería de Datos e Inteligencia Artificial Aplicada.

Saturnino Job Morales Escobar es profesor de tiempo completo en el Centro Universitario UAEM Valle de México, de la Universidad Autónoma del Estado de México. Candidato a Investigador del Sistema Nacional de Investigadoras e Investigadores y miembro del cuerpo académico de Inteligencia Computacional del CU Valle de México. Sus líneas de trabajo incluyen soluciones tecnológicas basadas en Inteligencia Artificial, Reconocimiento de Patrones y Procesamiento de Señales.