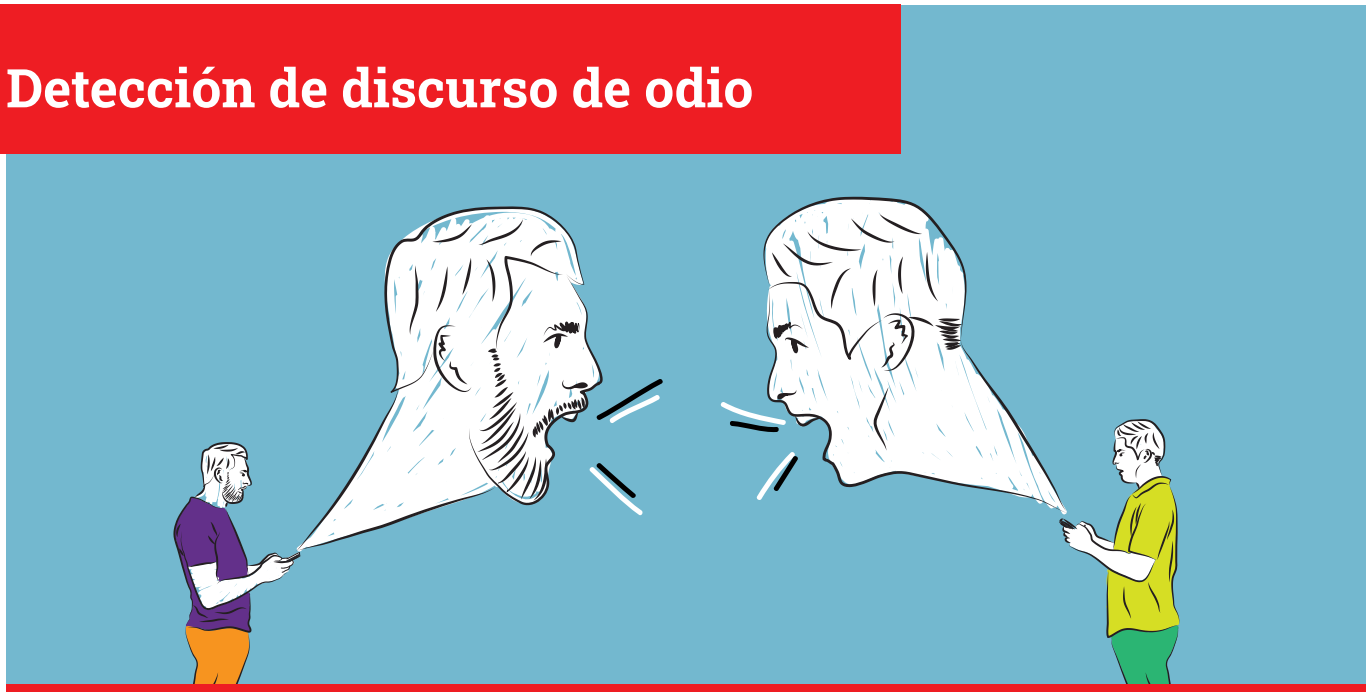
Detección de discurso de odio



AYMÉ ARANGO Estudiante de Doctorado del Departamento de Ciencias de la Computación de la Universidad de Chile

Las redes sociales se han convertido en un medio importante de interacción entre usuarios de todo el mundo. El contenido compartido puede ser de gran utilidad, como fuente de información inmediata que permite el análisis de eventos, estudio de fenómenos, la difusión de arte, ciencia, entre otras. Junto con esta información, también se encuentran manifestaciones de ciertos fenómenos comunicacionales como noticias falsas y discurso de odio que pueden producir efectos colaterales dañinos.

A pesar de que hay cierta discrepancia en cómo definir el término "discurso de odio", una de las definiciones más usadas es: expresiones derogatorias a individuos o grupos atendiendo a cierta característica como color de la piel, origen étnico, género, orientación sexual, entre otros.[1] La propagación de este tipo de contenido en los medios digitales tiene como efectos la molestia e intimidación de los usuarios. En casos extremos puede trascender el ámbito virtual y llegar a ocasionar daños físicos en individuos. Estudios recientes han encontrado vínculos entre el odio en las redes y los crímenes de odio [1]. Desde diversas disciplinas se trabaja para entender y tratar de identificar a tiempo este fenómeno.

Revisar el contenido publicado consiste en una ardua tarea para los proveedores de redes sociales. Debido al gran flujo de datos a analizar en un red social, y a su variedad, se requieren técnicas automatizadas para detectar este tipo de contenido y tomar medidas necesarias a tiempo. Dada la complejidad de la tarea, esto no ha podido lograrse satisfactoriamente hasta el momento.

Desde el punto de vista de la ciencia de datos, la detección de discurso de odio puede ser planteada como un problema de clasificación en el cual la entrada es un mensaje (*tweet*, comentario, fotografía, etc.) y la salida es la clasificación de éste como contenido odioso o no. Sin embargo, algunos investigadores consideran categorías más específicas y construyen modelos capaces de predecir el tipo específico de odio que está siendo expresado, como sexismo, racismo, xenofobia, entre otros.

Técnicas de inteligencia artificial se han venido utilizando para intentar resolver este problema. Específicamente, los modelos de aprendizaje automático han sido ampliamente utilizados como herramientas en la detección de discurso de odio [2, 3], incluyendo, en los últimos años, modelos basados en arquitecturas de redes neuronales [4]. Para que tales modelos "aprendan" a diferenciar el contenido "odioso" del contenido "normal", se necesitan datos previamente etiquetados. Idealmente, estos datos deberían contener ejemplos representativos de los diferentes tipos de expresiones de odio existentes. Obtener este tipo de datos etiquetados es costoso y debido a la información sensible que manejan y a políticas de cada

1 | https://www.encyclopedia.com/international/encyclopedias-almanacs-transcripts-and-maps/hate-speech.

plataforma, muy pocos conjuntos de datos son públicos y la mayoría son pequeños.[2] Adicionalmente, algunos de los conjuntos de datos publicados han sido reportados como sesgados [5], lo que reduce las posibilidades de utilizar datos de calidad, y como consecuencia, de construir buenos detectores de discurso de odio.

Como parte de mi tesis doctoral, junto con los profesores Bárbara Poblete y Jorge Pérez, estamos investigando técnicas para la construcción de modelos que sean generalizables a diferentes idiomas. Tal y como sucede en otras tareas relacionadas con el Procesamiento del Lenguaje Natural, la mayoría de los modelos desarrollados hasta el momento han sido principalmente explotados para resolver el problema en el idioma inglés. Como consecuencia, la gran parte de los recursos construidos son de utilidad solamente para este idioma, mientras la tarea avanza más lentamente para el resto. Analizando dos de los mejores modelos reportados en la literatura de idioma Inglés [6], encontramos que los resultados mostrados estaban sobreestimados debido a problemas experimentales, y uso de datos sesgados. Además, estos modelos presentan una pobre generalización a datos en el mismo idioma inglés y a datos en español.

Siendo el odio en medios digitales un fenómeno del cual hay evidencia a lo largo de todo el mundo, se requieren soluciones efectivas en los distintos idiomas para afrontar el problema. La idea de nuestro enfoque es aprovechar los recursos existentes (mayormente en inglés) y construir modelos generalizables a diferentes idiomas, ahorrando así el esfuerzo necesario en la creación de nuevos recursos para cada idioma separadamente. Para que los modelos de aprendizaje automático sean capaces de transferir conocimiento de un idioma a otro, se requieren representaciones de los datos a través de un conjunto de características que puedan ser comunes para diferentes idiomas. Ejemplo de esto pueden ser representaciones vectoriales multilingües o información que no esté directamente relacionada con un idioma específico. Particularmente, nuestro equipo de investigación ha trabajado en encontrar dichas características que sean comunes al odio en diferentes idiomas que nos permitan construir modelos generalizables. Bajo nuestro foco de atención, se encuentran aquellas representaciones que puedan ser extraídas del contexto del mensaje, del autor del mensaje (meta-información) y que por su naturaleza no estén atadas a un único idioma [7]. Además, estamos interesados en construir representaciones específicas para el lenguaje de odio, siendo este un fenómeno con características especiales donde ciertas palabras o expresiones pueden tomar connotaciones de odio, en dependencia del contexto. Dichas expresiones no son únicas y pueden depender no sólo del idioma, sino del contexto cultural en el que se exprese. Nos interesaría resaltar estas diferencias culturales en aras de construir modelos que generalicen mejor.

Este tipo de generalización presenta aún varios retos debido a las diferentes características de los idiomas y a la complejidad que puede tener la tarea, siendo el odio un fenómeno no sólo lingüístico, sino social y cultural. Definitivamente, todavía hay mucho que investigar en esta área. Los resultados aún no son concluyentes respecto a qué modelo o representación de datos resulta mejor para esta tarea y aunque se han logrado algunos avances, la tarea aún está por resolverse. ■

REFERENCIAS

[1] Williams ML, Burnap P, Javed A, Liu H, Ozalp S. Hate in the machine: anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. Br J Criminol (2020), 60(1), pp. 93–117.

[2] Anzovino, M., Fersini, E., and Rosso, P. Automatic Identification and Classification of Misogynistic Language on Twitter. In International Conference on Applications of Natural Language to Information Systems (2018), Springer, pp. 57–64.

[3] Papegnies, E., Labatut, V., Dufour, R., and Linares, G. Graph-based Features for Automatic Online Abuse Detection. In International Conference on Statistical Language and Speech Processing (2017), Springer, pp. 70–81.

[4] Gambäck, B., and Sikdar, U. K. Using Convolutional Neural Networks to Classify Hate-Speech. In Proceedings of the First Workshop on Abusive Language Online (2017), Association for Computational Linguistics, pp. 85–90.

[5] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the Association for Computational Linguistics (2019), pp. 1668–1678.

[6] Arango, A., Pérez, J., Poblete, B.: Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (2019), ACM, pp. 45–54.

[7] Arango, A., Pérez, J., & Poblete, B. Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation (extended version). Information Systems , 101584 (2020).

2 | https://github.com/aymeam/Datasets-for-Hate-Speech-Detection.