Spanish

La posibilidad de monitorear el contenido de odio en línea a partir de lo que escribe la gente se está convirtiendo en un asunto muy importante para varios actores, como gobiernos, empresas de TIC y profesionales de ONG's que implementan campañas de sensibilización en respuesta al preocupante aumento de los abusos y de la incitación al odio en línea. Al mismo tiempo, la detección automática del lenguaje abusivo (más conocido como *abusive language*) es un tema de creciente interés en el campo del Procesamiento del Lenguaje Natural (PLN), especialmente si el objetivo es identificar diversas formas de odio en las publicaciones de las redes sociales. El *abusive language* es un término genérico que se utiliza para definir los contenidos hostiles generados por usuarios, que intimidan o incitan a la violencia y al desprecio, dirigiéndose a grupos vulnerables en las redes sociales. Hoy en día, estos contenidos están muy extendidos, y se encuentran también en otros tipos de textos como los artículos y títulos de periódicos online.

La importancia de comprender y detectar automáticamente el discurso de odio se debe al aumento de las manifestaciones de actos violentos vinculados a conductas abusivas en línea, como el ciberacoso, el racismo, el sexismo y la homofobia. Se han implementado varios enfoques en los últimos años para apoyar la identificación y el monitoreo de estos fenómenos, lamentablemente estos están lejos de resolver el problema debido a la complejidad interna del lenguaje abusivo y las dificultades para detectar sus formas más implícitas.

En nuestra investigación de doctorado, hemos examinado las cuestiones relacionadas con la identificación automática del lenguaje abusivo en línea, investigando las diferentes maneras de hostilidad contra las mujeres, los inmigrantes y las comunidades culturales minoritarias, en idiomas como el italiano, el inglés y el español. El marco multilingüe nos ha permitido tener un enfoque comparativo para reflexionar sobre cómo se expresa el discurso de odio en varios idiomas, y cómo dichas expresiones se deben representar en el proceso automático del texto. El análisis de los resultados de los distintos métodos de clasificación de los mensajes en relación con la presencia del lenguaje abusivo, ha sacado a la luz algunas dificultades principalmente vinculadas a sus manifestaciones más implícitas. Por ejemplo, en los casos en que se utilizan figuras retóricas (como la ironía y el sarcasmo), cuando se fortalecen ideologías (como la ideología sexista) o esquemas cognitivos (como los estereotipos), o cuando se postulan contrarias a un tema de discusión.

Para abordar estas dificultades, hemos propuesto distintas soluciones que también se pueden aplicar a diferentes géneros textuales. En particular, hemos observado que los aspectos cognitivos y creativos del discurso del odio son más difíciles de deducir automáticamente de los textos. Al mismo tiempo, también son elementos muy recurrentes como el caso del sarcasmo un recurso retórico que tiende a

socavar la precisión de los sistemas. De hecho, por sus peculiaridades, el sarcasmo es adecuado para enmascarar mensajes ofensivos, especialmente en textos muy breves e informales como los publicados en Twitter. Nuestra hipótesis es que al informar al sistema sobre la presencia del sarcasmo, se mejoraría la identificación de los mensajes de odio, incluso cuando estos están disfrazados de sarcásticos. Para ello, es interesante estudiar cómo la introducción de conocimientos lingüísticos en modelos de detección puede ser útil para capturar los niveles de significado más implícitos.

Según la bibliografía retórica, el sarcasmo se considera una forma particular de ironía. De acuerdo con nuestro enfoque lingüístico y multilingüe, examinamos expresiones irónicas en contenido generado por usuarios en italiano y español, revelando los rasgos más universales del lenguaje irónico. Sobre esta base, nos hemos centrado en el italiano para validar nuestra hipótesis e investigar las características específicas del sarcasmo en contextos sensibles, como los debates en línea sobre temas sociales por ejemplo, la inmigración.

En concreto, hemos creado nuevos recursos que nos permitieron profundizar en nuestra hipótesis y desarrollar diversos enfoques para identificar dos maneras de lenguaje abusivo en tuits y títulos de periódicos: los discursos de odio (o *hate speech*) y los estereotipos. Nuestra idea es combinar de manera fructífera el conocimiento general de los modelos lingüísticos y la información lingüística obtenida mediante la extracción de elementos lingüísticos específicos o entrenando simultáneamente el sistema al reconocimiento del lenguaje irónico en una arquitectura multitarea. Los resultados experimentales confirman que hacer que los sistemas sean conscientes del sarcasmo mejora el reconocimiento del discurso de odio y los estereotipos en los textos de las redes sociales, como los tuits. Al informarles de elementos lingüísticos específicos, se vuelven más sensibles a la identificación de estereotipos tanto en los tuits como en los títulos de periódicos.

Los *corpora* utilizados en nuestros experimentos se propusieron como referencia para tareas compartidas en dos ediciones de EVALITA, la campaña de evaluación de herramientas de PLN para el italiano, ayudando a crear un nuevo estado del arte para estas tareas de detección en italiano. Además, el marco multidisciplinario y multilingüe de nuestros análisis nos permitió reflexionar también sobre los límites entre aspectos más generales y dominios más específicos que a menudo se superponen en los enfoques computacionales para identificar el discurso del odio y los fenómenos relacionados.