Algoritmos sofisticados para prevenir ciberdelitos

Mediante un novedoso modelo de inteligencia artificial, un grupo de investigadores del programa de Matemáticas Aplicadas y Ciencias de la Computación de la URosario logra identificar focos de posibles delitos informáticos en las redes sociales.

Por Mauricio Veloza
Fotos Milagro Castro

El 25 de mayo de 2020, en Minneapolis (Estados Unidos), el afroamericano **George Floyd** fue asesinado por un policía. Su muerte generó una gran ola de protestas en Estados Unidos y en el mundo entero que buscaban evidenciar los desmanes policiales y la segregación racial. La consigna "las vidas negras importan" repotenció el gran movimiento contra el racismo policial ***Black Lives Matter***.

La muerte de Floyd tuvo amplia resonancia en el planeta gracias a la difusión de mensajes por las redes sociales, en especial Twitter que fue 'inundando' con reflexiones, acusaciones y demandas con sentido político.

"Fuimos muy oportunistas. Justamente cuando estaba sucediendo el movimiento de *Black Lives Matter* encontramos que la mayoría de los mensajes provenían de esa red social, por eso los *hashtags* asociados a ese fenómeno se convirtieron en insumos principales para nuestra investigación", explica **Daniel Díaz López. "Twitter es una red social donde se muestra mucho inconformismo y donde se ven múltiples posiciones políticas"**, continúa.

Díaz López es profesor del programa de pregrado y de la **Maestría en Matemáticas Aplicadas y Ciencias de la Computación** de la **Escuela de Ingeniería, Ciencia y Tecnología** de la Universidad del Rosario y lidera el estudio **Desarrollo de capacidades de inteligencia cibernética para la prevención del delito** —aún no finalizado— que busca investigar mecanismos avanzados para la protección del ciberespacio y la prevención de ciberdelitos.

En otras palabras, su objetivo es **analizar la tecnología emergente para identificar a las personas que la utilizan con el propósito de cometer delitos en el ciberespacio, e identificar maneras para contrarrestar estas acciones.**

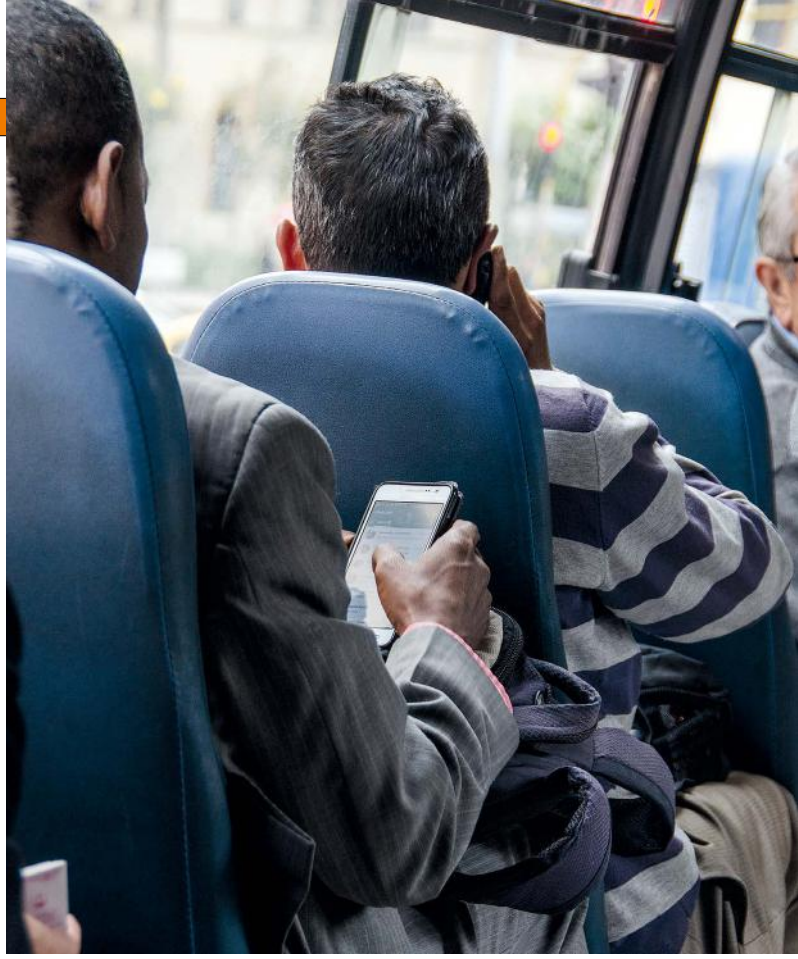
Un punto importante en esta investigación lo constituyen las redes sociales porque resultan ser el lugar propicio para expresar mensajes de odio y violencia que pueden dar paso a la ejecución de un delito. También porque son 'tribuna' para la información falsa que por sí misma puede ser una acción punible.

En el caso de Floyd, los investigadores analizaron 1.287 tuits, pero también se decidieron revisar otras movilizaciones sociales como las ocurridas en Colombia a finales de 2019. En ese momento en Twitter se publicaron informaciones falsas que generaron pánico e incertidumbre en buena parte de la población, como lo demostró gran parte de los 1.081 tuits analizados en este caso.

Uno de los casos de desinformación estudiado fue aquel donde se usó el *hashtag* ***#DCblackout***, que partió de una cuenta con solo tres seguidores y se convirtió en tendencia en poco tiempo. Difundió información falsa sobre una interrupción generalizada de comunicaciones en Washington, D.C., que provocó graves trastornos en esta ciudad y generó múltiples actos violentos.

"Hay una situación real y desafortunada: la internet carece de soberanía. Muchas veces las personas en redes sociales se sienten con más libertad o capacidad para lanzar amenazas o generar noticias falsas. Todo ello en un entorno físico sería mucho más difícil de hacer, porque

Daniel Díaz López, profesor en Matemáticas Aplicadas y Ciencias de la Computación de la Escuela de Ingeniería, Ciencia y Tecnología de la Universidad del Rosario, lidera una investigación para construir una solución basada en inteligencia artificial (IA), específicamente para el procesamiento del lenguaje natural, que apoye a las entidades del Estado en la prevención de ciberdelitos.



allí hay unas restricciones sociales mucho más fuertes. En el escenario digital, la gente es más 'valiente'; además está protegida por el anonimato", asegura el profesor Díaz López.

Desde su perspectiva, lo que se ve con frecuencia en redes sociales es que se generan movimientos para la promoción de odios y de violencia. Eso —dice— no es otra cosa que terrorismo, ya que se entiende como la generación de terror en la sociedad. El ciberterrorismo, por lo tanto, es esa proliferación de terror en el ciberespacio.

"Las redes sociales se convierten en un escenario muy interesante para analizar porque es ahí donde se gestan movimientos supremacistas blancos o que buscan incitar al sabotaje masivo contra los cuerpos de seguridad. Es allí donde crecen", afirma.

Por eso, junto con **Julián Ramírez** y **Alejandra Campo Archbold,** estudiantes del programa de pregrado en Matemáticas Aplicadas y Ciencias de la Computación, y **Julián Aponte Díaz**, oficial de la Armada Nacional de Colombia, decidió adelantar la investigación.

El objetivo de los profesionales es construir una solución basada en inteligencia artificial (IA), específicamente para el procesamiento del lenguaje natural, que apoye a las entidades del Estado en la prevención de ciberdelitos.

En líneas generales, la investigación propone el uso de un modelo de similitud fundamentado en el Procesamiento de Lenguaje Natural (NLP por su sigla en inglés) para monitorear las actividades sospechosas en las redes sociales. A través de este modelo, una agencia de seguridad del Estado puede buscar publicaciones que sean similares, identificar sospechosos y de esta manera anticiparse a la materialización o promoción de delitos cibernéticos.

> Los investigadores aplicarán el modelo a temas específicos como las fotomultas, la generación de pánico en las finanzas o cualquier tema que genere susceptibilidades y se exprese con determinadas emociones en las redes sociales, particularmente en Twitter donde la opinión de las personas queda al descubierto.

Así funciona el modelo de similitud

El NLP es un área de la IA que busca construir soluciones capaces de interpretar el lenguaje humano. Por ejemplo, cuando se le da un comando de voz al celular, él tiene que reconocer las palabras que se le dicen. "Esa función de reconocimiento utiliza NLP, que no es otra cosa que darle la capacidad a un modelo matemático de entender lo que una persona expresa", explica el profesor Díaz López.

Por su parte, Campo Archbold comenta que la investigación se basó en el ciclo de la ciencia de datos: primero se analizó el contexto y el estado del problema, luego se adquirieron los datos, enseguida se creó el modelamiento y, finalmente, se hizo el despliegue. "Extrajimos los datos a través de una aplicación que identifica etiquetas y efectuamos el preprocesamiento para aplicar luego el modelo de similitud", anota.

La primera tarea que cumplieron fue una reorganización de los tuits: descartaron los que tenían palabras confusas o mal escritas, limpiaron y organizaron los datos. Luego hicieron una vectorización, es decir, convirtieron las palabras en números. "Esto nos sirvió para crear un modelo de similitud que permitiera asociar tuits. En el proceso de depuración clasificamos tuits positivos y tuits negativos", asegura Ramírez.

De esa forma, emplearon conjuntos de datos con más de 500.000 vocablos que indicaban si había una intencionalidad

positiva o negativa. Así fueron entrenando el modelo para que identificara qué palabra es positiva o negativa. "Siempre hay un rango de error, que va disminuyendo cuando se va detectando la intencionalidad de la palabra", agrega el profesional.

Realmente se utilizan dos modelos de procesamiento de lenguaje natural: uno es el de similitud, que busca agrupar, de todo el universo de tuits capturados, aquellos que tienen más similitudes entre ellos.

El segundo modelo se aplica a esos grupos. "Es el de sentimientos y es para detectar el nivel de agresividad en cada grupo. En los grupos de tuits más agresivos que detectamos procuramos identificar a sus generadores y replicadores. Esa es la combinación ganadora", afirma Díaz López.

Aunque pueda parecer extraño **que la ciencia de datos analice los sentimientos humanos, el profesor asegura que es posible,** pues los conjuntos de datos utilizados son clasificados por humanos.

"Ese grupo de datos nos sirve para entrenar un modelo matemático, de forma que cuando le pasemos un tuit determinado él lo codifica y lo clasifica como positivo o negativo. Cuanto más grande sea el conjunto de datos, más preciso puede ser el modelo porque aprenderá más. Eso fue lo que hicimos para el segundo modelo", asegura.

Hacer este proceso manualmente sería muy dispendioso; al disminuir el tiempo de análisis, la respuesta de un agente del Estado para detectar dónde está el grupo agresivo o el foco de un posible ciberdelito puede ser más veloz. En suma, el modelo identifica aquellos nodos que pueden influir la ejecución de posibles hechos punibles y logra una acción inmediata.

Si se llega a demostrar que en las redes se orquestó un plan para cometer un delito, quienes lo hicieron pueden ser acusados de ciberdelito, pues usaron el ciberespacio para promover un delito en el espacio físico.

Una estrategia de ciberdefensa

Esa oportunidad de pronta identificación de posibles ciberdelitos que da el modelo que están desarrollando los investigadores del programa de Matemáticas Aplicadas y Ciencias de la Computación, llevó a plantear que puede ser aplicado por las Fuerzas Militares.

"Claramente se requiere que las fuerzas de seguridad del Estado monitoreen ese tipo de situaciones peligrosas para prevenir delitos en el marco de estrategia de ciberdefensa nacional", explica el profesor.

No obstante, como existe el riesgo de que estos mensajes en redes sociales se interpreten de forma equivocada y en lugar de ser focos de ciberdelitos sean simplemente manifestaciones espontáneas del derecho a la protesta social, se necesitan profesionales capacitados en el modelo.

"Es muy importante que los analistas de los datos puedan validarlos con un sentido crítico y con objetividad. Ningún modelo de este tipo funciona de manera autónoma; siempre debe existir un humano que hace la validación de lo que dice el modelo", anota Díaz López.

Otro riesgo que se podría presentar es el relacionado con traspasar la línea de privacidad y de autonomía de cada persona. Sin embargo, los investigadores aseguran que esa línea está establecida, puesto que existe una ley de inteligencia y contrainteligencia que establece los límites de las entidades del Estado para adelantar ese tipo de actividades.

NLP descubre la esencia de las palabras

El Procesamiento del Lenguaje Natural (NLP) es el área de inteligencia artificial que aborda la comunicación humana a través de modelos de aprendizaje automático computacional. En síntesis, les da a las palabras una representación matemática, con lo cual un modelo de NLP podría analizar la expresividad de una frase, interpretar el deseo de una persona a partir del uso de ciertas palabras o, incluso, establecer similitudes de intención entre oraciones. Por lo tanto, NLP ofrece un futuro prometedor para la comprensión del lenguaje humano, que puede ser útil en diferentes campos como servicio al cliente, publicidad, traducción de voz y elaboración de perfiles de sospechosos, entre otros.
En el contexto de la seguridad nacional puede ser útil para detectar campañas provenientes de Estados hostiles y organizaciones de ciberdelincuencia. Además, podría facilitar la resolución de casos relacionados con estrategias de desinformación contra personas u organizaciones privadas.
Como 'Manipulación social hostil' se conoce a la generación de violencia e inestabilidad a través de las redes sociales. La gran cantidad de información difundida de esa manera hace que sea difícil monitorear e identificar su origen. Por esta razón, las autoridades están viendo en la ciencia de datos un recurso ideal para recopilar, procesar y analizar datos que conduzcan a la identificación oportuna de ese tipo de amenazas.

Además, el proyecto incluyó solo información de fuentes abiertas, es decir, datos públicos difundidos en redes sociales y no información privada. Es decir, el mismo ejercicio que vienen haciendo de tiempo atrás las empresas de *marketing* para conocer el impacto de una nueva marca.

Hacia el futuro, aseguran Campo Archbold y Ramírez, el proyecto de investigación planea aumentar las características consideradas en el análisis de tuits. Esto permitirá hacer una evaluación más profunda de la información obtenida y detectar patrones avanzados de amenazas especializadas.

Así mismo, aplicarán el modelo a temas específicos como las fotomultas, la generación de pánico en las finanzas o cualquier tema que genere susceptibilidades y se exprese con determinadas emociones en las redes sociales, particularmente en Twitter donde la opinión de las personas queda al descubierto.