Un código para la identidad

11 septiembre, 2023

Un estudiante de la UNR participó en el desarrollo un algoritmo que permite digitalizar el archivo periodístico de Abuelas de Plaza de Mayo.

El estudiante de la Universidad Nacional de Rosario, Matías Naranjo Herper, obtuvo junto a dos compañeros el tercer puesto en el desafío “Inteligencia Artificial (IA) por la Identidad”, concurso que tuvo el objetivo de digitalizar y transcribir el archivo periodístico de Abuelas de Plaza de Mayo.

Esta convocatoria fue organizada por el Ministerio de Ciencia, Tecnología e Innovación de la Nación, junto con la Fundación Sadosky, y congregó a estudiantes de Computación, Ciencias de Datos y carreras afines de todo el país. “Supuestamente tenían digitalizada su biblioteca con recortes de diarios pero en realidad poseían imágenes con noticias escaneadas, lo que imposibilitaba el acceso total y sistematizado a ellas. La utilización de un OCR (Reconocimiento óptico de caracteres) no era una posibilidad debido a la complejidad y calidad de las imágenes.

Un OCR permite pasar imágenes a texto sin tener que copiarlo manualmente. El problema de esta herramienta es que al utilizar este proceso se generaban textos incomprensibles, mezclados con noticias que

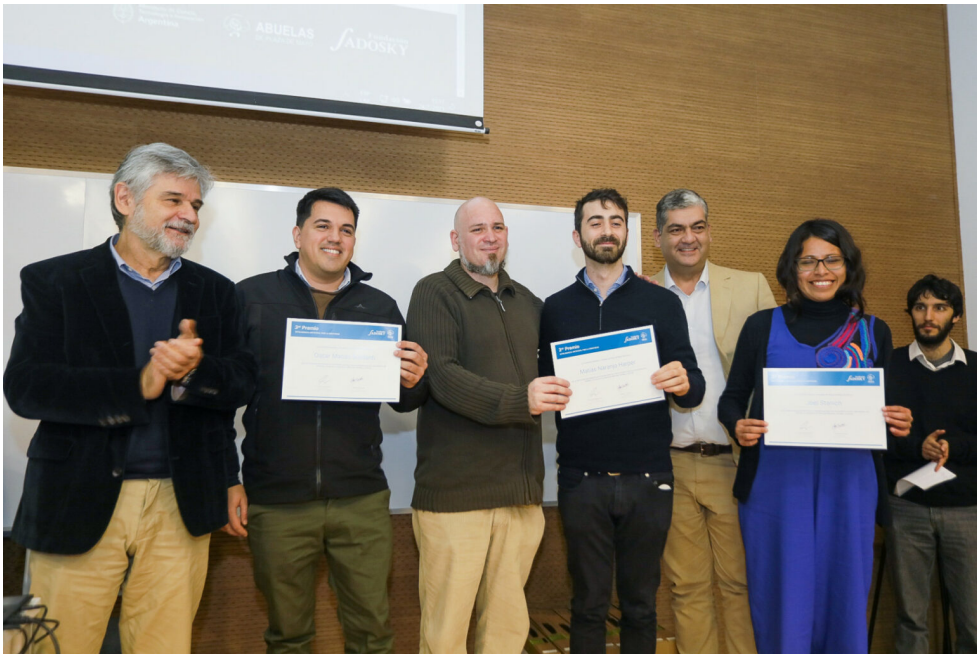
no tenían relación con la principal y con las publicidades propias que estaban incluídas en los recortes de diario. “Se planteó el desafío de poder encontrar una solución más efectiva. Dentro del desafío existieron algunos espacios cooperativos, se anotaron cerca de 200 personas, lo que permitió generar una pequeña base de datos de noticias transcriptas a mano que servía para chequear que los algoritmos que desarrollamos estén funcionando correctamente”.

A lo largo de 45 años de trabajo, las Abuelas de Plaza de Mayo produjeron una gran cantidad de documentación que registra las acciones que llevaron adelante en su búsqueda por restituir la identidad de las niñas y niños desaparecidos durante la última dictadura militar. Hoy en día, cuentan con un inmenso archivo de recortes periodísticos que preserva documentación de un gran valor histórico para la institución, sus familias y la sociedad. Además, constituye una fuente para la formación, la investigación y promueve el ejercicio de los derechos humanos. Las noticias recopiladas en tantos años de trabajo se alojaban en cerca de 30 gigabytes de imágenes, lo que presentaba un gran desafío para todos los equipos que se anotaron en esta convocatoria.

Naranjo Harper integró un equipo con otros dos jóvenes, Matias Bonfanti de la Provincia del Chaco y Joel Stanich de Córdoba. “Nuestra propuesta de solución fue hacer dos modelos de aprendizaje supervisado de reconocimiento de imagen. Trabajamos todo pensando en la noticia como una imagen. El primero de ellos es un modelo de segmentación: que permite reconocer cuantas noticias y publicidades hay en cada imagen. Las noticias son separadas y las publicidades descartadas. El modelo fué entrenado con mil imágenes etiquetadas a mano”, explicó y agregó: “El otro modelo es de clasificación: el mismo se encarga de clasificar las partes de las noticias (Titulo, cuerpo, epígrafes, bajada, copete, etc) y generar recuadros (bounding box) que pueden ser procesados individualmente. Este segundo modelo fue entrenado con 900 imágenes etiquetadas a mano”.

Esto permitió que se pueda utilizar un OCR en cada recuadro y ser transcrito automáticamente junto a su etiqueta. “Así pudimos hacer un archivo de texto donde se guardaba la información de cada noticia de manera ordenada,básicamente con una etiqueta de “título” teníamos el

texto que poseía el Título. Esta es una propuesta de solución porque la realidad es que hay un montón de falencias dentro de las propias noticias: había algunas que estaban cortadas al medio, otras manchadas, escaneos movidos, noticias que ni siquiera eran reconocibles para el ojo humano".



En el medio Naranjo Harper, estudiante de la UNR que obtuvo el tercer puesto.

Matías destacó que este desafío fué de gran importancia para poder generar en un futuro buscadores que permitan a cualquier persona investigar en este gigante archivo. "Por ejemplo, puedo buscar "Estela de Carlotto" y que me aparezcan todas las noticias donde figura el nombre. Ahí al hacer clic en la que nos interesa se puede acceder al texto completo y, además, a la imagen de la noticia para poder hacer un complemento".

Por este desarrollo, el equipo obtuvo el tercer premio. La entrega de reconocimientos fue realizada en la Facultad de Ciencias Exactas de la Universidad de Buenos Aires, con presencia del Ministro de Ciencia, Tecnología e Innovación, Daniel Filmus y la Vicepresidenta de Abuelas de Plaza de Mayo, Buscarita Roa. Durante el encuentro hubo mensajes grabados de la Presidenta de Abuelas de Plaza de Mayo, Estela de Carlotto, y de Taty Almeida, Madre de Plaza de Mayo Línea Fundadora. "Fue una experiencia muy interesante, recorrimos el Museo de la Memoria, conocimos a nietos recuperados, y pudimos dialogar con personas de distintos lugares del país".

El estudiante de la UNR resaltó que se enfocaron en poder llevar lo que

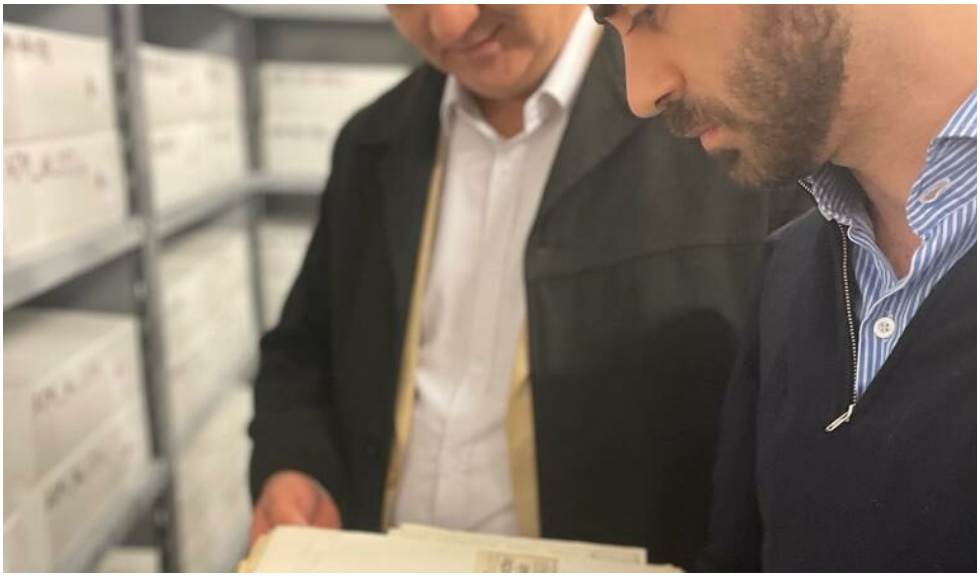
construyeron a las palabras más simples posibles porque muchas veces existe la idea de que “la programación es abstracta e inentendible y no es así”. A su vez añadió que “hay personas que hacen excelentes preguntas y eso es más importante que la capacidad de saber todo sobre código”, por lo que la experiencia de poder intercambiar palabras con personas de otros ámbitos, y sobre todo con Abuelas de Plaza de Mayo, conformaron un aprendizaje invaluable que hace que el trabajo realizado cobre un sentido aún mayor.

Un espacio de trabajo colaborativo

Los tres integrantes del equipo trabajaron durante todo el proceso a distancia, y fué recién en la entrega de premios que se vieron por primera vez de manera presencial. “Habíamos estudiado programación juntos y quedó muy buena relación. Hablamos mucho por videollamada y WhatsApp, éramos realmente un grupo de amigos. Sin embargo, nunca nos habíamos visto presencialmente”.

En este sentido, se optó por trabajar con un servicio de Google Cloud Platform, generando una computadora en la nube. “Todos nos conectamos a ese servidor por medio de un IP, y trabajamos ahí. Los tres llevamos el proceso de manera muy responsable: programamos en el mismo código, etiquetamos imágenes a mano, y teníamos reuniones en las que planificamos los pasos a seguir”.

El arduo proceso de trabajo llevó un mes y medio en donde los tres tuvieron que combinar estas tareas con el resto de responsabilidades laborales y sociales que ya contaban en su cotidianidad. “La mayor cantidad de trabajo se concentró en etiquetar las imagenes y en darle el broche final al modelo. Era la primera vez que nos encontramos frente a un desafío tan importante.”… “Trabajamos con muestras representativas del total de datos, lo que nos permitía que lo que probamos con cinco o diez imágenes y que nuestra computadora virtual permitía procesar en un tiempo razonable, pudiera luego trasladarse al total de imágenes con una computadora más poderosa (o mayor cantidad de tiempo)”.



Las noticias recopiladas se alojaban en cerca de 30 gigabytes de imágenes, lo que presentaba un gran desafío.

Un recorrido académico con muchas aristas

Matías Naranjo Harper es oriundo de Pergamino pero ya hace muchos años que vive en Rosario. Comenzó estudiando Ingeniería Mecánica en la Facultad de Ciencias Exactas, Ingeniería y Agrimensura de la UNR, carrera en la que estuvo casi tres años. En paralelo, en su último año comenzó a estudiar la Licenciatura en Tecnologías Aplicadas al Arte Sonoro, en la Escuela de Música de la Universidad.

“Cuando terminé el ciclo básico de ingeniería noté que mis intereses eran otros, por lo que decidí cambiarme definitivamente. Estoy a seis materias de recibirme de la Licenciatura en Tecnologías Aplicadas al Arte Sonoro, y desde hace casi dos años, me empecé a interiorizar en lo que es la Ciencia de Datos, que es a lo que me dedico en la actualidad”, reflexionó.

El objeto de esta carrera gira en torno a formar profesionales especializados en las diversas áreas en las que tiene incumbencia, desde el audio profesional, con la sonorización de espectáculos y medios de comunicación masiva, registro fonográfico, diseño de sonido para artes escénicas (audiovisuales, multimediales, etc.), hasta la asistencia de toda manifestación artística que demande un enfoque creativo de la manipulación del material sonoro, con un sustento estético firme que le permita al profesional ser un interlocutor válido entre el artista y la realización de la obra. “Tiene mucho de música pero también hay una parte más técnica porque cuenta con materias como acústica, matemática, física,