Vol. 1
Núm. 1
2024

REVISTA Datos, Políticas e Innovación Pública

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

REVISTA
Datos, Políticas
e Innovación Pública

Vera, Fabián; Briseño, Ramón A. ; Alonso Contreras, Guadalupe
Aplicación Web para identificar personas a partir de la similitud semántica
en descripciones físicas textuales utilizando técnicas de Procesamiento del
Lenguaje Natural
Datos, Políticas e Innovación Pública, vol. 1, núm. 1, marzo; 2024, pp. 53 - 60.
Instituto de Información Estadística y Geográfica

Aplicación Web para identificar personas a partir de la similitud semántica en descripciones físicas textuales utilizando técnicas de Procesamiento del Lenguaje Natural

Fabián Vera, Ramon A. Briseño , Guadalupe Alonso-Contreras

Universidad de Pamplona, Pamplona, Colombia, Universidad de Guadalajara, Centro Universitario de Ciencias Económico Administrativas, Doctorado en Tecnologías de Información, México, Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas (UPIITA) Instituto Politécnico Nacional CDMX, México\

fabian.vera@unipamplona.edu.com, alejandro.bmartinez@alumnos.udg.mx, galonsoc1700@alumno.ipn.mx

Resumen. En Jalisco, uno de los estados con mayor número de desapariciones en el país (Ibarra, 2022), persiste el desafío de identificar a Personas Fallecidas Sin Identificar (PFSI). Por lo que, el proyecto RENIA (Red neuronal de investigación y análisis para la búsqueda e identificación de personas) busca la iniciativa de abordar este problema por medio de Inteligencia Artificial. El problema al que nos enfrentamos es una tarea compleja debido al alto número de desapariciones en el estado, por lo tanto, el trabajo propuesto muestra una solución basada en la implementación de una aplicación web con uso de técnicas de Procesamiento de Lenguaje Natural (PLN) para la comparación de descripciones textuales de datos de personas desaparecidas, con el fin de colaborar con las autoridades gubernamentales y facilitar el proceso de búsqueda e identificación de personas en el estado de Jalisco. El modelo Universal Sentence Encoder implementado en el aplicativo web, como parte del proyecto RENIA, alcanzó una efectividad del 90% al encontrar registros buscados dentro de las 100 mejores similitudes de cada búsqueda en la base de datos de PFSI (IJCF-PFSI, s.f.), que cuenta con 4,278 registros. La capacidad que presenta para localizar registros específicos dentro de la extensa base de datos de PFSI aporta una eficacia considerable, agilizando el proceso de identificación y brindando así una herramienta eficiente para aumentar las posibilidades de encontrar a personas reportadas como desaparecidas.

Palabras clave. *Procesamiento de Lenguaje Natural (PLN), Identificación de personas, Inteligencia Artificial (IA), Análisis Semántico, Aprendizaje automático, Embeddings.*

1. Introducción

A raíz de diversas situaciones como conflictos armados, desastres naturales, migraciones, entre otros, se presentan cientos de miles de personas desaparecidas a nivel mundial. Según la doctora en antropología Isabel Beltrán Gil (adondevanlosdesaparecidos, 2022), las desapariciones en México están relacionadas con la impunidad y el incumplimiento de políticas públicas para abordar el problema. Esto se debe a la falta de capacidad y recursos de las instituciones gubernamentales en los procesos de búsqueda e identificación, lo que agrava la crisis forense en el país. La falta de registros precisos y una base de datos actualizada por parte del gobierno mexicano, como resalta Consuelo Morales Elizondo (Morales, 2015), activista y fundadora de Ciudadanos en Apoyo a los Derechos Humanos, dificulta tener información estadística precisa sobre el problema.

Por otro lado, según la investigación de Josué Ibarra, estudiante de Estudios Políticos y Gobierno en la UDG (Ibarra, 2022), entre 2000 y 2022, se han reportado alrededor de 12,949 desapariciones en Jalisco, siendo los últimos años los más críticos con 7,912 casos. Por lo cual, la zona metropolitana de Guadalajara registra un 17% de todas las desapariciones en México.
Además, se destaca el conflicto entre las desapariciones forzadas y el incidente de los tráileres de la muerte en Jalisco, ya que, durante el gobierno de Aristóteles Sandoval, se utilizaron dos tráileres para transportar los 322 cadáveres que superaban la capacidad de las morgues (Vargas, 2020).

A través del proyecto RENIA, se propone abordar el problema de identificación de personas en condición de desaparición en Jalisco mediante la Inteligencia Artificial y el Procesamiento de Lenguaje Natural, con el objetivo de realizar búsquedas e identificar la similitud semántica entre descripciones textuales de las personas desaparecidas mediante un aplicativo web. El Procesamiento de Lenguaje Natural se encargará de la capacidad de las computadoras para manejar e interpretar el lenguaje humano y realizar diversas tareas como la traducción o el análisis de textos (Berryhill, 2019), como la similitud semántica, definida como una medida del grado en el que dos fragmentos de texto tienen el mismo significado.

Este trabajo toma como base de datos las fichas de personas fallecidas sin identificar (PFSI) del SEMEFO Jalisco (IJCF-PFSI, s.f.), las cuales son contrastadas con el significado semántico de descripciones de características físicas introducidas por el usuario en la aplicación web propuesta. Se implementa el modelo pre entrenado de PLN Universal Sentence Encoder (USE) para calcular la similitud entre el texto ingresado por el usuario y los registros de personas desaparecidas en la base de datos recolectada. Asimismo, la aplicación web demostró que esta búsqueda por medio

PLN puede ser una alternativa eficiente para usuarios sin conocimientos técnicos de IA, ya que ofrecerá una interfaz clara y sin filtros de búsqueda.

2. Planteamiento del problema

De acuerdo con Prieto Sferrazza Taibi (Sferrazza, 2021), la desaparición forzada de personas es considerada una de las más atroces violaciones a los derechos humanos, generando consecuencias perjudiciales tanto para la persona desaparecida, sus familiares y la sociedad en su totalidad. Por lo que, puede conceptualizarse como el arresto, detención, secuestro o cualquier forma de privación de la libertad por parte de agentes del Estado. Esta acción se acompaña de la negativa a reconocer la privación de libertad, ocultamiento de la suerte o paradero de la persona desaparecida, sustrayéndola a la protección de la ley.

Las víctimas de desapariciones son conscientes de que sus familias desconocen su paradero, reduciendo las posibilidades de recibir ayuda. Al quedar separados de la sociedad, se encuentran privadas de todos sus derechos y a merced de sus aprehensores, siendo el caso. Los familiares y amigos de estas personas pasan por una angustia similar, enfrentando la incertidumbre acerca del destino de la víctima. Vivir en constante preocupación por la causa de la desaparición es una angustia mental, ignorando si la víctima vive aún y de ser así, dónde se encuentra recluida, en qué condiciones y cuál es su estado de salud (“La desaparición forzada en méxico: una mirada desde los organismos del sistema de naciones unidas”, 2019).

Cepeda y Leetoy en el artículo ‘De víctimas a expertas: estrategias de agencia cívica para la identificación de desaparecidos en México’ (Cepeda, 2021), destacan la crisis de violencia sin precedentes en México. Jalisco es uno de los más afectados, con registros que indican 7,045 personas no localizadas y 2,238 desaparecidas desde 1995 hasta abril de 2020.

El colectivo de mujeres Por Amor a Ellxs (Facebook, s.f.), en la Zona Metropolitana de Guadalajara, ha implementado estrategias en la solución de la crisis de desapariciones. Utilizan redes sociales, especialmente Facebook, para amplificar la información que recolectan, extendiendo su alcance a otras familias en la misma situación. En su página comparten información sobre qué hacer en caso de la desaparición de un ser querido, resaltando la difusión de nombres, descripciones físicas y señas particulares de personas no reclamadas en las morgues. Desde su inicio en 2016, su labor ha contribuido a la identificación de más de 100 personas. La página se enriqueció con la participación de profesionales que asistieron a las instalaciones del SEMEFO para examinar fotografías de cuerpos no reclamados, con el objetivo de proporcionarles identidad y buscar a sus desaparecidos.

Además, el periódico El País en un video documental del año 2018, “Las familias de Jalisco en busca de sus desaparecidos” (Ibarra, 2022), presenta testimonios de diferentes personas buscando a sus desaparecidos. Consuelo Velázquez, entre ellos, menciona cómo le negaron tomar fotografías en el SEMEFO para difundirlas entre colectivos de búsqueda e identificación en Jalisco. Destaca la ausencia de difusión adecuada, lo que llevó a la necesidad de recurrir a descripciones manuscritas para compartir información entre colectivos y permitir que las personas sin identificar fueran reconocidas por sus familiares. Esta falta de herramientas tecnológicas de uso masivo para la difusión de información sensible dificulta aún más este proceso.

Para la propuesta de aplicativo web de este trabajo se toma como base de datos el Registro PFSI de SEMEFO Jalisco (IJCF-PFSI, s.f.), ya que posee un campo de búsqueda en su sitio web. Sin embargo, dentro de su buscador, se presenta una limitación ya que los usuarios que buscan personas desaparecidas desconocen las palabras exactas utilizadas por las autoridades para describir los cuerpos de personas fallecidas sin identificar.

3. Trabajos relacionados

Se ha revisado la literatura con el objetivo de identificar trabajos donde se produzcan aplicaciones de PLN enfocadas en la clasificación de textos y similitud semántica. Por ejemplo, se identifica que en (Merayo, 2019) se aplica PLN para identificar contenido inapropiado en texto, utilizando diversas técnicas como machine learning y deep learning, destacando máquinas de soporte vectorial. Esto demuestra la aplicabilidad de la técnica a redes sociales y plataformas como YouTube para identificar textos no regulados. Por otro lado, en (Hu, 2021) se aborda la clasificación de género de los nombres con modelos de aprendizaje automático basados en caracteres. Se utilizan algoritmos como Class Scaled Logistic Regression, Deep Neural Network, Long Short-Term Memory, Char-BERT y Name embedding, mostrando un enfoque interesante en la identificación.

En (Li, 2023) se emplean redes de gráficas convolucionales para hacer una clasificación del texto en una fuente heterogénea. El uso de grafos representa las conexiones que hay entre los pesos en la clasificación de los textos.
Hemos encontrado que el esquema de encriptación de búsqueda semántica difusa (FSSE), descrito en (Liu, 2020), admite la búsqueda de múltiples palabras clave en datos encriptados en la nube. Permitiendo la búsqueda difusa y expansiones semánticas, utilizando huellas dactilares de palabras clave y distancia de Hamming para mejorar la precisión. De manera similar, en (Nemeshaev, 2021) se discute la posibilidad de utilizar la búsqueda semántica para

nombrar a un experto en sistemas de automatización, describiendo características de la búsqueda por etiquetas, la ontología del dominio y el algoritmo de búsqueda semántica.

En (Jeong, 2022) se observa el enfoque hacia el análisis de conjunto de datos y el Procesamiento de Lenguaje Natural para la tarea de pregunta-respuesta basada en diálogos. Aborda las limitaciones de los modelos de lenguaje pre-entrenados que no suelen considerar el razonamiento de sentido común, por lo que se presenta el modelo Diálogo-QA con Razonamiento de Sentido Común (DQACR), el cual aprovecha la búsqueda semántica y el aprendizaje continuo para mejorar el razonamiento de sentido común y pérdida de información.

En cuanto al problema de Desambiguación del Sentido de Palabras (WSD en sus siglas en inglés) se aborda en (Nodehi, 2022) mediante técnicas de PLN y algoritmos metaheurísticos. La propuesta implica el uso de una red neuronal basada en grafos de WordNet para generar incrustaciones de palabras y sentidos simultáneos, mejorando la precisión mediante un método para agrupar y mapear sentidos en el grafo.

Por otro lado, considerando trabajos relacionados directamente con encontrar la similitud semántica de dos textos, se encontró que en (Balaha, 2021) se lleva a cabo la calificación automática de exámenes, comparando la similitud entre las respuestas de los estudiantes y las respuestas de referencia. Este trabajo comparó el rendimiento de varios modelos como BERT, Glove, FastText, Word2Vec y USE, donde éste último mostró mejor exactitud. En (Mahajan, 2020) se propone un ejercicio de identificación de grados de similitud entre registros de visitas periódicas de chequeos de salud a un hospital por parte de una persona. El trabajo destaca que las fichas de visitas suelen ser muy similares, pero es importante resaltar cuáles de ellas carecen de similitud en las descripciones textuales para identificar cambios en la salud del paciente. El experimento se llevó a cabo con el modelo pre entrenado de PLN BERT.

Adicionalmente, en (Sheth, 2021) se utiliza una aplicación con Universal Sentence Encoder para encontrar similitud en las descripciones de las características del recurso humano de una organización con las descripciones de capacitaciones. De esta forma, se busca identificar a las personas adecuadas para recibir entrenamiento en áreas específicas. En (Vowinckel, 2023), el contexto de las solicitudes de patentes, es fundamental identificar el estado de la técnica relevante y emplear estos documentos para evaluar la novedad e inventiva de la invención reivindicada. En este propósito, se utiliza el modelo BERT, que permite generar representaciones vectoriales de texto, facilitando el uso de similitud vectorial como indicador de la similitud semántica entre textos.

Finalmente, con el objetivo de conocer más sobre trabajos relacionados con identificar personas a partir de similitudes semánticas, se tiene que en (Zhou, 2023) se presenta una propuesta para encontrar imágenes de peatones utilizando PLN, buscando integrar características textuales de imagen y abordar la falta de información semántica en las características de las personas y analizando sus propiedades específicas en las imágenes y completándose con detalles semánticos. Se presenta una red de complemento de información que en (Frikha, 2021) se expone como un método de búsqueda de personas desaparecidas en áreas públicas, basado en atributos semánticos en sistemas de vigilancia. Utiliza clasificadores de atributos semánticos profundos basados en redes neuronales convolucionales para aprender y reconocer características en condiciones no controladas. En (Martin-Rodilla, 2019) se menciona la aplicación de técnicas de PLN en el análisis de informes forenses en el contexto de desapariciones forzadas durante la dictadura militar brasileña. Propone un sistema de extracción de información que identifica entidades nombradas y terminología clasificatoria e indexadora, asistiendo a los investigadores en la búsqueda de patrones en los informes de autopsia.

4. Propuesta

La propuesta se basa en una aplicación web que utiliza el modelo USE de procesamiento de lenguaje natural, el cual codifica textos en vectores de alta dimensión que se pueden utilizar para calcular su similitud semántica. Con dicho modelo, un usuario internauta puede realizar una búsqueda en la aplicación web de una persona desaparecida escribiendo una descripción textual de características físicas, a lo que la aplicación responderá con las fichas de personas desaparecidas (provenientes del formulario PFSI para esta aplicación) ordena de mayor a menor coincidencia semántica.

Para la implementación de la aplicación web se llevaron a cabo 3 pasos fundamentales: primero, la recolección de información y generación de la base de datos; segundo, pruebas de desempeño del modelo USE en similitud semántica de textos con la base de datos recopilada; tercero, el desarrollo de la aplicación web para búsquedas textuales con similitudes semánticas en la base de datos recolectada. Mismos pasos que se describen a continuación.

4.1 Base de datos

Como ya se mencionó anteriormente, la base de datos del aplicativo web se tomó del formulario PFSI (IJCF-PFSI, s.f.) que dispone SEMEFO Jalisco para la identificación de personas fallecidas sin identificar. El formulario PFSI descrito, cuanta con 9 campos: Id, Fecha ingreso, Sexo, Probable nombre, Edad, Tatuajes, Indumentarias, Señas particulares

y Delegación IJCF, los cuales pueden contener, o no, información de los cuerpos. Tatuajes, Indumentarias y Señas particulares son campos que tienen descripciones textuales. Con la finalidad de utilizar la similitud semántica y no una búsqueda por medio de filtros, por cada registro todos los campos, excepto el campo Id, se concatenaron en uno solo llamado descripción y se guardaron en una base de datos MySQL. Por consiguiente, la base de datos del aplicativo cuenta con 4,278 registros; cada registro sólo tiene dos campos: el Id del registro y el campo Descripción, el cual será utilizado para analizar las similitudes semánticas con las búsquedas ingresadas por el usuario en el aplicativo web.

4.2 Arquitectura del modelo USE

Antes de explicar la arquitectura del modelo USE, es importante definir primero algunos conceptos clave como:

Tokenización: En procesamiento del lenguaje natural, se basa en la idea de convertir una secuencia de texto en partes más pequeñas (“What is Tokenization? Types, Use Cases, Implementation”, s.f.).

Incrustación: se refiere a una representación vectorial de un objeto, como una palabra, un documento o una imagen. Esta representación captura características y relaciones importantes del objeto en un espacio de menor dimensión, permitiendo que las máquinas los manipulen y analicen de manera más eficiente.

La arquitectura del codificador universal de frases (USE) incluye un modelo de codificación formulado como una red de promediación profunda (DAN). A continuación, se explica paso a paso la arquitectura del modelo utilizado para el desarrollo del prototipo:

4.2.1 Codificador de red de promediado profundo (DAN):

El codificador de red de promediado profundo (DAN), un componente del codificador Universal Sentence Encoder (USE) (Cer, 2018), está diseñado para generar incrustaciones de frases de forma eficiente. La arquitectura se explica paso a paso a continuación:

1. Incrustaciones de entrada: El codificador DAN toma las incrustaciones de entrada para palabras y bi-gramas y las promedia juntas (Cer, 2018).
2. Red neuronal: Las incrustaciones promediadas se introducen en una red neuronal profunda (DNN).
3. Incrustación de frases: La DNN procesa las incrustaciones promediadas para producir incrustaciones de frases de 512 dimensiones.
4. Entrenamiento e implementación: El codificador DAN se entrena de forma similar al codificador basado en transformadores y se implementa en TensorFlow (Cer, 2018).
5. Eficiencia y compensaciones: El codificador DAN está optimizado para una inferencia eficiente, ofreciendo una precisión ligeramente reducida en comparación con el codificador basado en transformador (Cer, 2018). Este compromiso entre precisión y recursos computacionales lo hace adecuado para diversas tareas de procesamiento del lenguaje natural y aplicaciones de aprendizaje por transferencia.

Otro de los puntos fundamentales de la arquitectura son las tareas de transferencia en el contexto del codificador universal de frases (USE), ya que implican la aplicación de incrustaciones de frases pre-entrenadas a diversas tareas de procesamiento del lenguaje natural, ofreciendo una visión del impacto de la selección e integración de modelos en el rendimiento de las tareas.

Esta arquitectura proporciona un enfoque flexible y eficiente para generar incrustaciones de frases, atendiendo a las diversas necesidades de las tareas de procesamiento del lenguaje natural.

4.3 Pruebas de desempeño del modelo USE

El modelo USE está entrenado y optimizado para procesar texto que consiste en más de una palabra, como oraciones, frases o párrafos cortos. Su entrenamiento abarca diversas fuentes de datos y tareas, con el objetivo de adaptarse dinámicamente a la comprensión del lenguaje natural. La entrada al modelo es un texto de longitud variable, y la salida es un vector de 512 dimensiones. Se aplica el modelo de referencia STS (Semantic Textual Similarity) para medir la similitud semántica, evaluado mediante el benchmark STS. Este benchmark proporciona una medida de cómo las puntuaciones de similitud, calculadas mediante incrustaciones de oraciones, se relacionan con los juicios humanos.

Por otro lado, se emplea la correlación de Pearson (una prueba que mide la relación estadística entre dos variables continuas) para evaluar la calidad de las puntuaciones de similitud generadas por la máquina en comparación con las evaluaciones humanas (adondevanlosdesaparecidos, 2022). Este codificador se distingue de otros modelos de incrustación de nivel de palabra, porque se entrenó en una serie de tareas de predicción de lenguaje natural que requieren modelar el significado de secuencias de palabras individuales (adondevanlosdesaparecidos, 2022).

Las pruebas de rendimiento del modelo USE implican analizar qué tan bien el modelo relaciona una descripción textual

con su registro similar escrito con palabras diferentes. Para llevar a cabo estas pruebas, se seleccionaron 40 registros de manera aleatoria, con la condición de que tuvieran más de 30 palabras (garantizando así que tuvieran una descripción textual). Después, esos 40 registros se parafrasearon tres veces con palabras distintas y con diferentes cantidades de palabras, utilizando la plataforma de IA Chat GPT (ChatGPT, s.f.). Se solicitó el parafraseo con una cantidad similar de palabras, el parafraseo con el 30% menos de palabras y el parafraseo con el 30% más de palabras. Posteriormente, se compararon las descripciones parafraseadas una a una, evaluando la similitud semántica con el modelo USE y los 4,278 registros de la base de datos. Al identificar cada registro parafraseado con su ID, se pudo determinar si el modelo lo clasificó entre los 100 IDs de registros con mayor similitud mediante el coeficiente de correlación de Pearson. El hecho de que el modelo detecte similitud de un registro parafraseado dentro de los 100 IDs de registros con mejor similitud se consideró como un acierto del modelo. De las 40 pruebas por cada uno de los tres tipos de parafraseo, se obtuvo el porcentaje de aciertos del modelo. El porcentaje de aciertos del modelo ayudó a estimar cuánto más fácil resulta realizar una búsqueda con la aplicación web en comparación con hacerlo manualmente en la base de datos cuando no se conocen las palabras exactas con las que se registró el cuerpo de una persona fallecida sin identificar en SEMEFO.

4.4 Desarrollo de la aplicación web

Diagrama de casos de uso

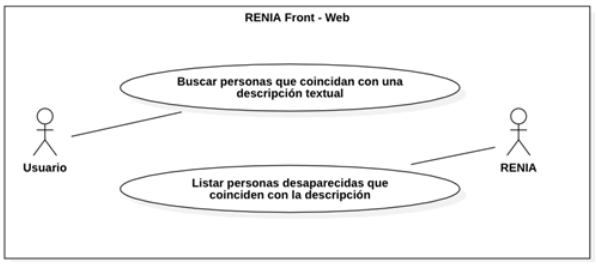

Figura 1. Diagrama de casos de uso del aplicativo web

El desarrollo del aplicativo se dividió en Front-End y Back-End, los cuales conforman módulos de una arquitectura orientada a servicios (siendo el cliente el Front-end y el servidor el Back-End) como se muestra en la Figura 1 y Figura 2, respectivamente. El Back-End implementa una API escrita en lenguaje Python, que proporciona un end-point donde se calcula la similitud semántica de la búsqueda del usuario con los registros de la base de datos a través del modelo USE, devolviendo los registros ordenados de mayor a menor similitud semántica. Por otro lado, el Front-End está diseñado con el framework React para crear interfaces de usuario. La Figura 3 presenta la interfaz gráfica de búsqueda y obtención de resultados del aplicativo web. En el repositorio https://github.com/alexAgnus/RENIAtexto.git (Briseño, s.f.) se puede descargar el código de la implementación para pruebas y mejoras del aplicativo.

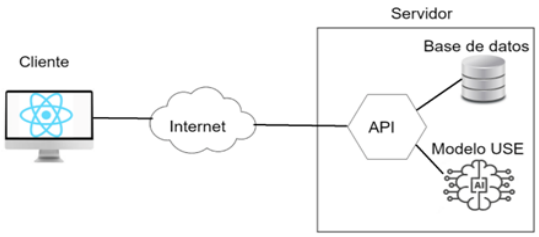

Figura 2. Arquitectura cliente-servidor de la aplicación web del proyecto RENIA.

Figura 3. Interfaz gráfica de la aplicación web en el apartado de búsqueda de personas desaparecidas a través de similitud semántica de textos.

5. Resultados

Con el objetivo de parafrasear los 40 registros seleccionados de manera aleatoria y asegurando que tuvieran más de 30 palabras, se proporcionaron las siguientes instrucciones a Chat GPT: *reescribe las siguientes oraciones con diferentes palabras, manteniendo el significado semántico; reescribe las siguientes oraciones con diferentes palabras manteniendo el significado semántico y disminuyendo el número de palabras en un 30%; y reescribe las siguientes oraciones con diferentes palabras manteniendo el significado semántico y aumentando el número de palabras en un 30%. Dichas instrucciones generaron tres bases de datos de registros parafraseados: una con la misma cantidad de palabras, otra con menos palabras y una tercera con más palabras, respectivamente.*

De las 120 pruebas totales realizadas, el modelo acertó en 95 ocasiones, con un 79% de efectividad al encontrar el registro buscado de manera parafraseada en los primeros 100 registros con mayor similitud. En las 40 pruebas de búsqueda con parafraseo y una cantidad similar de palabras distintas, el modelo acertó en 32 ocasiones con un 80% de efectividad. En las búsquedas con parafraseo disminuyendo la cantidad de palabras, el modelo acertó en 36 ocasiones (90% de efectividad). Finalmente, en las búsquedas con parafraseo aumentando el número de palabras, el modelo acertó en 27 ocasiones (67.5% de efectividad) como se muestra en la Tabla 1.

Tipo de búsqueda	Aciertos de 40 búsquedas	Porcentaje de efectividad
Parafraseo con la misma cantidad de palabras distintas	32	80%
Parafraseo con el 30% menos de palabras	36	90%
Parafraseo con el 30% más de palabras	27	67.5%

Tabla 1. Resultados de las pruebas de desempeño del modelo USE.

6. Discusión y conclusiones

Las pruebas de PLN utilizando similitud semántica de textos realizados en este trabajo, utilizando el modelo USE y la base de datos de personas desaparecidas sin identificar (PFSI) de SEMEFO Jalisco, demostraron ser una herramienta potente en la identificación de personas a través de descripciones textuales de rasgos físicos. También se observó que al reescribir un registro con palabras distintas y reducir en un 30% la cantidad de palabras para su posterior búsqueda de similitud semántica, se obtiene una mayor efectividad en las coincidencias que mantener o aumentar el número de palabras del registro original. Esto podría deberse a que, al parafrasear con palabras distintas en menor cantidad, el registro conserva las palabras clave con más peso en la interpretación del significado semántico. Por otro lado, el aumentar palabras que no aporten sentido semántico podría sesgar el significado del texto a buscar.

El mejor resultado que puede alcanzar la aplicación web es que en un 90% de las búsquedas de personas, donde se utilicen en su mayoría palabras clave en las descripciones, la persona buscada se encuentre dentro de los primeros 100 registros devueltos. Considerando que la base de datos de esta aplicación contiene 4,278 registros, explorar sólo

los primeros 100 en 9 de cada 10 búsquedas agiliza el proceso de identificación de personas.

Posiblemente, la mejor forma de identificar personas por descripciones textuales sea combinar técnicas tradicionales para comparación de similitud sintáctica con técnicas de comparación en el significado semántico, con el objetivo de mejorar la exactitud en las búsquedas.

7. Trabajo futuro

Como trabajo futuro, se plantea realizar una comparación entre el modelo USE utilizado en la presente propuesta y las ventajas que brindan las relaciones semánticas presentes en una base de datos basada en grafos, como NEO4J. Se busca discutir las fortalezas y debilidades de ambos enfoques. Además, se buscará mejorar el rendimiento de la aplicación mediante la creación de un modelo híbrido, que incorpore diferentes paradigmas de similitud textual, tales como el sintáctico y el orientado a grafos.

8. Referencias

adondevanlosdesaparecidos. (23 de junio de 2022). Desaparición de personas: un paradigma del crimen perfecto. A dónde van los desaparecidos. Recuperado de: https://adondevanlosdesaparecidos.org/2022/06/22/desaparicion-de-personas-un-paradigma-del-crimen-perfecto/

Balaha, H. M., & Saafan, M. M. (2021). Automatic Exam Correction Framework (AECF) for the MCQs, Essays, and Equations Matching. IEEE Access, 9, 32368-32389. Recuperado de: https://doi.org/10.1109/ACCESS.2021.3060940

Berryhill, J., Heang, K. K., Clogher, R., & McBride, K. (noviembre de 2019). Hello, World: Artificial intelligence and its use in the public sector. OECD. Recuperado de: https://doi.org/10.1787/726fd39d-en

Briseño, R. A. (s.f.). Alexagnus/Reniatexto. GitHub. Recuperado el 4 de noviembre de 2023, de https://github.com/alexAgnus/RENIAtexto

Cepeda, A., & Leetoy, S. (enero de 2021). De víctimas a expertas: estrategias de agencia cívica para la identificación de desaparecidos en México. Íconos - Revista de Ciencias Sociales, (69). Recuperado de: https://doi.org/10.17141/iconos.69.2021.4197

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., ... & Sung, Y.-H. (2018). Universal Sentence Encoder. arXiv preprint arXiv:1803.11175.

ChatGPT. (s.f.). Recuperado el 31 de agosto de 2023, de https://chat.openai.com

Facebook. (s.f.). Recuperado el 31 de agosto de 2023, de https://www.facebook.com/PorAmorAEllxs

Frikha, M., Fendri, E., & Hammami, M. (enero de 2021). Deep Semantic Attributes for People Search. Procedia Computer Science, 192, 90-99. Recuperado de: https://doi.org/10.1016/j.procs.2021.08.010

Hu, Y., Hu, C., Tran, T., Kasturi, T., Joseph, E., & Gillingham, M. (7 de febrero de 2021). What's in a Name? -- Gender Classification of Names with Character Based Machine Learning Models. arXiv.org. Recuperado de: https://arxiv.org/abs/2102.03692v1

Ibarra, J. (5 de mayo de 2022). Jalisco registra números rojos en desaparición de personas. Zona Docs. Recuperado de: https://www.zonadocs.mx/2022/05/05/jalisco-registra-numeros-rojos-en-desaparicion-de-personas/

IJCF-PFSI. (s.f.). Recuperado el 29 de agosto de 2023, de https://cienciasforenses.jalisco.gob.mx/registro_pfsi.php

Jeong, S., Oh, D., Park, K., & Lim, H. (enero de 2022). Considering Commonsense in Solving QA: Reading Comprehension with Semantic Search and Continual Learning. Applied Sciences, 12(9). Recuperado de: https://doi.org/10.3390/app12094099

LA DESAPARICIÓN FORZADA EN MÉXICO: UNA MIRADA DESDE LOS ORGANISMOS DEL SISTEMA DE NACIONES UNIDAS. (2019). Recuperado de https://www.cndh.org.mx/sites/default/files/documentos/2019-09/lib_DesaparicionForzadaMexicoUnaMirada.pdf

Li, H., Yan, Y., Wang, S., Liu, J., & Cui, Y. (marzo de 2023). Text classification on heterogeneous information network via enhanced GCN and knowledge. Neural Computing and Applications, 35(20), 14911-14927. Recuperado de: https://doi.org/10.1007/s00521-023-08494-0

Liu, G., Yang, G., Bai, S., Zhou, Q., & Dai, H. (2020). FSSE: An Effective Fuzzy Semantic Searchable Encryption Scheme Over Encrypted Cloud Data. IEEE Access, 8, 71893-71906. Recuperado de: https://doi.org/10.1109/ACCESS.2020.2966367

Mahajan, D., et al. (noviembre de 2020). Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning. JMIR Medical Informatics, 8(11), e22508. Recuperado de: https://doi.org/10.2196/22508

Martin-Rodilla, P., Hattori, M. L., & Gonzalez-Perez, C. (julio de 2019). Assisting Forensic Identification through Unsupervised Information Extraction of Free Text Autopsy Reports: The Disappearances Cases during the Brazilian Military Dictatorship. Information, 10(7). Recuperado de: https://doi.org/10.3390/info10070231

Merayo-Alba, S., Fidalgo, E., González-Castro, V., Alaiz-Rodríguez, R., & Velasco-Mata, J. (2019). Use of Natural Language Processing to Identify Inappropriate Content in Text. En Hybrid Artificial Intelligent Systems (pp. 254-263). Springer International Publishing. Recuperado de: https://doi.org/10.1007/978-3-030-29859-3_22

Morales, E. C. (2015). La desaparición: un problema que impacta a las personas vulnerables.

Nemeshaev, S., Barykin, L., & Dadteev, K. (enero de 2021). Selection of experts for scientific and technical expertise based on semantic search. Procedia Computer Science, 190, 643-646. Recuperado de: https://doi.org/10.1016/j.procs.2021.06.102

Nodehi, A. K., & Charkari, N. M. (septiembre de 2022). A metaheuristic with a neural surrogate function for Word Sense Disambiguation. Machine Learning with Applications, 9, 100369. Recuperado de: https://doi.org/10.1016/j.mlwa.2022.100369

Sferrazza Taibi, P. (julio de 2021). La búsqueda de personas desaparecidas: derecho humano de las víctimas y obligación internacional del Estado. Estudios constitucionales, 19(1), 265-308. Recuperado de: https://doi.org/10.4067/S0718-52002021000100265

Sheth, D., Gupta, A. R., & D'Mello, L. (noviembre de 2021). Using Universal Sentence Encoder for Semantic Search of Employee Data. En 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA) (pp. 1-4). Recuperado de: https://doi.org/10.1109/ICCICA52458.2021.9697114

Vargas Fara González, D. F., Tapia, F., Gallardo, A., & Samantha. (11 de noviembre de 2020). Jalisco: La verdad de los "tráileres de la muerte". Crisis Forense. Recuperado de: https://quintoelab.org/crisisforense/jalisco-la-verdad-de-los-traileres-de-la-muerte/

Vowinckel, K., & Hähnke, V. D. (junio de 2023). SEARCHFORMER: Semantic patent embeddings by siamese transformers for prior art search. World Patent Information, 73, 102192. Recuperado de: https://doi.org/10.1016/j.wpi.2023.102192

What is Tokenization? Types, Use Cases, Implementation. (s.f.). Recuperado el 18 de diciembre de 2023, de https://www.datacamp.com/blog/what-is-tokenization

Zhou, H., Li, F., Tian, X., & Huang, Y. (2023). Feature semantic alignment and information supplement for Text-based person search. Frontiers in Physics, 11. Recuperado de: https://www.frontiersin.org/articles/10.3389/fphy.2023.1192412

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Jalisco

www.iieg.gob.mx
Dirección de Información de Gobierno, Seguridad Pública e Impartición de Justicia