**PROPUESTA DE TEMA DE TESIS DE DOCTORADO**

**DOCTORADO EN CIENCIAS DE LA INGENIERÍA**

**TÍTULO DEL TEMA DE TESIS PROPUESTO**

**Interfaz de Lenguaje Natural para la Consulta de Información en un Lago de Datos Hospitalario**

**PROPONENTE**

**Jonathan Zavala Díaz**

**DIRECTOR DE LA TESIS**

**Dr. Juan Carlos Olivares Rojas**

**Palabras clave:** Lenguaje natural, NPL, Interfaz, español, Hospitalario.

**Índice**

**1. Resumen 1**

**2. Introducción 2**

**3. Antecedentes 2**

**4. Marco Teórico 2**

**5. Objetivos 2**

***5.1 Objetivo General* 2**

***5.2 Objetivos Particulares* 2**

**6. Metas 2**

**7. Impacto 2**

**8. Metodología 2**

**9. Programa de actividades, calendarización 2**

**10. Productos Entregables 2**

**11. Vinculación con otras instituciones, empresas o sectores 2**

**12. Referencias 2**

**1. Resumen**

El área de la salud es una de las áreas más beneficiadas con el uso de las TICs en los procesos de cuidado de la salud y diagnóstico de los pacientes. Al igual que muchas otras áreas, los datos de salud crece de forma rápida y en grandes volúmenes en poco tiempo. Esto presenta enormes ventajas, pero también muchos retos como el detalle de poder realizar consultas y predicciones en poco tiempo. Los lenguajes de consultas generalmente no son amigables ni dinámicos para los usuarios finales. En los últimos años el avance en el procesamiento del lenguaje natural y de la inteligencia artificial a permitiendo la masificación de aplicaciones como los chatbots y asistentes digitales que permiten una comunicación de forma natural en diversos idiomas como el inglés. Desafortunadamente, la información médica es de un contexto muy específico, así como los desarrollos existentes están muy enfocado en otros idiomas como el inglés, dejando a los especialistas de la salud hispanoparlantes muy rezagado.

**2. Introducción**

Con los avances y crecimiento en las Tecnologías de la Información y la Comunicación (TICs), grandes volúmenes de datos se siguen acumulando velozmente en los bancos de datos de cada organización. Mas sin embargo la mayor parte de esta información nunca se puede utilizar para generar un beneficio real para la organización, para ello se debe ser capaz de convertir la información en conocimiento. El conocimiento se puede utilizar para aliviar el proceso de toma de decisiones por parte de los expertos en el área dominio [1].

En los últimos 20 años, la recopilación y el almacenamiento de datos hospitalarios se ha incrementado enormemente con el uso generalizado de los sistemas de información clínica, los cuales contienen grandes cantidades de datos sobre la salud y atención medica de los pacientes [2]. Una amplia gama de estos datos se encuentra comúnmente dentro de narrativas clínicas las cuales son del tipo de datos no estructurados. Debido a la naturaleza de estos tipos de datos, un Data Lake es el tipo de almacenamiento que mejor se adapta a ello, ya que esta diseñado para almacenar datos sin procesar (Estructurados, no estructurados, semiestructurados y binarios). Los informes narrativos permiten la flexibilidad de expresión como dudas, negaciones o hipótesis diagnósticas y la representación compleja de enfermedades, examen clínico, historial del paciente y antecedentes médicos familiares [3]. La mayoría de los registros médicos actuales conservan un gran elemento de texto libre. Si bien esto es atractivo para la mayoría de los usuarios finales debido a la flexibilidad de expresión, crea desafíos para el uso continuo de la información contenida en las notas [4]. El procesamiento del lenguaje natural (NLP, por sus siglas en inglés) es una forma de aprendizaje automático que se puede utilizar en este contexto para procesar y analizar elementos de texto libre, por lo cual puede ayudar en la predicción de los resultados de los pacientes, aumentar los sistemas de clasificación de hospitales y generar modelos de diagnóstico que detectan enfermedades crónicas en etapa temprana [4].

El NLP ahora se usa cada vez más en la medicina para mejorar la utilización de registros de salud electrónicos no estructurados y para proporcionar una forma de comunicación con los pacientes para responder preguntas y realizar consultas [5]. Usar el lenguaje natural es una forma más fácil para recuperar información de registros de salud. Las computadoras no pueden entender el lenguaje natural por lo que se necesita una interfaz; esa es la razón para desarrollar una interfaz de lenguaje natural de consulta de información. La interfaz de lenguaje natural es capaz de traducir la consulta de lenguaje natural dada por el usuario a una equivalente en lenguaje de consulta. Se han desarrollaron interfaces de lenguaje natural para bases de datos y así convertir el lenguaje natural a una consulta SQL y obtener el resultado correspondiente de la base de datos. Todavía hay mucho trabajo de investigación en el campo de la interfaz de lenguaje natural y se están desarrollando nuevas interfaces para las bases de datos que brindan respuestas más precisas.

Con el avance en el poder de procesamiento del hardware se han desarrollado herramientas basadas en lenguaje natural para consulta de información, más sin embargo dichas herramientas están diseñadas para un sistema de base de datos sin tomar en cuenta datos no estructurados, donde se necesitan otras técnicas para analizar la información contenida en texto libre [6]. Debido a que los desarrollos más recientes de interfaces de lenguaje natural se centran en entender la pregunta y trasladarla a un lenguaje de consulta como SQL con la finalidad de obtener la información de una base de datos, se necesita crear una herramienta que pueda interpretar la pregunta en lenguaje natural y realizar con la ayuda de NLP y Machine Learning el análisis en datos no estructurados como narrativas clínicas para encontrar información requerida por el usuario. En la literatura de los últimos años han propuesto herramientas basadas en lenguaje natural para consulta de información, más sin embargo dichas herramientas están desarrolladas para idiomas principalmente en inglés y algunos otros idiomas específicos como francés, afgano y cingalés [2], [7], [8]. Por lo cual nace la necesidad de crear una herramienta especial para el idioma español.

Lo expuesto anteriormente ha sido la principal motivación para llevar a cabo el trabajo de investigación que se propone en este protocolo. Cuyo objetivo principal es desarrollar soluciones basadas en tecnologías de procesamiento de lenguaje natural, tecnologías de la información y Machine Learning para convertir la información en conocimiento que pueda ayudar a encontrar patrones de alguna enfermedad en narrativas clínicas. Para cumplir con este objetivo, se ha seguido una metodología propuesta en este protocolo.

**3. Antecedentes**

Un gran número de investigadores han enfocado sus esfuerzos en desarrollar herramientas basadas en lenguaje natural para consulta de información como Interfaces en Lenguaje Natural (NLI, por sus siglas en inglés), dichas herramientas enfocadas a diversos dominios y aplicando diversas técnicas de procesamiento de lenguaje. A continuación, se presentan los trabajos más relevantes encontrados en esta área de NLP.

En el trabajo publicado por Rencis [9] se presenta un lenguaje natural controlado, más sin embargo solo es el primer paso para la creación de una interfaz basada en lenguaje natural para consulta de información. En los trabajos [10] y [11] interfaces de lenguaje natural para visualización de datos ambos en el idioma inglés.

Otra herramienta es Doc'EDS presentada en [2], una herramienta de búsqueda semántica francesa para consultar documentos de salud de un Data Warehouse clínico, si bien el sistema proporciona una interfaz fácil de usar, está diseñada únicamente para el idioma francés. En [7] , [8] y [12] presentan interfaces en leguaje natural en las cuales traducen las consultas de lenguaje natural a un lenguaje de consulta para eztraer información de bases de datos, recalcando de estos trabajos el enfoque a datos estructurados solamente.

La herramienta MedCAT presentada en **[13]** se trata de un el kit de herramientas de anotación de conceptos médicos de código abierto que proporciona un novedoso algoritmo de aprendizaje automático auto-supervisado para extraer conceptos utilizando cualquier vocabulario de conceptos y una interfaz .

En el trabajo publicado por Trivedi et al. [14] presentan NLPReViz: una herramienta interactiva para el procesamiento del lenguaje natural en texto clínico, utilizan la técnica Bag-of-words de NLP, también se encuentra desarrollado para el idioma inglés.

En el trabajo publicado por Rojas et al. [15] presentan Clinical Flair, un modelo de lenguaje de dominio específico entrenado en narrativas clínicas en español. Clinical Flair, un modelo de lenguaje a nivel de carácter para la PNL clínica en español.

Los autores en [16] utilizan técnicas de procesamiento de lenguaje natural como base de su algoritmo de desidentificación para eliminar datos personales de las historias clínicas y así proteger su identidad.

En la Tabla 1 se presentan los trabajos mencionados anteriormente destacando sus características como el idioma el ámbito, técnicas utilizadas y si desarrollan una interfaz de lenguaje natural.

Tabla 1.- Trabajos encontrados en la literatura sobre NLP y herramientas basadas en lenguaje natural.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trabajo | NLI | Idioma | Ámbito | Técnicas o herramientas utilizadas |
| Rencis [9] | No | Letón | Biomédico | Lenguaje natural controlado |
| Álvarez et al. [16] | Sí | Español | Biomédico | Librería Flair |
| Setlur et al. [10] | Sí | Inglés | Demografía/ Geografía | Gramática probabilística |
| Pressat-Laffouilhère et al. [2] | Sí | Francés | Biomédico | Algoritmos de procesamiento del lenguaje natural |
| Peduru-Hewa et al. [7] | Sí | Cingalés | Genérico | Método propuesto de conversión a SQL |
| S. Karimi, A. A. Rasel & M. S. Abdullah [8] | Sí | Afgano | Genérico | Técnica de mapeo de palabras |
| Yu & Silva [11] | Sí | Inglés | Genérico | Algoritmos de procesamiento del lenguaje natural |
| KraljevicYu et al. [13] | Sí | Inglés | Biomédico | Named Entity Recognition |
| Das & Balabantaray [12] | Sí | Inglés | Genérico | Técnica de mapeo de palabras |
| Trivedi et al. [14] | Sí | Inglés | Biomédico | Bag-of-words |
| Rojas et al. [15] | No | Español | Biomédico | Named Entity Recognition |

**4. Marco teórico**

El lenguaje natural se refiere a la forma en que las personas se comunican y el procesamiento de lenguaje natural se encarga de analizar y procesar el lenguaje humano a través de herramientas y tecnologías de software. Esto involucra distintas áreas de la computación, tales como inteligencia artificial, lingüística computacional, etc. Mediante el procesamiento del lenguaje natural es posible procesar documentos de texto, mensajes SMS, email, páginas web, etc. y organizar el conocimiento para realizar tareas como análisis de sentimientos, análisis de contexto, generación de resúmenes, traducción automática, sistemas de diálogos, etc. [17]

Un concepto importante en el procesamiento del lenguaje natural son las **word embeddings** ya que se ha demostrado que si son previamente entrenadas son de gran utilidad para tareas posteriores de NLP, tanto por su capacidad para ayudar al aprendizaje y la generalización con información aprendida de datos no etiquetados, como por la relativa facilidad de incluirlos en cualquier proceso de aprendizaje [18].

En la última década, junto con el crecimiento del aprendizaje profundo, las representaciones (embeddings) basadas en redes neuronales han reemplazado casi por completo a los modelos convencionales basados en conteo y han dominado el campo. Dado que las **word embeddings** neuronales generalmente se entrenan con algún tipo de objetivo de modelado de lenguaje, como predecir una palabra faltante en un contexto, también se conocen como modelos predictivos [19]. Las incrustaciones de palabras fueron popularizadas por Word2vec, en [20] utilizan Word2Vec y BERT.

Existen herramientas avanzadas para el procesamiento del lenguaje natural como Apache OpenNLP [21], Stanford CoreNLP [22] y Stanza [23].

OpenNLP:

La biblioteca Apache OpenNLP es un conjunto de herramientas basado en el aprendizaje automático para el procesamiento de texto en lenguaje natural. OpenNLP es compatible con las tareas de NLP más comunes, como la tokenización, la segmentación de oraciones, el etiquetado de partes del discurso, la extracción de entidades nombradas, la fragmentación, el análisis, la detección de idiomas y la resolución de correferencias.

CoreNLP:

Stanford CoreNLP es nuestro conjunto de herramientas de Java que proporciona una amplia variedad de herramientas de PNL. CoreNLP es su ventanilla única para el procesamiento de lenguaje natural en Java. CoreNLP permite a los usuarios derivar anotaciones lingüísticas para el texto, incluidos límites de tokens y oraciones, partes del discurso, entidades nombradas, valores numéricos y de tiempo, análisis de dependencia y circunscripción, correferencia, sentimiento, atribuciones de citas y relaciones. CoreNLP actualmente admite 8 idiomas: árabe, chino, inglés, francés, alemán, húngaro, italiano y español.

Stanza:

Stanza es una nueva biblioteca de NLP de Python que incluye una canalización de NLP neuronal multilingüe y una interfaz para trabajar con Stanford CoreNLP en Python. Stanza es una colección de herramientas precisas y eficientes para el análisis lingüístico de muchos lenguajes humanos. Desde texto sin procesar hasta análisis sintáctico y reconocimiento de entidades, Stanza trae modelos de NLP de última generación a los idiomas de su elección.

**5. Objetivos**

**5.1 Objetivo general**

Desarrollar un interfaz de Lenguaje Natural para la Consulta de Información en un Lago de Datos Hospitalario

**5.2 Objetivos particulares.**

1. Investigar diversas técnicas y herramientas de lenguaje natural en español para el área de la salud.
2. Estudiar de Técnicas de conversión a lenguajes de consulta de datos.
3. Implementar un lago de datos hospitalarios provenientes de diversas fuentes y en diversos formatos.
4. Desarrollar interfaces en lenguaje natural (español) que permita la obtención de datos de sistemas hospitalarios en forma sencilla y funcional.
5. Realizar modelos de aprendizaje automático y de análisis de datos que permitan mostrar información de pronósticos.
6. Desarrollar una interfaz de usuario para visualizar la información proveniente de la interfaz de lenguaje natural

**6. Metas.**

1. 1 tabla con al menos 10 técnicas y herramientas de lenguaje natural en español para el área de la salud, resaltando sus características, evaluarlas, para elegir 3 de ellas con las cuales trabajar.
2. 1 tabla con diferentes Técnicas de conversión a lenguajes de consulta de datos y elegir la que tenga mayores prestaciones y se adapte a nuestros requerimientos.
3. 1 implementación lago de datos hospitalarios provenientes de diversas fuentes y en diversos formatos.
4. 1 interfaz en lenguaje natural en el idioma que permita la obtención de datos de sistemas hospitalarios en forma sencilla y funcional.
5. 1 implementación de modelos de aprendizaje automático y de análisis de datos que permitan mostrar información de pronósticos.
6. 1 interfaz de usuario para visualizar la información proveniente de la interfaz de lenguaje natural

**7. Impacto.**

Impacto o beneficio en la solución a un problema relacionado con el sector productivo o la generación del conocimiento científico o tecnológico. Sustente la realización de su proyecto respecto a la magnitud del problema, la trascendencia de su estudio, su factibilidad, vulnerabilidad e impacto (social, económico, ambiental, etc.) en congruencia con la línea de investigación del programa de doctorado (½ cuartilla).

**8. Metodología.**

La metodología propuesta a seguir durante el desarrollo de este proyecto de tesis se divide en cuatro partes principales: en la primera se desarrolla un estudio del estado del arte que permita conocer los esfuerzos de investigación más relevantes dentro de las áreas de interés del proyecto; la segunda parte consiste en la formalización de los métodos propuestos en este trabajo para interfaces de lenguaje natural enfocadas a bases de conocimiento; en la tercera etapa, se lleva a cabo la implementación de los métodos propuestos; finalmente, en la cuarta parte se lleva a cabo la validación de la propuesta en las bases de conocimiento de DBpedia y MusicBrainz.

* Estudio del estado del arte.

En esta parte de la metodología se llevó a cabo un análisis de todos aquellos desarrollos de última tecnología realizados en los contextos de PLN e interfaces de lenguaje natural, principales tecnologías involucradas en este trabajo de tesis.

* + Procesamiento de lenguaje natural: Análisis de los diferentes niveles de PLN, así como las diversas aplicaciones de esta tecnología.
  + Interfaces de lenguaje natural: Análisis de las principales arquitecturas utilizadas en el desarrollo de este tipo de aplicaciones y de los esfuerzos de investigación más sobresalientes enfocados en proveer soluciones de este tipo.
* Formalización de la propuesta.

Esta parte de la metodología contempla el desarrollo de una ontología enfocada a la descripción de la estructura sintáctica de la pregunta, así como de su contexto en términos de la base de conocimiento del dominio. Toda la información a almacenar en esta ontología será obtenida mediante el proceso de análisis de preguntas en lenguaje natural diseñado en este trabajo, el cual estará basado en técnicas tales como el análisis de dependencias, lematización y la búsqueda de sinónimos. Finalmente, la creación de consultas SPARQL a partir de una representación semántica de la pregunta, permitirá generar el patrón de grafo a ser obtenido de la base de conocimiento para dar respuesta a la pregunta provista.

* Implementación de la propuesta.

Esta etapa consiste en la implementación de la interfaz de lenguaje natural propuesta por medio de herramientas de PLN. La interfaz implementada agrupa el modelo ontológico independiente del dominio para la representación de preguntas en lenguaje natural, el proceso de análisis de preguntas basado en técnicas de PLN y el proceso de generación de consultas SPARQL a partir de la representación semántica de la pregunta en lenguaje natural.

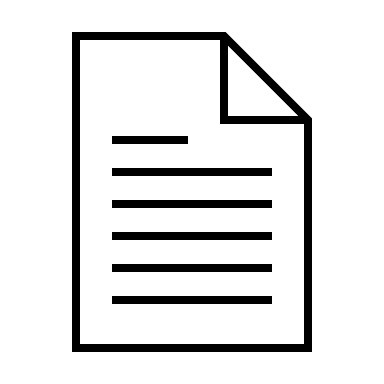
* Validación de la propuesta.

Finalmente, esta parte de la metodología contempla la validación de la interfaz de lenguaje natural implementada en bases de conocimiento basadas en Linked Data. En concreto, la interfaz desarrollada se aplicará sobre el conjunto de datos de DBpedia, un esfuerzo comunitario para extraer conocimiento estructurado y multilingüe de Wikipedia, y el conjunto de datos de MusicBrainz, una fuente de información ampliamente utilizada en el dominio de la música.

Arquitectura propuesta

La interfaz que se propone recibe una consulta formulada en lenguaje natural por los usuarios en idioma español, después la traduce al lenguaje de consulta y mediante ésta, extrae de la ontología del usuario la información solicitada. Su arquitectura está dividida en dos módulos principales: a) el módulo generador de conocimiento (MGC) y b) el módulo de la interfaz (MI). El MGC se encarga de generar el conocimiento que el MI necesita para responder a las consultas formuladas en lenguaje natural por los usuarios. En la Sección 3.1 se describe el MGC y en la Sección 3.2 se describe el MI. En la Fig. 1 se presenta la arquitectura general de la interfaz propuesta.

Icono

Descripción generada automáticamente

**Narrativas clínicas**

**Clasificaciion de la pregunta???**

**Modulo generador de conocimiento**

**Base de conocimiento / Data Lake**

**Tratamiento semántico**

**Interfaz de Lenguaje Natural**

**Modulo NLP**

**Consulta en lenguaje natural (español)**

**Respuesta a consulta**

Modulo NLP

* Tokenización
* Etiquetado Gramatical
* Lematización

Tratamiento semántico

* Identificación de entidades nombradas
* Búsqueda de sinónimos
* Detección de elementos de la base del conocimiento
* Análisis de dependencias
* Desambiguación

Modulo generador de conocimiento

* ---
* Etiquetado Gramatical
* Lematización

Diagrama

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente

**9. Programa de actividades, calendarización.**

Gráfico

Descripción generada automáticamente

**10. Productos entregables.**

* Software - Interfaces en lenguaje natural (español) que permita la obtención de datos de sistemas hospitalarios en forma sencilla y funcional.
* Método para modelos de aprendizaje automático y de análisis de datos que permitan mostrar información de pronósticos.
* Software - Interfaz de Lenguaje Natural para la Consulta de Información en un Lago de Datos Hospitalario
* Presentación en congreso internacional
* 1 publicación en el índice JCR
* Documento de tesis
* Certificado Toefl con 550 puntos

**11. Vinculación con otras instituciones, empresas o sectores.**

Pendiente…

**12. Referencias.**

[1] E. Rencis, “Application of a Configurable Keywords-Based Query Language to the Healthcare Domain,” *J. Adv. Inf. Technol.*, vol. 12, no. 2, pp. 142–147, 2021, doi: 10.12720/jait.12.2.142-147.

[2] T. Pressat-Laffouilhère *et al.*, “Evaluation of Doc’EDS: a French semantic search tool to query health documents from a clinical data warehouse,” *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 34, Dec. 2022, doi: 10.1186/s12911-022-01762-4.

[3] N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, and A. Burgun, “Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse,” *J. Am. Med. Informatics Assoc.*, vol. 24, no. 3, pp. 607–613, May 2017, doi: 10.1093/jamia/ocw144.

[4] S. Locke, A. Bashall, S. Al-Adely, J. Moore, A. Wilson, and G. B. Kitchen, “Natural language processing in medicine: A review,” *Trends Anaesth. Crit. Care*, vol. 38, pp. 4–9, Jun. 2021, doi: 10.1016/j.tacc.2021.02.007.

[5] J. Wang *et al.*, “Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed,” *J. Med. Internet Res.*, vol. 22, no. 1, p. e16816, Jan. 2020, doi: 10.2196/16816.

[6] E. U. Reshma and P. C. Remya, “A review of different approaches in natural language interfaces to databases,” in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2017, pp. 801–804. doi: 10.1109/ISS1.2017.8389287.

[7] D. S. Peduru Hewa and C. Farook, “A Sinhala Natural Language Interface for Querying Databases Using Natural Language Processing,” in *2021 21st International Conference on Advances in ICT for Emerging Regions (ICter)*, Dec. 2021, pp. 213–218. doi: 10.1109/ICter53630.2021.9774794.

[8] S. Karimi, A. A. Rasel, and M. S. Abdullah, “Natural Language Query and Control Interface for Database Using Afghan Language,” in *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, Aug. 2022, pp. 1–8. doi: 10.1109/INISTA55318.2022.9894168.

[9] E. Rencis, “Towards a natural language-based interface for querying hospital data,” in *Proceedings of 2018 International Conference on Big Data Technologies - ICBDT ’18*, 2018, pp. 25–28. doi: 10.1145/3226116.3226133.

[10] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang, “Eviza: A natural language interface for visual analysis,” in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, Oct. 2016, pp. 365–377. doi: 10.1145/2984511.2984588.

[11] B. Yu and C. T. Silva, “FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System,” *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 1–11, Jan. 2020, doi: 10.1109/TVCG.2019.2934668.

[12] A. Das and R. C. Balabantaray, “MyNLIDB: A Natural Language Interface to Database,” in *2019 International Conference on Information Technology (ICIT)*, Dec. 2019, pp. 234–238. doi: 10.1109/ICIT48102.2019.00048.

[13] Z. Kraljevic *et al.*, “Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit,” *Artif. Intell. Med.*, vol. 117, p. 102083, Jul. 2021, doi: 10.1016/j.artmed.2021.102083.

[14] G. Trivedi, P. Pham, W. W. Chapman, R. Hwa, J. Wiebe, and H. Hochheiser, “NLPReViz: an interactive tool for natural language processing on clinical text,” *J. Am. Med. Informatics Assoc.*, vol. 25, no. 1, pp. 81–87, Jan. 2018, doi: 10.1093/jamia/ocx070.

[15] M. Rojas, J. Dunstan, and F. Villena, “Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing,” in *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 2022, pp. 87–92. doi: 10.18653/v1/2022.clinicalnlp-1.9.

[16] C. Álvarez *et al.*, “Estudio longitudinal para el desarrollo de modelos predictivos de complicaciones crónicas de la diabetes mellitus tipo 2,” *Komput. Sapiens*, vol. 3, pp. 10–15, 2022, [Online]. Available: http://komputersapiens.smia.mx/publicaciones.php#KSXIV-III

[17] F. Pech-May, L. A. López-Gómez, and J. Magaña-Govea, “Procesamiento de lenguaje natural con aprendizaje profundo,” *Komput. Sapiens*, vol. 2, pp. 56–61, 2019, [Online]. Available: http://komputersapiens.smia.mx/publicaciones.php#KSXI-II

[18] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “FLAIR: An easy-to-use framework for state-of-the-art NLP,” in *Proceedings of the 2019 Conference of the North*, 2019, pp. 54–59. doi: 10.18653/v1/N19-4010.

[19] M. T. Pilehvar and J. Camacho-Collados, “Word Embeddings,” 2021, pp. 25–40. doi: 10.1007/978-3-031-02177-0\_3.

[20] J.-C. Hernández-Hernández, D. Juárez-Morales, J.-J. Guzmán-Landa, and G. J. Hoyos-Rivera, “Análisis de Sentimientos en Twitter,” *Komput. Sapiens*, vol. 2, pp. 59–63, 2022, [Online]. Available: http://komputersapiens.smia.mx/publicaciones.php#KSXIV-II

[21] “Apache OpenNLP.” https://opennlp.apache.org/

[22] “CoreNLP.” https://stanfordnlp.github.io/CoreNLP/

[23] “Stanza.” https://stanfordnlp.github.io/stanza/