INSTITUTO TECNOLÓGICO SUPERIOR DE TEZIUTLÁN

Tesis

"Asistente virtual con implementación de PNL en plataforma de intercambio económico cultural para el fortalecimiento de lenguas indígenas"

PRESENTA:

ADOLFO ORTIZ BAUTISTA

CON NÚMERO DE CONTROL

19TE0477

PARA OBTENER EL TÍTULO DE:

INGENIERO EN INFORMATICA

CLAVE DEL PROGRAMA ACADÉMICO

IINF-2010-220

DIRECTOR (A) DE TESIS:

M. EDGAR DEGANTE AGUILAR

"La Juventud de hoy, Tecnología del Mañana"

TEZIUTLÁN, PUEBLA, FEBRERO 2024

PRELIMINARES

Agradecimientos

Deseo transmitir mi agradecimiento más profundo y sincero a todos aquellos que, de una forma u otra, aportaron de manera significativa a la culminación de esta investigación y tesis.

Primero que nada, deseo expresar mi gratitud a mi familia, quienes me proporcionaron un apoyo constante en cada fase de este recorrido académico. Sus palabras de motivación me impulsaron a superar los retos y dificultades que surgieron en el trayecto.

Mi sincero agradecimiento también va dirigido a mi asesor académico, por su paciencia, dedicación y sus valiosos conocimientos que fueron fundamentales a lo largo de este proceso para dar forma a mis ideas y guiar mi investigación.

No puedo dejar de mencionar a todos mis profesores y amigos, quienes compartieron conmigo sus experiencias y conocimientos a lo largo de la carrera.

Finalmente, pero no menos importante, deseo expresar mi agradecimiento a todos aquellos que, directa o indirectamente, tuvieron un impacto en mi educación y en la elaboración de esta tesis.

Resumen

El siguiente trabajo de investigación se centra en el desarrollo y el entrenamiento de un Asistente Virtual utilizando tecnología de Procesamiento del Lenguaje Natural (PNL). El asistente será implementado en una plataforma de intercambio económico y cultural con el objetivo principal de fortalecer las lenguas indígenas.

En el panorama actual, las lenguas indígenas se encuentran en peligro de extinción debido a la falta de recursos y el predominio de idiomas más extendidos. Este estudio aborda este problema proporcionando una solución tecnológica que facilite la comunicación y promueva la conservación de las lenguas indígenas.

Para alcanzar este objetivo, se realizó el entrenamiento y creación de un Asistente Virtual fundamentado en PNL. Este asistente es capaz de realizar traducciones del español al náhuatl. El modelo de traducción emplea técnicas sofisticadas de procesamiento de lenguaje natural para incrementar la calidad de las traducciones y asegurar una experiencia eficaz.

Esta investigación presenta un enfoque innovador para el fortalecimiento de lenguas indígenas, por medio del entrenamiento y desarrollo de un Asistente Virtual con PNL en una plataforma de intercambio económico y cultural. Esta herramienta se presenta como un recurso valioso para la preservación de las lenguas indígenas.

Introducción

Nos hace mención Monsiváis (2019) que, México se encuentra dentro de los países con mayor cantidad de pueblos indígenas y por consiguiente de hablantes de distintas lenguas. Sin embargo, estos pueblos enfrentan desafíos significativos para su integración y desarrollo.

Aunque las etnias en México contribuyen significativamente a la cultura y economía del país, lamentablemente, enfrentan discriminación constante hacia su identidad étnica y cultural, esta marginación proviene principalmente de la población mestiza de México.

Por lo tanto, muchos de estos individuos no tienen conocimientos del idioma predominante en México, el español, lo que contribuye a su exclusión. Uno de los desafíos más grandes es la barrera del idioma.

Además, las lenguas indígenas son un reflejo de la rica diversidad cultural de México. Sin embargo, con la llegada de la era digital y la globalización, están en peligro de extinción. Por lo tanto, es esencial buscar estrategias para preservarlas y fortalecerlas. Esta tarea es de suma importancia.

Como nación pluricultural y multilingüe, México tiene el deber de preservar su riqueza en diversidad lingüística. Existe una responsabilidad significativa en lograr que las lenguas indígenas coexistan con el español, promoviendo así la interacción entre los diferentes grupos étnicos. En este escenario, la tecnología puede desempeñar un papel fundamental.

Es crucial subrayar que la diversidad de lenguas no solo realza la cultura, sino que también puede ser un motor para la innovación y el progreso. Cada lengua proporciona una visión única del mundo, y al promover su uso y estudio, se

pueden explorar nuevas formas de pensamiento y creatividad. Además de facilitar la comunicación, el aprendizaje de estas lenguas puede tener otros beneficios.

El uso de la tecnología resulta ser una estrategia favorable para el desarrollo de asistentes virtuales con la capacidad de comprender y generar textos en lenguas indígenas (español a náhuatl) y viceversa.

Se sugiere una solución que implica la creación de un asistente virtual que emplea el Procesamiento de Lenguaje Natural (PNL). Este asistente no solo facilitará la comunicación en lenguas indígenas, sino que también ayudará a reforzar estas lenguas al fomentar su uso en un ambiente digital.

La integración de PNL hará posible que el asistente virtual tenga la capacidad de comprender y generar textos en lenguas indígenas de tal forma que facilitaría la interacción entre distintos usuarios al estar presente en una plataforma de intercambio económico cultural.

El valor de aprender un segundo idioma no se restringe solo a los idiomas extranjeros, ya que puede proporcionar una mayor interacción y un acceso más extenso al aprendizaje de un idioma indígena. Al promover su estudio, se puede mitigar la disminución en el número de hablantes y, por ende, prevenir la desaparición de estas lenguas.

Índice

Capítulo I
Generalidades del proyecto

1.1 Descripción de la empresa

1.1.1 Antecedentes

El Instituto Tecnológico Superior de Teziutlán, ubicado en Teziutlán, Puebla, México, es una institución educativa que tiene sus raíces en diversas industrias locales como la minería, la fruticultura, la ganadería y la confección de ropa. Se estableció en 1994 a través de negociaciones con la Secretaría de Educación Pública, aunque comenzó a operar un año antes, en 1993, ofreciendo programas en Ingeniería Industrial y Administración. Fue la primera institución tecnológica descentralizada en Puebla, bajo la dirección inicial de José Emilio Guillermo Ortega Balbuena.

Originalmente, el instituto se ubicaba en el "Centro de Bachillerato Tecnológico, Industrial y de Servicios No. 44", pero debido al crecimiento de la matrícula estudiantil, tuvo que trasladarse en dos ocasiones. Primero, se mudó a una granja avícola y luego, en 2001, a un terreno cercano a una antigua mina de cobre, después de sufrir daños por la depresión tropical "IRENE".

A lo largo del tiempo, el Instituto Tecnológico Superior de Teziutlán ha mejorado significativamente su infraestructura, talleres y laboratorios. En 2006, logró la certificación ISO 9001-2008 y obtuvo acreditaciones en varias disciplinas, incluyendo Informática, Administración, Ingeniería Industrial e Ingeniería en Sistemas Computacionales.

Actualmente, el instituto ofrece una variedad de programas de estudio, que incluyen:

- Ingeniería en Gestión Empresarial

- Ingeniería en Industrias Alimentarias
- Ingeniería en Sistemas Computacionales
- Ingeniería Industrial
- Ingeniería Informática

• Ingeniería Mecatrónica

El instituto se dedica a proporcionar una educación de alta calidad, actualizada y orientada al servicio, que se adapta a las necesidades de la sociedad. Para ello, utiliza sus recursos de manera eficiente y cumple con sus programas de trabajo.

Teléfono

(231)3114000.

1.1.2 Misión y visión

Misión

El instituto Tecnológico Superior de Teziutlán tienen como Misión, formar Profesionales que se constituyan en agentes de cambio y promuevan el desarrollo integral de la sociedad, mediante la implementación de procesos académicos de calidad.

Visión

Llegar a ser la Institución de Educación Superior Tecnológica más reconocida en el Estado de Puebla, que ofrezca un proceso de Enseñanza – Aprendizaje certificado, comprometido con la excelencia académica y la formación integral del Alumno, contribuyendo al desarrollo sustentable, económico, político y social de nuestro Estado.

1.1.3 Valores

Integridad

Actuar con rectitud, honestidad, honradez y transparencia, de manera congruente, sin engaños, ni falsedades en la realización de sus funciones.

Compromiso

Cumplir con la sociedad ofreciéndoles profesionales capaces y comprometidos con su región y el Estado para satisfacer las necesidades presentes y futuras.

Creatividad

Mantener una actitud constructiva, considerando la mejora continua y la innovación.

Lealtad

Ajustar su actuación al compromiso personal con los objetivos del ITST, de tal modo que se refleje y fortalezca el conjunto de logros del Instituto.

Actitud de servicio

Fomentar en el alumno el deseo de servir a su comunidad y su identificación plena con el instituto a colaborar en todas y cada una de las actividades programadas, así como la aplicación de las políticas y procedimientos una vez que se integren al sector productivo.

Legalidad

Conocer y cumplir la normativa aplicable a las actividades relativas a su ámbito de competencia.

1.1.4 Macro localización y Micro localización

Macro localización

El Instituto Tecnológico Superior de Teziutlán está ubicado en el municipio de Teziutlán, perteneciente al estado de puebla.

Ilustración 1.1 *Municipio de* Teziutlán, Pue.

Fuente: EcuRed, 2023

Micro localización

El Instituto Tecnológico Superior de Teziutlán tiene su ubicación en: Fracción I y II S/N, Aire Libre, Teziutlán, con código postal 73960.

Ilustración 1.2 *Micro localización* Tec de Teziutlán vista de mapa

Fuente: *Google Maps, 2023.*

1.2 Problemas de investigación a resolver

La capacitación de un Asistente Virtual que utiliza el Procesamiento del Lenguaje Natural (PNL) en una plataforma de intercambio cultural y económico para el fortalecimiento de las lenguas indígenas es un tema de gran importancia para la preservación y difusión de la diversidad lingüística de México.

Sin embargo, existen varios retos para el desarrollo de este tipo de asistentes, como la escasez de recursos lingüísticos, la variabilidad dialectal, la adaptación al contexto y la evaluación de la calidad.

Por lo tanto, el problema que se plantea en esta investigación es: ¿Cómo diseñar e implementar un Asistente Virtual que utilice el PNL para apoyar el aprendizaje y la revitalización de las lenguas indígenas?

1.2.1 Preservación de lenguas indígenas

El desafío principal se encuentra en la amenaza persistente de extinción de lenguas indígenas, como el náhuatl, debido a la escasez de recursos y al predominio de idiomas más extendidos. Este trabajo de investigación se enfoca en cómo la creación de un Traductor de Español a Náhuatl puede contribuir de manera efectiva a la conservación y revitalización de este idioma, así como de otras lenguas indígenas.

1.2.2 Calidad de las traducciones

Un desafío crítico es garantizar la precisión y calidad de las traducciones proporcionadas por el Asistente Virtual. La investigación se centrará en cómo lograr traducciones precisas y contextualmente relevantes, dado que las lenguas

indígenas pueden tener particularidades culturales y contextuales que deben ser reflejadas en las traducciones.

1.2.3 Impacto en la revitalización cultural

Un aspecto importante es evaluar el impacto de la plataforma en la revitalización cultural y en la identidad de las comunidades indígenas. Esto implica examinar si el uso del náhuatl en la comunicación diaria se fortalece y si se promueve un mayor intercambio cultural y económico entre las comunidades indígenas y la sociedad en general.

Este estudio se enfocará en estos desafíos de investigación con el objetivo de brindar un entendimiento más detallado de cómo la implementación de un Asistente Virtual con PNL puede ayudar a la conservación de las lenguas indígenas y al fortalecimiento de los vínculos culturales y económicos en un mundo cada vez más globalizado.

1.3 Preguntas de investigación

¿Puede un Asistente Virtual con tecnología de Procesamiento del Lenguaje Natural (PNL) mejorar significativamente la traducción precisa del español al náhuatl y viceversa?

¿Puede un Asistente Virtual contribuir a la preservación de las lenguas indígenas y al fortalecimiento de las conexiones culturales en un mundo globalizado?

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar un servicio de traducción para plataforma digital de intercambio económico cultural mediante técnicas de Procesamiento de Lenguaje Natural y Machine Learning.

1.4.2 Objetivos Específicos

- Analizar técnicas de Machine Learning para traducción en lenguas indígenas.
- Desarrollar los mecanismos de selección y preparación de los datasets.
- Determinar la metodología adecuada para el entrenamiento de datasets.
- Entrenamiento de datasets.
- Generación de resultados.

1.5 Justificación de la investigación

En la actualidad, las Tecnologías de la Información y la Comunicación (TIC) se encuentran presentes en múltiples entornos de la vida diaria de una persona, pudiendo encontrarse en el trabajo, en el hogar, en el colegio e incluso en el propio vehículo (Pérez, 2020).

La omnipresencia de las Tecnologías de la Información y Comunicación (TIC) en nuestra cotidianidad, junto con su constante desarrollo y evolución, ha llevado a una adaptación gradual del ser humano a un entorno completamente digital.

En el mundo globalizado actual, la habilidad de comunicarnos de manera efectiva con personas de diferentes culturas es esencial. Superar las barreras del idioma se vuelve crucial, especialmente en contextos de intercambio económico y cultural.

La creación de un Asistente Virtual que utiliza el Procesamiento del Lenguaje Natural (PLN) en una plataforma orientada a promover el intercambio cultural y económico, con un enfoque particular en el uso de lenguas indígenas.

Para comprender mejor esta propuesta, es importante visualizar cómo se integraría un Asistente Virtual habilitado con PLN en una plataforma digital diseñada para facilitar el intercambio económico y cultural. Esta solución tiene el potencial de generar un cambio significativo.

Al aprovechar tanto el PLN como el Aprendizaje Automático (ML), este Asistente Virtual no solo se limitaría a proporcionar traducciones, sino que también permitiría una comunicación fluida y natural en lenguaje humano.

El Aprendizaje Automático, en este contexto, se convertiría en la base para mejorar de forma continua el desempeño del Asistente, adaptándolo a las necesidades cambiantes y enriqueciendo la experiencia de los futuros usuarios.

Dentro del ámbito de las lenguas indígenas, que tienen un valor cultural y lingüístico profundo, esta solución no solo sería una respuesta a la barrera del idioma, sino que también podría ser un impulsor clave para el intercambio económico y cultural entre comunidades diversas.

La convergencia del PLN y el ML en este Asistente Virtual no solo abre nuevas posibilidades para una comprensión más profunda entre comunidades diversas, sino que también contribuye a la preservación y el enriquecimiento de las lenguas indígenas en un mundo cada vez más interconectado.

CAPÍTULO II
MARCO TEÓRICO

2.1 Fundamentos teóricos

En este capítulo, se explican los términos más importantes para entender el proceso de formación y desarrollo de un modelo de traducción.

2.1.1 Asistente Virtual

La tecnología avanzada se ha transformado en un elemento crucial en nuestra sociedad contemporánea, emergiendo como respuesta a la demanda de hacer más sencilla la vida cotidiana de las personas.

Los asistentes virtuales son herramientas que facilitan la interacción entre las personas y las máquinas. Están basados en inteligencia artificial, lo que les permite entender los comandos de voz de forma natural (Elías & Morán, 2023).

Los asistentes virtuales pueden tener un papel relevante en cómo procesamos la información que nos ofrecen estos aparatos (Pérez, 2020).

Las barreras técnicas están disminuyendo cada vez más, y la interacción con los dispositivos puede facilitar y superar ciertos obstáculos de comportamiento social, como la timidez y la inseguridad.

Un asistente virtual es una aplicación de software creada para realizar tareas específicas o brindar servicios interactuando con los usuarios, generalmente a través de interfaces de conversación de texto o voz.

Soofastaei (2021) afirma que un asistente virtual inteligente (IVA) o asistente personal inteligente (IPA) es un agente de software que puede hacer tareas o servicios para una persona basándose en comandos o preguntas. El avance de la calidad de los algoritmos de aprendizaje de la inteligencia artificial (IA) amplia la

aplicación de los IVA en distintas áreas. Los IVA tienen cada vez más capacidades y usos.

2.1.2 Inteligencia Artificial

La Inteligencia Artificial se ha convertido en uno de los temas de discusión más destacados en los últimos años, tanto en el ámbito tecnológico como empresarial, sin olvidar el efecto que tendrá en la forma en que los humanos interactuaremos, realizaremos nuestro trabajo y estaremos interconectados.

Jofre (2023) señala que la inteligencia artificial no es una novedad en varios ámbitos humanos, ya que se ha integrado en distintas tecnologías desde hace tiempo. Lo que la hace diferente es su facilidad de uso y, sobre todo, la posibilidad de que cualquier persona con acceso a internet pueda usarla de forma libre y gratuita.

El campo de la inteligencia artificial (IA) ha tenido avances increíbles en los años recientes, que nos llevan a un nuevo panorama que muestra transformaciones radicales en la vida humana (Terrones Rodríguez, 2020).

Dichos avances están impregnando las diversas esferas que conforman la vida humana, profesional, política, social, económica y nos sitúan en una etapa de transición a un nuevo tiempo que depara asombrosos desafíos.

Moreno Padilla (2019) indica que en el pasado, la inteligencia artificial, la realidad virtual, la programación y la simulación eran consideradas como componentes de la ciencia ficción y de mundos futuribles que solo existían en las obras de Isaac Asimov, Arthur C. Clarke, Stanisław Lem y H. G. Wells.

Estos autores nos presentaron las posibilidades ilimitadas de las máquinas y cómo estas se convertirían en el futuro en una parte integral de nuestras vidas y en potentes instrumentos que han transformado el mundo de diversas maneras.

Según Sánchez et al.(2021), la definición de Inteligencia Artificial (IA) es compleja debido a la imprecisión inherente al concepto de inteligencia. En un sentido más informal, se considera que una máquina posee IA cuando puede replicar las capacidades cognitivas que normalmente asociamos con el cerebro humano, tales como la creatividad, la capacidad de aprender, la comprensión, la percepción del entorno y el uso del lenguaje.

Conforme a lo indicado por Hernández & Cruz (2022), la Inteligencia Artificial (IA) ha empezado a infiltrarse en diversas áreas y sectores. En la actualidad, encontramos aplicaciones de IA en nuestros dispositivos móviles, en las plataformas de transmisión de contenido y en la atención al cliente, lo cual se ha convertido en una actividad estratégica esencial para las organizaciones.

2.1.3 Procesamiento de Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN) es una disciplina que se enfoca en comprender la estructura y funcionamiento del lenguaje humano, la creación de nuevo lenguaje y todas las actividades relacionadas con el manejo del lenguaje(*Procesamiento del lenguaje natural (PLN) - GPT-3.*, 2021).

Entre estas tareas se incluye la creación de texto nuevo, la traducción de un idioma a otro, la formulación y respuesta de preguntas, la generación de resúmenes, los chatbots, entre otros.

El Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) tiene sus raíces en el trabajo de físicos soviéticos durante los primeros años de la Guerra

Fría. Su objetivo inicial era la traducción de documentos, pero los análisis y modelos de esa época tenían limitaciones debido a la falta de conocimiento lingüístico y al bajo rendimiento computacional de la época (Contreras, 2001).(Alvarenga Alcaraz et al., 2020)

2.1.4 Machine Learning

En el contexto de la inteligencia artificial, los modelos lingüísticos se refieren a las estructuras de redes neuronales que se utilizan para procesar y generar textos, entre otras funciones. Actualmente, existen modelos que tienen la capacidad de realizar tareas como escribir libros, programar y mantener conversaciones de manera fluida (Sevilla Salcedo et al., 2022).

No obstante, para alcanzar los grandes modelos actuales, la investigación en NLP ha atravesado numerosos avances intermedios esenciales que forman parte de las mega redes neuronales que se utilizan hoy en día.

Un área de la Inteligencia Artificial que ha cobrado relevancia en los últimos años es el aprendizaje automático (Machine Learning). En este campo, un sistema adquiere la habilidad de realizar tareas a través de ejemplos o por medio de un proceso de ensayo y error (Sánchez et al., 2021).

Por otro lado, el concepto de aprendizaje automático fue acuñado por Arthur Samuel en 1959. Este científico lo definió como la habilidad de una computadora para aprender sin necesidad de una programación explícita y constante por parte de los programadores (Sánchez et al., 2021).

Otros autores también han caracterizado el aprendizaje automático como una aplicación de la inteligencia artificial (IA) que dota a los sistemas de la habilidad de aprender y perfeccionarse de la experiencia de manera autónoma sin necesidad de

programación explícita. Se enfoca primordialmente en la creación de programas informáticos que pueden acceder y utilizar datos de manera autónoma.

2.1.5 Aprendizaje Supervisado

La característica distintiva de estos algoritmos es que el entrenamiento se lleva a cabo con una solución en mente (es decir, está orientado hacia una solución) (Aragones & Graells, 2019).

Por otro lado, los métodos de aprendizaje supervisado se distinguen por la presencia de información adicional en el conjunto de datos, la cual facilitará la adaptación del modelo a estos datos (Aragones & Graells, 2019).

Podemos visualizar esta información adicional como una etiqueta vinculada a cada una de las observaciones que tenemos a nuestra disposición. De esta forma, el algoritmo puede ajustar y adaptar su configuración utilizando la información contenida en estas etiquetas.

En otras palabras, el algoritmo recibe un conjunto de variables (una imagen, las características de un objeto), junto con la solución asociada al problema presentada como una variable objetivo (una etiqueta que describe el contenido de la imagen, una etiqueta con el grado de calidad del objeto). Entre estas técnicas, podemos encontrar una amplia gama de propuestas.

2.1.6 Aprendizaje No Supervisado

Siguiendo lo expuesto por Aragones & Graells (2019), en los métodos no supervisados, no contamos con información adicional (no existe una variable que se pueda identificar como objetivo). En escenarios donde todas las variables son igualmente importantes y el objetivo es agrupar, los métodos no supervisados resultan ser la opción más adecuada.

Este enfoque permite adquirir conocimiento de la estructura original de los datos mediante la realización de agrupaciones de observaciones similares (se asignarán al mismo grupo), mientras que las observaciones no similares se asignarán a grupos diferentes.

Este tipo de algoritmos resuelve situaciones en las que no existe una variable objetivo, todas las variables tienen la misma relevancia o peso.

Conforme a lo planteado por Aragones & Graells (2019), el propósito principal es agrupar observaciones similares y, al mismo tiempo, separar aquellas que son diferentes en grupos distintos. Las diversas variantes de los métodos no supervisados se detallan en otro módulo.

Los métodos no supervisados se distinguen generalmente por el hecho de que todas las variables o características de nuestro conjunto de datos se consideran con el mismo peso o la misma relevancia.

De acuerdo con Aragones & Graells (2019), los métodos que no cuentan con información adicional (como una variable objetivo o respuesta), a diferencia de los métodos supervisados, se conocen como métodos no supervisados.

2.1.7 Aprendizaje reforzado

Este tipo de aprendizaje no será abordado, pero ha despertado y sigue despertando gran interés en la comunidad científica debido a su habilidad para adaptarse a problemas complejos.

En esencia, se trata de instruir a un agente para que adquiera conocimientos en un entorno (que simboliza el problema) mediante la ejecución de diversas operaciones y la observación y evaluación de los resultados. De este modo, puede

descubrir las mejores estrategias para resolver el problema en cuestión (Aragones & Graells, 2019).

Según Ochoa Toro (2023) el aprendizaje por refuerzo se distingue de las otras dos ramas del aprendizaje automático, es decir, el aprendizaje supervisado y el aprendizaje no supervisado. Esto se debe a que, a diferencia del aprendizaje supervisado, no existe un valor de salida esperado. Además, a diferencia del aprendizaje no supervisado, no se intenta clasificar grupos basándose en las distancias entre los datos.

En la definición de este modelo de aprendizaje automático, se identifican los elementos que lo componen. Estos elementos son el agente y el entorno, que interactúan entre sí a través de las acciones que realiza el agente, el estado del entorno y las recompensas que el entorno otorga al agente.

2.1.8 Open AI

A partir de enero de 2023, se publicaron varios artículos de noticias que destacaban uno de los progresos más notables en la Inteligencia Artificial (IA) hasta la fecha. Se referían a una innovadora forma de chat automatizado (también conocido como chatbot) que, desde finales de 2022, generó gran entusiasmo por su habilidad para producir respuestas exactas sobre cualquier tema, utilizando un lenguaje extraordinariamente similar al humano (Pacheco, 2023).

"El año 2023 comienza con una innovación de alcances poco previstos hasta el momento, el surgimiento de una inteligencia artificial de acceso abierto al público: ChatGPT." (Jofre, 2023).

Según Jofre (2023), ChatGPT, también conocido como Generative Pre-training Transformer, es una forma de inteligencia artificial creada por Open AI, una

empresa fundada por Elon Musk y Sam Altman, y respaldada por varios inversores, incluyendo a Microsoft. Esta IA tiene la capacidad de interactuar a través de un chatbot, y sus respuestas son notables por su realismo y similitud con el lenguaje humano. Se desarrolla a partir del Procesamiento del Lenguaje Natural (PLN), una subdisciplina de la inteligencia artificial que se centra en la comprensión del lenguaje humano por parte de las máquinas.

En este caso, se introdujo un modelo robusto y potente que no requería un ajuste fino y que con unas pocas demostraciones era capaz de resolver tareas complejas con un buen rendimiento.

Desde su lanzamiento, el modelo se convirtió en un punto de referencia en el modelado del lenguaje y fue uno de los primeros en mostrar un rendimiento eficaz en nuevas tareas sin necesidad de entrenamiento específico (Sevilla Salcedo et al., 2022).

Al estar entrenado como un modelo de lenguaje para comprender y generar textos, ChatGPT tiene la capacidad de realizar un sin fin de tareas, algunas de estas son:

- "Generación de textos coherentes y bien estructurados, abarcando diferentes géneros narrativos y con diferente grado de extensión a partir de una instrucción inicial simple." (Jofre, 2023).
- "Resolución de problemas planteados y respuestas a preguntas específicas, devolviendo resultados coherentes y en su gran mayoría pertinentes." (Jofre, 2023).
- "Cálculos básicos y resolución de problemas matemáticos como ecuaciones lineales, cuadráticas y polinómicas entre otras." (Jofre, 2023).

- "Análisis de datos y generación de código informático a través de instrucciones delimitadas." (Jofre, 2023).

2.1.9 JSONL

JSONL es un formato basado en texto que utiliza la extensión de archivo .jsonl y es esencialmente el mismo que el formato JSON, excepto que los caracteres de nueva línea se utilizan para delimitar los datos JSON. También se conoce con el nombre de JSON Lines (*JSONL*, 2022).

Los archivos JSONL pueden ser importados y vinculados a través de Manifold, que también ofrece la posibilidad de exportar tablas en formato JSONL. En el formato GeoJSONL, se utiliza JSONL.

Un archivo JSONL contiene solo una tabla. Cuando se manejan archivos de gran tamaño en dispositivos con memoria RAM limitada, la lectura de un archivo JSONL permite analizarlo dinámicamente una línea a la vez. Aunque el archivo en sí puede ser de cualquier tamaño, cada línea no debe superar los 2 GB.

Un archivo JSON que se genera en formato de líneas JSON se conoce como archivo JSONL. Los datos estructurados se describen en un lenguaje simple y el principal uso de los archivos JSONL es transmitir datos estructurados que deben manejarse un registro a la vez.

JSON Lines, una variante de JSON, permite a los desarrolladores almacenar entradas de datos estructurados en una sola línea de texto, facilitando la transmisión de datos a través de protocolos como TCP o UNIX Pipes.

JSON Lines es un formato ideal para archivos de registro y ofrece una forma flexible de transferir mensajes entre procesos cooperativos, según el sitio web que respalda el formato (jsonlines.org). Se integra bien con las tuberías de shell y los

programas de procesamiento de texto que tienen una interfaz de estilo UNIX. Los archivos JSONL son similares a los archivos NDJSON en su estructura.

2.1.10 Transformer

El Transformer es un modelo de Aprendizaje Profundo que se creó en 2017 para hacer traducciones entre idiomas naturales. Gracias a las innovaciones que implemento, sobre todo la de auto-atención, se han podido crear prototipos que poseen una noción intuitiva del contexto y que entienden el sentido y las estructuras ocultas del lenguaje (Bender et al., 2023).

En el año 2020, OpenAI lanzó GPT-3, un modelo ya entrenado con un enfoque en la generación de lenguaje. Este modelo mostró resultados prometedores, generando textos de una calidad tan alta que es complicado discernir si fueron escritos por un humano o una máquina.

Se puede sostener que el código fuente es un texto generado en un lenguaje formal, por lo que podría ser creado utilizando herramientas basadas en estos modelos.

Los modelos que se fundamentan en Transformers están evidenciando, a través del impacto de sus logros, ser la tecnología idónea para el procesamiento de secuencias extensas en el Procesamiento del Lenguaje Natural (PLN)(Bender et al., 2023).

En el año 2018, Google introdujo uno de los primeros modelos fundamentados en Transformers, conocido como Bidirectional Encoder Representations from Transformers (BERT). En su versión más amplia, este modelo cuenta con 340 millones de hiperparámetros (Sevilla Salcedo et al., 2022).

Basado exclusivamente en bloques codificadores de transformers, el diseño del modelo se enfocó en el aprendizaje transferible, preentrenando un modelo para una tarea específica, para luego, mediante un ajuste fino, especializarlo para cualquier otra tarea. Gracias a esta técnica, el modelo podía adaptarse a diferentes tareas a medida que se afinaba.

2.1.11 Visual Studio Code

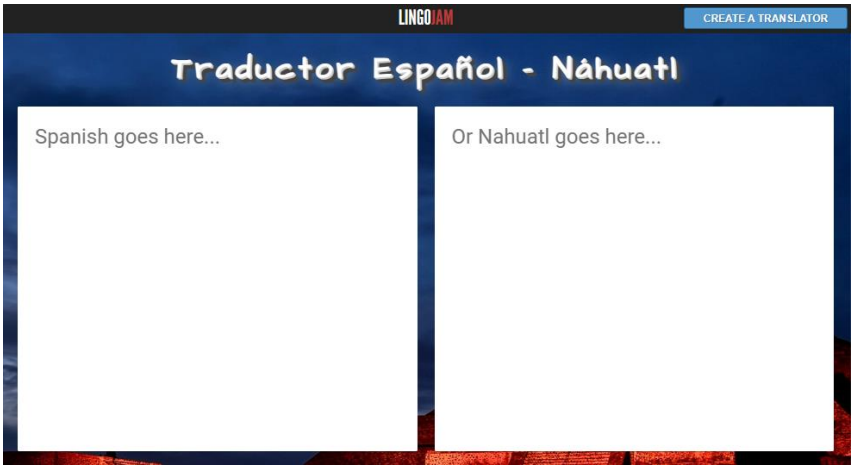
Visual Studio es un potente instrumento de desarrollo que facilita la realización de todo el proceso de desarrollo en un único lugar. Se trata de un entorno de desarrollo integrado (IDE) integral que se puede emplear para redactar, editar, depurar y compilar el código, y posteriormente desplegar la aplicación.

Además de las funciones de edición y depuración de código, Visual Studio proporciona compiladores, herramientas de autocompletado de código, gestión de código fuente, extensiones y una serie de otras características diseñadas para optimizar cada etapa del proceso de desarrollo de software (anandmeg, 2023).

2.2 Estado del arte

A continuación, se presentan algunos traductores existentes del español al náhuatl.

Ilustración 2.1 *Traductor LingoJam*

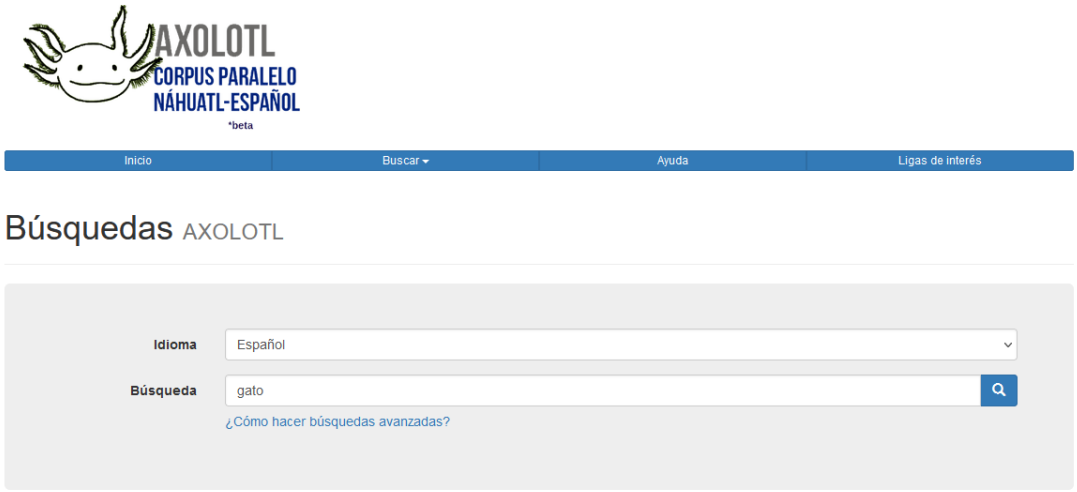


Fuente: LingoJam, 2023.

La precisión del traductor puede variar al traducir expresiones o palabras dependiendo del contexto, por lo que es necesario familiarizarse con la estructura del idioma mexicano. Si se desea traducir algo de español a náhuatl, se utilizará la variante conocida como mexicano de la Huasteca con la ortografía INALI 2018.

Si se desea traducir algo de náhuatl a español, se puede traducir de cualquier variante u ortografía que haya existido en la historia del náhuatl. El desarrollo de este traductor continúa, por lo que aún hay mucho vocabulario por añadir.

Ilustración 2.2 *Traductor Axolot*



Fuente: Axolot, 2023.

El Axolotl, un corpus paralelo español-náhuatl, es una colección digital que reúne varias fuentes con contenido paralelo en estos dos idiomas.

Se define un corpus paralelo como aquel que consta de textos en un idioma fuente junto con sus respectivas traducciones en uno o más idiomas destino. En otras palabras, cada texto tiene su correspondiente traducción.

2.2.1 CRISP-DM

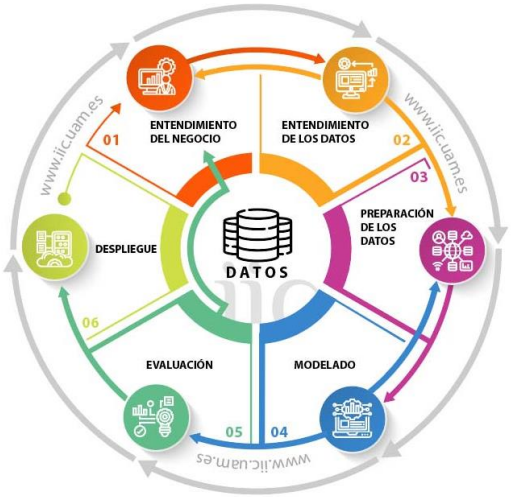
Hay varias metodologías que ayudan al análisis de datos para obtener información que se transforme en conocimiento: una de ellas es la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) que, aunque es la más usada para proyectos de minería de datos y tiene más de veinte años de existencia, no es muy reconocida en el ámbito laboral de muchas organizaciones de diferentes tipos en México (Espinosa-Zúñiga, 2020).

La metodología CRISP-DM es una de las más utilizadas en la actualidad para la implementación de proyectos de minería de datos. Fue lanzada en 1997 con el apoyo financiero del Programa de Investigación y Desarrollo en Tecnologías de Información de la Unión Europea (ESPRIT) y fue liderada inicialmente por cinco empresas:

SPSS (una empresa dedicada al software estadístico que más tarde sería comprada por IBM), Teradata (una empresa especializada en Inteligencia de Negocios), Daimler AG (una empresa automotriz con un equipo significativo de Minería de Datos), NCR (una de las empresas más grandes en informática en ese momento que más tarde produciría su propio software de Minería de Datos) y Ohra (una compañía de seguros que en esos años comenzaba a explorar el uso potencial de la minería de datos).

La primera versión de la metodología se dio a conocer en Bruselas en 1999 y se publicó ese mismo año como una guía paso a paso para la minería de datos.

Ilustración 2.3 *Fases de la Metodología CRISP-DM*



Fuente: Medium, 2023.

La metodología CRISP-DM se compone de seis etapas:

Comprensión del problema o negocio

Este es el paso más crucial, ya que una comprensión incorrecta del problema o del negocio hará que las siguientes etapas sean inútiles.

Las tareas principales de este paso son:

- Identificación del problema: Implica comprender y definir el problema, así como identificar los requisitos, suposiciones, limitaciones y ventajas del proyecto.
- Determinación de objetivos: Define los objetivos que se pretenden alcanzar al sugerir una solución basada en un modelo de minería de datos.
- Evaluación de la situación actual: Describe la situación actual antes de implementar la solución de minería de datos propuesta, con el fin de tener un punto de referencia que permita medir el éxito del proyecto.

Comprensión de datos

Las tareas clave de esta fase son:

- Recolección de datos: Implica la obtención de los datos necesarios para el proyecto, identificando las fuentes, las metodologías utilizadas para su recolección, los problemas encontrados durante su adquisición y cómo se resolvieron estos problemas.
- Descripción de datos: Se refiere a la identificación del tipo, formato, volumen y significado de cada dato.
- Exploración de datos: Consiste en la aplicación de pruebas estadísticas básicas para conocer las propiedades de los datos con el objetivo de comprenderlos de la mejor manera posible.

Preparación de datos

Esta es la fase que suele tomar más tiempo en el proyecto, y en ella se eligen y modifican los datos según los resultados de la fase previa para usarlos en la etapa de modelado.

Las actividades principales de esta fase son:

- Limpieza de datos: Se usan varias técnicas, como la normalización de datos, la discretización de campos numéricos, el tratamiento de valores ausentes, el manejo de duplicados y la imputación de datos.
- Creación de indicadores: Se generan indicadores a partir de los datos existentes que aumenten la capacidad predictiva de los datos y ayuden a encontrar patrones interesantes para el modelado.

- Transformación de datos: Se cambia el formato o la estructura de algunos datos sin modificar su significado, con la finalidad de usar una técnica específica en la etapa de modelado.

Modelado

En esta fase se construye el modelo de minería de datos. Las actividades principales de esta fase son:

- Elección de técnica de modelado: Se elige la técnica más apropiada según el problema a solucionar, los datos disponibles, las herramientas de minería de datos disponibles y el dominio de la técnica elegida.
- Elección de datos de prueba: Para algunos tipos de modelos, es preciso separar el conjunto de datos en datos de entrenamiento y de validación.
- Creación del modelo: Se obtiene el modelo óptimo mediante un proceso iterativo de modificación de sus parámetros.

Evaluación del modelo

En esta fase, se evalúa la calidad del modelo mediante el análisis de varias métricas estadísticas. Los resultados se comparan con los obtenidos anteriormente o se analizan con la ayuda de expertos en el campo del problema. Dependiendo de los resultados de esta fase, se decide avanzar a la última fase de la metodología, volver a una de las fases anteriores o incluso comenzar un nuevo proyecto desde cero.

Implementación del modelo

Esta fase utiliza el conocimiento obtenido a través del modelo mediante acciones específicas. Es crucial documentar los resultados de forma comprensible para el

usuario final y garantizar que todas las fases de la metodología estén adecuadamente documentadas para revisar el proyecto y extraer lecciones del proceso. Además, es importante supervisar las acciones para identificar oportunidades de mejora o incluso nuevos desafíos.

2.2.2 SEMMA

Gavilán (2021) explica que SEMMA es una metodología creada por el SAS Institute en 2012 para hacer Data Mining, que es anterior a CRISP-DM, otra metodología en la que también participó SAS (actualmente IBM). Aunque ambas metodologías se enfocan en el Data Mining, que es el precursor del Machine Learning, sus propuestas siguen siendo útiles en el ámbito más amplio del Machine Learning, según su opinión.

SEMMA establece un proceso que se compone de cinco fases:

Ilustración 2.4 *Fases de la Metodología SEMMA*

Fuente: Creación propia, 2023.

Muestreo: Se trata de obtener datos y, de ser necesario, de crear tablas u otras estructuras para guardarlos. Es importante tener en cuenta el tamaño de las muestras para asegurar que el conjunto de datos sea lo bastante grande para ser representativo, pero también lo bastante pequeño para permitir su procesamiento.

Exploración: Consiste en entender los datos disponibles mediante visualización, agrupamiento, etc., de forma que se puedan hallar relaciones, tendencias y

cualquier otra información que ofrezca una visión general de los datos y del fenómeno subyacente.

Modificación: Implica trabajar con los datos para prepararlos para la etapa de modelado. Esto se consigue realizando pequeñas transformaciones, seleccionando variables e incluso creando nuevas a partir de los datos brutos.

Modelado: Es posiblemente la etapa más científica y valiosa (aunque no necesariamente la que toma más tiempo), donde se elige el modelo más adecuado (el que mejor predice la variable de salida a partir de las variables de entrada). Esto se hace eligiendo entre las familias de modelos disponibles (como redes neuronales, árboles de decisión, regresión logística, etc.) y modificando el modelo en las opciones elegidas.

Evaluación: Se refiere a la evaluación del desempeño del modelo en su conjunto en términos de confiabilidad, utilidad, etc.

Gavilán (2021) explica que se trata de un proceso definido a un nivel muy alto, que no se distingue mucho de los que ya hemos visto (sobre todo del CRISP-DM) y que, en gran parte es sentido común o conocimiento básico de la forma de trabajar con datos.

2.2.3 KDD

KDD (Knowledge Discovery in Databases) fue propuesta por Fayyad en 1996 y consta de cinco fases: selección, preprocesamiento, transformación, minería de datos y evaluación e implantación. KDD es un proceso iterativo e interactivo, lo que permite a los investigadores refinar sus resultados a medida que adquieren una mejor comprensión de los datos.

2.2.4 Análisis Comparativo de Metodologías

Las metodologías KDD, SEMMA y CRISP-DM son las más destacadas en el campo de la minería de datos. A pesar de que todas ellas ofrecen un esquema organizado para el análisis y la minería de datos, existen diferencias fundamentales entre ellas que pueden afectar la decisión de cuál utilizar para un proyecto en particular.

KDD es un proceso iterativo e interactivo, lo que permite a los investigadores refinar sus resultados a medida que adquieren una mejor comprensión de los datos.

SEMMA es una metodología más simplificada y lineal en comparación con KDD, lo que puede facilitar su implementación.

CRISP-DM es flexible y se puede personalizar fácilmente para adaptarse a las necesidades específicas de un proyecto.

Para el entrenamiento del modelo de español a náhuatl, CRISP-DM es la mejor opción por varias razones:

- Flexibilidad: CRISP-DM es flexible y se puede personalizar fácilmente para adaptarse a las necesidades concretas.
- Iterativo: Al igual que KDD, CRISP-DM es un proceso iterativo, lo que permite revisar y mejorar el modelo a medida que se avanza.
- Ampliamente adoptado: CRISP-DM es ampliamente adoptado en la industria, lo que significa que hay muchos recursos y soporte disponibles.

CAPÍTULO III
DESARROLLO Y METODOLOGÍA

3.1 Diseño y metodología CRISP-DM

En el capítulo que sigue, se desglosarán las acciones ejecutadas en cada una de las cuatro etapas seleccionadas de la metodología CRISP-DM para la realización de este proyecto. Las fases elegidas son las siguientes:

Comprensión de los datos

En esta fase, se realizó un análisis exhaustivo de los datos disponibles. Se identificaron las características clave, se exploraron las relaciones entre las variables y se comprendieron las estructuras de los datos.

Preparación de los datos

Esta etapa implicó la limpieza y transformación de los datos para su posterior análisis. Los datos se depuraron de posibles errores, se normalizaron y se estructuraron de manera que pudieran ser procesados eficientemente por los algoritmos de Machine Learning.

Modelado

Aquí, se seleccionó el algoritmo de Machine Learning más adecuado para el problema en cuestión. Se entrenó el modelo utilizando los datos preparados y se ajustaron los parámetros del modelo para optimizar su rendimiento.

Evaluación

Finalmente, se evaluó el rendimiento del modelo. Se utilizaron métricas apropiadas para medir la precisión del modelo y se realizaron pruebas para asegurar que el modelo funcionaba como se esperaba.

Cada una de estas fases jugó un papel crucial en el desarrollo del proyecto y se abordarán en detalle en el próximo capítulo. Esto proporcionará una visión completa de las actividades realizadas y los resultados obtenidos en cada etapa del proyecto.

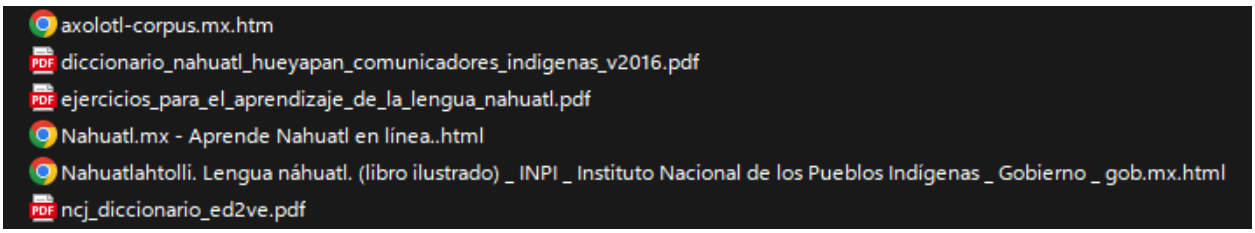
3.2 Comprensión de los Datos

En esta fase, se realizó la recolección inicial de datos. Este paso es fundamental ya que proporciona la oportunidad de tener un primer contacto con el problema, familiarizarse con los datos y evaluar su calidad. Con una comprensión sólida de los datos, se puede avanzar con confianza en las fases siguientes de la metodología.

3.2.1 Recolectar los Datos Iniciales

Para la obtención inicial de datos, se recurrió a distintas fuentes de información en español y náhuatl que se juzgaron pertinentes y de confianza para el proyecto.

Ilustración 3.1 *Fuentes de información español y náhuatl*



Fuente: Creación propia, 2023.

Un diccionario del español al náhuatl, que alberga 13 mil entradas con sus correspondientes traducciones, ejemplos y anotaciones gramaticales. Este diccionario permitió la adquisición de un vocabulario extenso y diverso de ambos

idiomas, además de facilitar el entendimiento de ciertas reglas y estructuras lingüísticas.

Un portal web del estado de Puebla, que brinda datos sobre la historia, la cultura, el turismo y la actualidad de esta región, donde una porción significativa de la población habla náhuatl. Este portal web facilito la extracción de textos de diversas temáticas y estilos, en español como en náhuatl.

La plataforma Axolotl de la UNAM, que funciona como un repositorio digital de recursos lingüísticos del náhuatl y otras lenguas indígenas de México. Esta plataforma permitió el acceso a textos académicos, literarios, periodísticos y cotidianos, en español y náhuatl, así como a corpus paralelos, anotados y alineados, que facilitan la comparación y traducción entre idiomas.

De cada una de las fuentes mencionadas, se obtuvo una muestra representativa de los datos, considerando factores como la cantidad, la calidad, la diversidad y la autenticidad de los mismos.

Para lograr esto, se emplearon diversos métodos de recopilación de datos, como la extracción de textos de sitios web, la descarga de archivos PDF, la conversión de formatos, la selección de segmentos, entre otros.

3.2.2 Descripción de los datos

Los datos recopilados para el proyecto consisten en textos en español y náhuatl, que se originan de fuentes distintas:

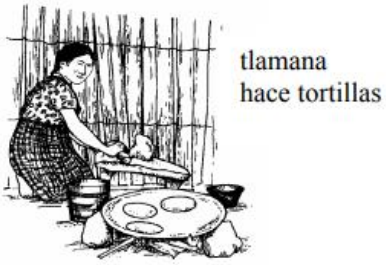
- Un diccionario.
- un sitio web.
- Una plataforma.

A continuación, se presentan algunos aspectos que se pueden describir sobre los datos:

Formato

Los datos se encuentran en formato digital, principalmente en formatos PDF, HTML. Algunos datos también contienen imágenes, que se presentan en formatos JPG, PNG.

Ilustración 3.2 *Traducción español - náhuatl*



Fuente: Diccionario Hueyapan, 2023.

Tipo

Los datos son predominantemente textuales, lo que significa que contienen información escrita en forma de palabras, frases, oraciones y párrafos. Algunos datos también son de tipo multimedia, lo que implica que contienen información visual y en forma de imágenes.

Estructura

Los datos tienen una estructura variable, dependiendo de la fuente y el tema. Algunos datos tienen una estructura simple, como las entradas del diccionario, que constan de una palabra en español, su traducción en náhuatl, un ejemplo y una nota gramatical. Otros datos tienen una estructura compleja, como los textos del sitio web y la plataforma, que pueden tener títulos, subtítulos, encabezados, listas, tablas, gráficos, enlaces, etc.

Contenido

Los datos presentan una diversidad de contenidos, cubriendo variadas temáticas y estilos. Algunos datos poseen un contenido informativo, como los textos del sitio web, que proporcionan datos sobre la historia, la cultura, el turismo y la actualidad de Puebla.

Otros datos poseen un contenido académico, como los textos de la plataforma, que exponen investigaciones, análisis y reflexiones sobre el náhuatl y otras lenguas indígenas.

Otros datos poseen un contenido literario, como los textos del diccionario, que contienen extractos de obras clásicas y contemporáneas en español y náhuatl.

Los atributos o variables que constituyen los datos son los elementos que caracterizan y distinguen cada texto. A continuación, se presentan algunos de los atributos que se pueden identificar y clasificar según su naturaleza:

Idioma

Es el atributo que señala el idioma en el que se redactó el texto, que puede ser español o náhuatl. Es un atributo categórico, es decir, que tiene un número limitado de posibles valores.

Fuente

Es el atributo que señala la fuente de la que se obtuvo el texto, que puede ser el diccionario, el sitio web o la plataforma. Es un atributo categórico, es decir, que tiene un número limitado de posibles valores.

Tema

Es el atributo que señala el tema principal del texto, que puede ser diverso, como historia, cultura, turismo, actualidad, educación, salud, economía, etc. Es un atributo categórico, es decir, que tiene un número limitado de posibles valores.

Estilo

Es el atributo que señala el estilo del texto, que puede ser formal, informal, académico, literario, periodístico, cotidiano, etc. Es un atributo categórico, es decir, que tiene un número limitado de posibles valores.

Longitud

Es el atributo que señala el número de palabras que contiene el texto. Es un atributo numérico, es decir, que tiene un valor cuantitativo y continuo.

Complejidad

Es el atributo que señala el grado de dificultad del texto, que puede depender de factores como el vocabulario, la gramática, la sintaxis, la coherencia, la cohesión, etc. Es un atributo numérico, es decir, que tiene un valor cuantitativo y continuo.

El objetivo o la variable de resultado que se busca predecir con el asistente virtual es la traducción de un texto del español al náhuatl. Es decir, se espera que el asistente virtual sea capaz de recibir un texto en uno de los dos idiomas y devolver un texto equivalente en el otro idioma, que mantenga el significado, el estilo y la calidad del texto original. La traducción es un atributo textual, es decir, que tiene un valor cualitativo y discreto.

3.2.3 Verificación de la calidad de los datos

Para la verificación de la calidad de los datos, se evaluó la calidad de los datos recopilados, utilizando diversos criterios para medir su idoneidad y fiabilidad para

el proyecto. Estos son algunos de los criterios que se utilizaron para medir la calidad de los datos:

Completitud

Es el grado en el que los datos poseen todos los valores esperados o necesarios para el análisis. Un dato se considera incompleto si le falta algún valor o si tiene un valor nulo o vacío. Por ejemplo, un texto que no tiene traducción en el otro idioma, o que tiene una traducción incompleta o truncada, se considera un dato incompleto.

Consistencia

Es el grado en el que los datos son coherentes y compatibles entre sí y con las reglas o estándares establecidos. Un dato se considera inconsistente si tiene un valor que contradice o viola algún principio o norma. Por ejemplo, un texto que tiene una traducción que no respeta la gramática o la ortografía del otro idioma, o que tiene una traducción que cambia el sentido o el tono del texto original, se considera un dato inconsistente.

Precisión

Es el nivel en el que los datos reflejan con exactitud la realidad o la veracidad de los hechos o fenómenos que representan. Un dato se considera impreciso si contiene un valor que incluye un error o una desviación. Por ejemplo, un texto que tiene una traducción que incluye una palabra o expresión incorrecta o inadecuada, o que tiene una traducción que pierde o distorsiona la información o el mensaje del texto original, se considera un dato impreciso.

Actualidad

Es el nivel en el que los datos están actualizados o son pertinentes para el análisis. Un dato se considera desactualizado si tiene un valor que no coincide con el momento o el contexto en el que se realiza el análisis. Por ejemplo, un texto que tiene una traducción que utiliza una palabra o expresión obsoleta o anticuada, o que tiene una traducción que no se ajusta a la situación o al propósito del análisis, se considera un dato desactualizado.

Relevancia

Es el nivel en el que los datos son relevantes o significativos para el análisis. Un dato se considera irrelevante si tiene un valor que no contribuye o que resta valor al análisis. Por ejemplo, un texto que tiene una traducción que utiliza una palabra o expresión demasiado general o específica, o que tiene una traducción que no se relaciona o se ajusta al tema o al estilo del análisis, se considera un dato irrelevante.

3.3 Preparación de los Datos

3.3.1 Selección de datos

Para la selección de datos, se decidió qué datos se utilizarían en el análisis, basándose en criterios de importancia, calidad, relevancia y restricciones.

- Se incluyeron los textos que tenían una traducción completa y precisa en el otro idioma, ya que eran los más relevantes y útiles para entrenar y evaluar el asistente virtual de traducción.
- Se excluyeron los textos que tenían una traducción incompleta, incorrecta o irrelevante, ya que podrían afectar negativamente el rendimiento y la calidad del asistente virtual de traducción.

- Se incluyeron los textos que tenían un tema y un estilo acorde con el objetivo y el alcance del proyecto, ya que eran los más pertinentes y significativos para el análisis y el modelado.
- Se excluyeron los textos que tenían un tema o un estilo que no se ajustaban al objetivo o al alcance del proyecto, ya que podrían ser irrelevantes o inapropiados para el análisis y el modelado.

3.3.2 Limpieza de datos

Para la limpieza de la información, se mejoró su calidad al nivel requerido por las técnicas de análisis seleccionadas. Se detectaron y solucionaron los problemas que afectaban a la información, como la ausencia, el error, la duplicación, el ruido o el sesgo de los mismos.

Estos son algunos de los métodos y técnicas que se utilizaron para limpiar los datos:

- Se validaron los datos utilizando reglas o criterios de validación, como la longitud, el formato, el tipo, el rango, etc. de los valores. Se eliminaron o corrigieron los valores que no cumplían con las reglas o criterios de validación.
- Se eliminaron los datos duplicados, es decir, los textos que tenían la misma traducción en el otro idioma, o que eran idénticos o muy similares entre sí.
- Se corrigieron los datos erróneos, es decir, los textos que tenían una traducción que contenía algún tipo de error ortográfico, gramatical, sintáctico, semántico, etc. Se utilizaron herramientas o algoritmos de detección y corrección de errores, como el corrector ortográfico, el corrector gramatical, el traductor, etc.

- Se sustituyeron o imputaron los datos faltantes, es decir, los textos que no tenían traducción en el otro idioma, o que tenían una traducción incompleta o truncada.

- Se eliminó o redujo el ruido y el sesgo de los datos, es decir, los textos que tenían una traducción que contenía algún tipo de ruido o distorsión, como el ruido de fondo, las interferencias, las ambigüedades, las contradicciones, las opiniones, las emociones, etc.

3.3.3 Integración de datos

Para la integración de datos, se combinaron los datos de diferentes fuentes o formatos en un solo conjunto de datos, se unificaron y homogeneizaron los datos para evitar redundancias o inconsistencias.

Estos son algunos de los métodos y técnicas que se utilizaron para integrar los datos:

- Se fusionaron los datos de diferentes fuentes, como el diccionario, el sitio web y la plataforma.

- Se unieron los datos de diferentes formatos, como el PDF, el HTML y el TXT.

- Se concatenaron los datos de diferentes idiomas, como el español y el náhuatl.

- Se alinearon los datos de diferentes textos, como el texto original y el texto traducido.

Se llevo a cabo la adhesión de los datos obtenidos de múltiples fuentes de información en un único archivo de Excel.

Ilustración 3.3 *Integración de datos*

Question	Answer
¿Cómo te llamas?	¿kenin timo toga?.
Este pueblo ¿Cómo se llama?	Inin altepetl ¿kenin itoga?.
Burro	Axno.
Pollo	Pio.
Gato	Mistle.
Cerdo	Pitsotl.
Armadillo	Ayotochi.
Tejón	Texon.
Coyote	Koyotl.
Víbora	Koatl.
Piojo	Mejtoli.
Buey	Kuague.
Hormiga	Tsigatl.
Cien pies	Petlasolkoatl.
Pez	Michi.

Fuente: Creación propia, 2023.

3.3.4 Formateo de datos

La información contenida en el archivo Excel puede ser convertida en una matriz de objetos en formato JSON, donde cada objeto representa una fila de la tabla. Este proceso permite que los datos sean extraídos en un formato coherente y fácil de leer.

El formato JSON es ampliamente aceptado y utilizado por diversas plataformas y lenguajes de programación, lo que facilita la integración de los datos con otros sistemas o aplicaciones.

Para los modelos de aprendizaje automático, como el modelo Davinci de Open AI, es necesario que los datos de entrada estén en un formato específico. En este contexto, el formato JSONL (JSON Lines) es útil porque cada línea del archivo es

un objeto JSON válido, lo que permite una lectura y procesamiento eficiente de los datos, especialmente cuando se trata de grandes conjuntos de datos.

Estos son algunos de los métodos y técnicas que se utilizaron para formatear los datos:

Ilustración 3.4 *TransformData*

```
async function TransformData(){
// Leer el archivo Excel ubicado en la ruta especificada
var workbook = xlsx.readFile("src/shared/traduccion.xlsx");

// Obtener los nombres de las hojas del libro de trabajo Excel
var shet_name_list = workbook.SheetNames;

 // Convertir los datos de la primera hoja del libro a formato JSON
var xlData = xlsx.utils.sheet_to_json(workbook.Sheets[shet_name_list[0]]);

// Iterar sobre cada elemento (fila) en los datos convertidos a JSON
for (const item of xlData){

    // Crear un objeto JSON con una estructura específica usando los datos extraídos del archivo Excel
    var object = `{"prompt": "${item.Question} ->", "completion": "${item.Answer} END"}`;

    // Agregar el objeto JSON creado al archivo 'traduccion.jsonl' ubicado en la ruta especificada
    await fs.appendFileSync("src/shared/traduccion.jsonl", object, "utf8", function(){})

    // Agregar una nueva línea al final del objeto JSON para separar cada entrada
    await fs.appendFileSync("src/shared/traduccion.jsonl", "\r\n", "utf8", function(){})
}

}
```

Fuente: Creación propia, 2023.

Este código representa una función asíncrona en JavaScript, denominada `TransformData()`, cuyo propósito es convertir datos de un archivo Excel a formato JSONL y posteriormente almacenarlos en un archivo `.jsonl`.

Ilustración 3.5 *Formato Jsonl*

```
{"prompt": "Arqueólogo ->", "completion": "Tlalnepantlalistli END"}
{"prompt": "Artesano ->", "completion": "Tlamachtinime END"}
{"prompt": "Diseñador ->", "completion": "Tlapanquetzaliztli END"}
{"prompt": "Piernas ->", "completion": "Kostle. END"}
{"prompt": "Un joven barre su calle ->", "completion": "Se piltontle tlachpana iniojtle. END"}
{"prompt": "El señor lejos camina ->", "completion": "In tlagatsintle uejka nejnemi. END"}
{"prompt": "El pájaro canta bonito ->", "completion": "In tototl mu kuigatia kuagualtsin. END"}
{"prompt": "Un caballo come zacate ->", "completion": "Se kaguayo tlagua sagatl. END"}
{"prompt": "Los muchachos aprisa van a la escuela ->", "completion": "In piltonkogonej mu tlaloa yue tlamachtiloyan. EN
{"prompt": "Las niñas estudian diariamente ->", "completion": "Soakogonej momachtia mojmostlan. END"}
{"prompt": "Mi mamá cuenta el dinero ->", "completion": "No male kipoa in melio. END"}
{"prompt": "Mi papá trabaja en el cerro ->", "completion": "No pale teguiti pan tepetl. END"}
{"prompt": "Mis hermanos siembran maíz ->", "completion": "No ikniuan kitoga tlaolle. END"}
{"prompt": "¿Quién vendrá a nuestro pueblo? ->", "completion": "¿Aquin gualas pan to altepetl?. END"}
{"prompt": "Algunos vendrán y otros se irán ->", "completion": "Siguin gualaske an siguin yaske. END"}
{"prompt": "¿Por qué te enojas? ->", "completion": "¿Tlega ti gualani?. END"}
{"prompt": "¿Tienes hambre? ->", "completion": "¿Ti apismigui?. END"}
```


Fuente: Creación propia, 2023.

Al transformar los datos a formato JSONL, se puede automatizar el proceso de preparación de datos para el entrenamiento del modelo. Esto no solo ahorra tiempo, sino que también garantiza la reproducibilidad de los experimentos de modelado.

3.4 Modelado

Con el objetivo de crear un asistente virtual de traducción del español al náhuatl, una tarea de generación de lenguaje natural, se optó por usar el aprendizaje supervisado.

El aprendizaje supervisado es una metodología de la inteligencia artificial y el aprendizaje automático en la que un modelo se capacita utilizando un conjunto de datos con etiquetas. En este escenario, los datos etiquetados consistirían en pares de oraciones en español y náhuatl.

3.4.1 Escoger la Técnica de Modelado

Primero, es crucial entender el concepto de un "token". En términos más detallados, un token se puede considerar como una unidad de texto que consta de alrededor de cuatro caracteres o 0.75 palabras en textos en inglés.

Consideremos el siguiente texto como ejemplo: "Los bots en el ámbito de 'call center'".

Cada palabra o signo de puntuación se cuenta como un token. Por ejemplo, "Los" es un token, "bots" es otro token ya que tiene cuatro caracteres. Las palabras "en" y "el" son dos tokens adicionales. La palabra "ámbito", que tiene seis caracteres, se puede considerar como dos tokens. La palabra "de" es otro token. Es importante notar que las comillas también se cuentan como un token, ya que forman parte del contexto de la frase completa.

Para agregar más detalle, un token no es necesariamente una palabra completa. Puede ser una parte de una palabra, un signo de puntuación, o incluso un espacio en blanco, dependiendo del contexto y de cómo se esté analizando el texto. En el procesamiento del lenguaje natural, los tokens son las unidades básicas de análisis. Al dividir el texto en tokens, podemos analizar y entender mejor la estructura y el significado del texto.

OpenAI proporciona cuatro modelos de entrenamiento distintos: Ada, Babbage, Curie y Davinci, cada uno diseñado para satisfacer diferentes necesidades. Para este proyecto, se seleccionó el modelo Davinci debido a su superior precisión en las respuestas. Aunque los otros modelos también son eficaces, pueden existir ligeras diferencias en los resultados que generan.

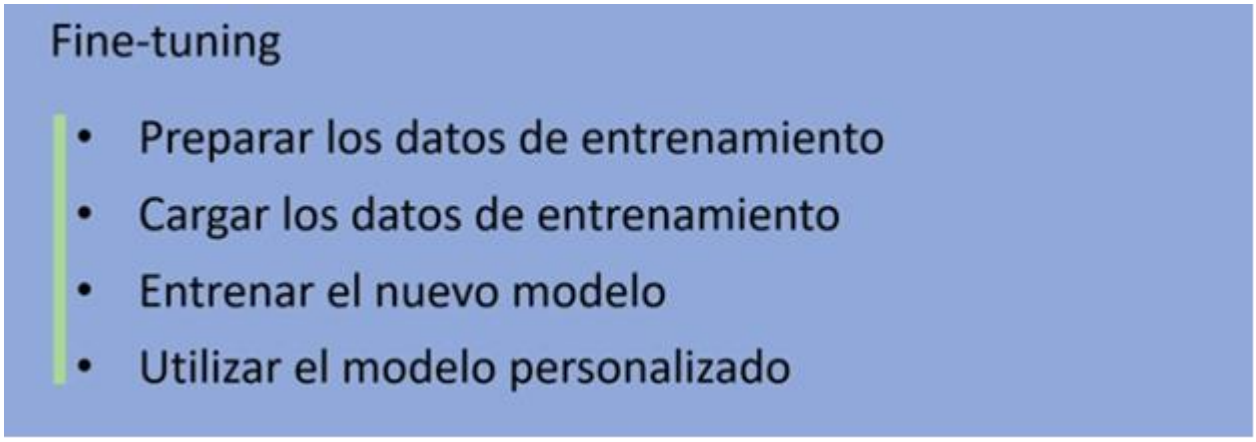
El "fine-tuning" es un procedimiento empleado en el aprendizaje profundo. Esta técnica mejora y personaliza los modelos de aprendizaje profundo para tareas concretas, utilizando un modelo preentrenado y ajustándolo con datos adicionales

pertinentes a la tarea en cuestión. En este contexto, se aplicará el "fine-tuning" para desarrollar y entrenar un modelo de traducción.

Es relevante destacar que el "fine-tuning" es una técnica efectiva que permite aprovechar el conocimiento adquirido por el modelo preentrenado, lo que reduce el tiempo y los recursos necesarios para entrenar el modelo desde cero. En este proyecto, el "fine-tuning" permitirá adaptar el modelo Davinci de OpenAI para realizar traducciones precisas del español al náhuatl.

El "fine-tuning" consta de cuatro etapas fundamentales:

Ilustración 3.6 *Etapas de Fine-tuning*



Fuente: Creación propia, 2023.

Preparación de los datos de entrenamiento

Este paso implica la recopilación y el procesamiento del conjunto de datos que se utilizará para el entrenamiento. En el contexto del proyecto en cuestión, será necesario compilar un conjunto de palabras y frases en español, junto con sus correspondientes traducciones al náhuatl.

Carga de los datos de entrenamiento

En esta fase, se importarán los datos procesados al entorno de entrenamiento. Para el proyecto en cuestión, esto implicaría cargar las palabras y frases en español y náhuatl en la plataforma o herramienta que se esté utilizando para el entrenamiento.

Entrenamiento del nuevo modelo

En este punto, se lleva a cabo la afinación del modelo preentrenado con los datos específicos. Se aplicarán algoritmos de aprendizaje automático para que el modelo aprenda las correlaciones entre el español y el náhuatl, basándose en el conjunto de datos proporcionado.

Uso del modelo personalizado

Una vez que el modelo ha sido entrenado, puede ser implementado para realizar traducciones del español al náhuatl. Su rendimiento será evaluado y se realizarán ajustes según sea necesario. En este proyecto, la afinación permitirá adaptar el modelo Davinci de OpenAI para realizar traducciones precisas del español al náhuatl.

3.4.2 Generar el Plan de Prueba

Para el diseño del test, se definió el método o el criterio para evaluar la calidad y la validez del modelo. Dado que el conjunto de datos era grande y variado, se optó por usar un método de división del conjunto de datos en entrenamiento y prueba. Este método consiste en separar los datos en dos subconjuntos: uno para entrenar el modelo y otro para probarlo.

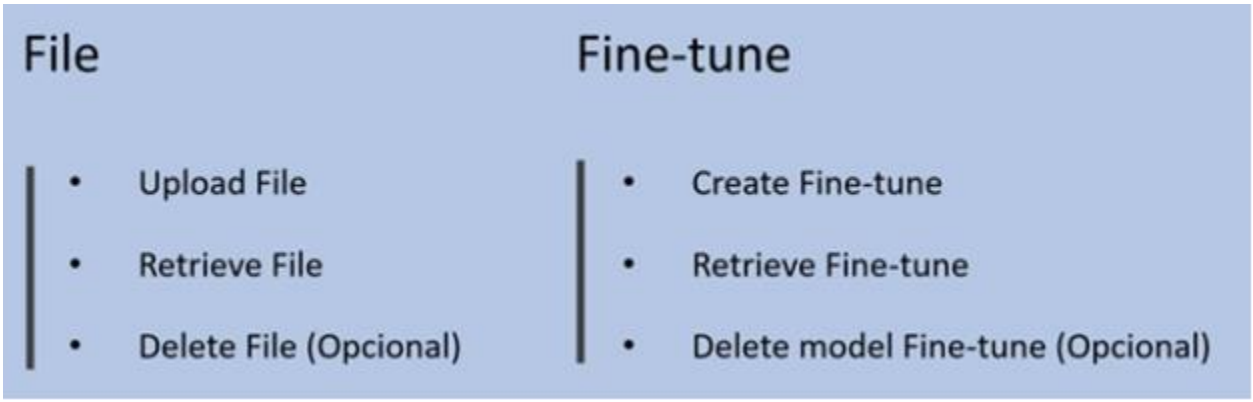
Dentro de este método, se escogió usar una proporción de 80/20, es decir, que el 80% de los datos se usaron para el entrenamiento y el 20% para la prueba. Para

diseñar el test, se tuvo en cuenta la cantidad y la representatividad de los datos, asegurándose de que los subconjuntos fueran balanceados y homogéneos.

3.4.3 Construir el Modelo

Para la construcción del modelo, se llevó a cabo lo siguiente:

Ilustración 3.7 *Proceso de entrenamiento*



Fuente: Creación propia, 2023.

Upload File

Facilita la carga de un archivo al sistema.

Ilustración 3.8 *UploadFile*

```
async function UploadFile(){
    // Utilizamos 'await' para esperar a que se complete la creación del archivo
    // usando la API de OpenAI. El método createFile se utiliza junto con un stream
    // de lectura del sistema de archivos para cargar el archivo 'traduccion.json'.

    const response = await openai.createFile(fs.createReadStream("src/shared/traduccion.jsonl"), "fine-tune");

    // La respuesta del proceso anterior se devuelve como resultado de la función.
    return response;

}
```

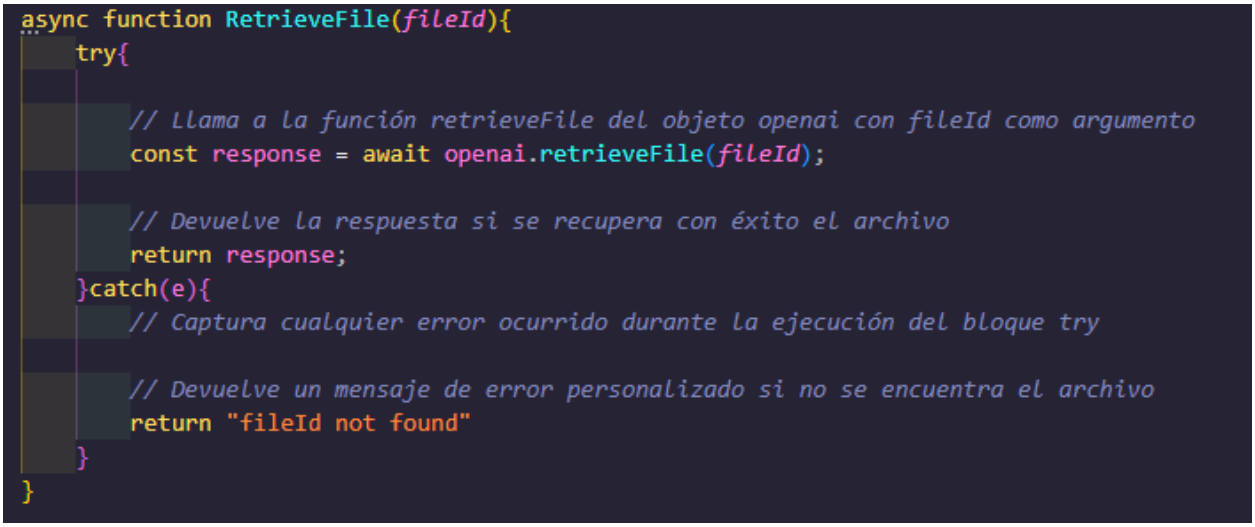
Fuente: Creación propia, 2023.

Este código representa una función asíncrona en JavaScript, denominada UploadFile, la función emplea la API de OpenAI para subir un archivo.

Retrieve File

Brinda la posibilidad de obtener un archivo que se ha subido previamente.

Ilustración 3.9 *RetrieveFile*



```
async function RetrieveFile(fileId){
    try{

        // Llama a la función retrieveFile del objeto openai con fileId como argumento
        const response = await openai.retrieveFile(fileId);

        // Devuelve la respuesta si se recupera con éxito el archivo
        return response;
    }catch(e){
        // Captura cualquier error ocurrido durante la ejecución del bloque try

        // Devuelve un mensaje de error personalizado si no se encuentra el archivo
        return "fileId not found"
    }
}
```

Fuente: Creación propia, 2023.

Esta función tiene como objetivo obtener un archivo a través de su fileId, utilizando para ello la API de OpenAI. En caso de que la función logre recuperar el archivo de manera exitosa, retorna la respuesta obtenida. Sin embargo, si se presenta algún error durante la recuperación del archivo, la función captura este error y retorna el mensaje "fileId not found".

Create Fine-tune

Permite la creación de un ajuste fino o personalización en el modelo o sistema.

Ilustración 3.10 *CreateFineTune*

```
async function CreateFineTune(fileId){
    try{
        // Intentar crear un modelo afinado con el archivo de entrenamiento proporcionado
        const response = await openai.createFineTune({
            training_file: fileId, // Archivo de entrenamiento
            model: "davinci", // Modelo base a afinar
            suffix: "question-answer-01" // Sufijo para identificar el modelo afinado
        });
        return response; // Retornar la respuesta si tiene éxito
    }catch(e){
        // En caso de error, retornar el estado y los datos del error
        return {status: 400, data: e}
    }
}
```

Fuente: Creación propia, 2023

La función `CreateFineTune` en JavaScript acepta un `fileId` como entrada, que es el identificador de un archivo de entrenamiento usado para afinar un modelo.

Esta función utiliza la API de OpenAI para afinar un modelo llamado "davinci". El sufijo "question-answer-01" se usa para distinguir el modelo afinado.

Si el proceso de afinación del modelo se realiza con éxito, la función retorna la respuesta obtenida de la API de OpenAI. En caso de que surja un error durante el proceso de afinación, la función maneja este error y retorna un objeto con un estado de 400 y los detalles del error.

Retrieve Fine-tune

Facilita la recuperación de los detalles o configuraciones de un ajuste fino que se ha creado previamente.

Ilustración 3.11 *RetriveFineTune*

```
async function RetrieveFineTune(fineTuneId){
    try{

        // Intentar obtener la respuesta de la API usando await para esperar que se complete
        const response = await openai.retrieveFineTune(fineTuneId);

        // Devolver la respuesta si es exitosa
        return response;
    }catch(e){
         // Capturar cualquier error ocurrido durante el proceso

        // Devolver un objeto con estado 400 y detalles del error
        return {status: 400, data: e}
    }
}
```

Fuente: Creación propia, 2023

La función `RetrieveFineTune` en JavaScript acepta un `fineTuneId` como entrada, que es el identificador de un modelo afinado. Este identificador se usa para obtener información de afinación fina.

Dentro de la función, se hace uso de la API de OpenAI para obtener la información de afinación fina. Si el proceso de recuperación es exitoso, la función retorna la respuesta obtenida de la API de OpenAI. En caso de que surja un error durante el proceso de recuperación, la función maneja este error y retorna un objeto con un estado de 400 y los detalles del error.

3.5 Alcance y enfoque de la investigación

La investigación que se presenta tiene como objetivo principal el desarrollo de un modelo de traducción basado en Machine Learning, que permita la traducción de textos del español al náhuatl. Los datos que se emplean para el entrenamiento y

evaluación del modelo se obtienen de una variedad de fuentes, incluyendo diccionarios del municipio de Hueyapan, sitios web culturales del estado de Puebla y repositorios de Axolot.

La metodología de la investigación se basa en un enfoque cuantitativo, descriptivo y correlacional. El enfoque cuantitativo se aplica en la recolección y análisis de los datos numéricos que se utilizan para el entrenamiento del modelo de Machine Learning. Por otro lado, el enfoque descriptivo se emplea para comprender las características de los datos y cómo estas pueden influir en el rendimiento del modelo. Finalmente, el enfoque correlacional se utiliza para entender la relación entre los datos de entrada (español) y los datos de salida (náhuatl).

Es relevante destacar que, más allá del desarrollo de un modelo de traducción eficiente, esta investigación también tiene como propósito contribuir al rescate y preservación de la lengua náhuatl mediante el uso de la tecnología.

3.6 Hipótesis

El Asistente Virtual utilizando PNL proporcionará una traducción más precisa y coherente en lenguas indígenas en comparación con las herramientas de traducción en línea existentes.

La hipótesis planteada al inicio de esta investigación fue parcialmente confirmada. El Asistente Virtual desarrollado, utilizando técnicas de Procesamiento de Lenguaje Natural (PNL), demostró tener la capacidad de proporcionar traducciones en lenguas indígenas.

Sin embargo, debido a varias limitaciones encontradas durante la realización del proyecto, no se pudo confirmar de manera concluyente que las traducciones

proporcionadas sean más precisas y coherentes en comparación con las herramientas de traducción en línea existentes.

Las limitaciones encontradas incluyen la complejidad de la lengua, recursos limitados, la calidad y cantidad de datos disponibles, el tiempo, y los desafíos de adaptabilidad y generalización. Estas limitaciones impactaron en la capacidad de entrenar y optimizar el Asistente Virtual para lograr una mayor precisión y coherencia en las traducciones.

A pesar de estas limitaciones, los resultados obtenidos en esta investigación proporcionan una base sólida para futuras investigaciones en el campo de la traducción de lenguas indígenas utilizando técnicas de PNL. Se espera que con el avance de la tecnología y la disponibilidad de más recursos, se pueda superar estas limitaciones y confirmar completamente la hipótesis en el futuro.

3.7 Diseño y metodología de la investigación

En este proyecto, la investigación cuantitativa se realiza a través de la recolección y análisis de datos numéricos. Estos datos provienen de los textos en español y náhuatl que se extraen de diccionarios del municipio de Hueyapan, páginas de cultura del estado de Puebla y repositorios de Axolot. Los textos se transforman en representaciones numéricas, como vectores de palabras, utilizando técnicas de procesamiento de lenguaje natural. Estos datos numéricos se utilizan para entrenar y evaluar el modelo de Machine Learning.

La investigación descriptiva en este proyecto implica el estudio de las características de los datos de entrenamiento. Esto incluye la longitud de las frases, la frecuencia de las palabras, la complejidad de las frases, entre otros. El objetivo es entender cómo estas características pueden influir en el rendimiento

del modelo de traducción. Por ejemplo, se podría analizar si las frases más largas o más complejas son más difíciles de traducir para el modelo.

La investigación correlacional en este proyecto busca entender la relación entre los datos de entrada (español) y los datos de salida (náhuatl). Por ejemplo, se podría investigar si la longitud de las frases en español se correlaciona con la longitud de las frases en náhuatl, o si la complejidad de las frases en español se correlaciona con la complejidad de las frases en náhuatl. Este análisis puede ayudar a entender qué aspectos de los datos de entrada son más relevantes para predecir los datos de salida.

3.8 Selección de la muestra

Para llevar a cabo este proyecto de investigación, es necesario obtener elementos específicos de los conjuntos de datos de texto bilingüe disponibles en repositorios seleccionados, tanto públicos como privados. Es crucial verificar si el volumen de datos es suficiente y accesible para entrenar nuestro modelo de Machine Learning.

En este caso de estudio, limitamos la información seleccionando únicamente los conjuntos de datos que contienen textos en los idiomas de interés para nuestro proyecto, como el español y el náhuatl. De esta manera, podemos realizar un análisis de información más enfocado y obtener resultados confiables de los elementos seleccionados.

Nuestra población de estudio se compone de los textos bilingües relevantes para el dominio de aplicación de nuestro modelo de traducción, como textos generales, médicos y legales, disponibles en los idiomas de interés.

Dependiendo del tiempo y la naturaleza de los elementos poblacionales, utilizamos técnicas como el muestreo aleatorio simple para determinar la población a evaluar.

Por ejemplo, podríamos seleccionar aleatoriamente un subconjunto específico de textos para formar nuestro conjunto de entrenamiento y otro subconjunto para formar nuestro conjunto de prueba.

3.9 Recolección de datos

En este proyecto, la recolección de datos implica obtener textos en español y náhuatl de diversas fuentes, incluyendo diccionarios del municipio de Hueyapan, páginas de cultura del estado de Puebla y repositorios de Axolot.

3.9.1 Selección del instrumento

El instrumento principal para la recolección de datos fue un documento de Excel. Los datos, que consisten en palabras o frases en español y su respectiva traducción al náhuatl, se obtuvieron manualmente de diversas fuentes, incluyendo diccionarios del municipio de Hueyapan, páginas de cultura del estado de Puebla y repositorios de Axolot. Estos datos se copiaron y pegaron en el documento de Excel.

3.9.2 Aplicación del instrumento

El instrumento que se utilizó para la recolección de datos fue un documento de Excel que se elaboró para este propósito. En este documento, se crearon dos columnas: una para el texto en español y otra para el texto en náhuatl.

Luego, se buscaron en diversas fuentes, como diccionarios del municipio de Hueyapan, páginas de cultura del estado de Puebla y repositorios de Axolot, los textos que se consideraron relevantes para la investigación. Se copiaron y pegaron los textos en el documento de Excel, asegurándose de que coincidieran el español y el náhuatl.

De esta manera, se obtuvo un corpus paralelo de textos en ambos idiomas que serviría para entrenar el modelo de traducción. La aplicación del instrumento implicó la revisión manual de las fuentes de datos y la copia y pegado de los datos relevantes en el documento de Excel. Este proceso requirió una cuidadosa revisión para asegurar la precisión de los datos recolectados.

3.9.3 Preparación de datos

Una vez recolectados los datos, deben ser preparados para el análisis. Esto puede implicar la limpieza de los datos (por ejemplo, eliminando caracteres no deseados), la transformación de los textos en representaciones numéricas (por ejemplo, utilizando técnicas de word embedding), y la división de los datos en un conjunto de entrenamiento y un conjunto de prueba.

3.10 Análisis de datos

El análisis de datos en este proyecto implica el entrenamiento y la evaluación del modelo de Machine Learning. Esto puede implicar la experimentación con diferentes algoritmos de Machine Learning, la optimización de los parámetros del modelo, y la evaluación del rendimiento del modelo utilizando métricas apropiadas (por ejemplo, la precisión de la traducción).

CAPÍTULO IV
RESULTADOS

4.1 Objetivos y metas alcanzadas

A continuación, se presentan los resultados obtenidos del entrenamiento de un modelo de traducción.

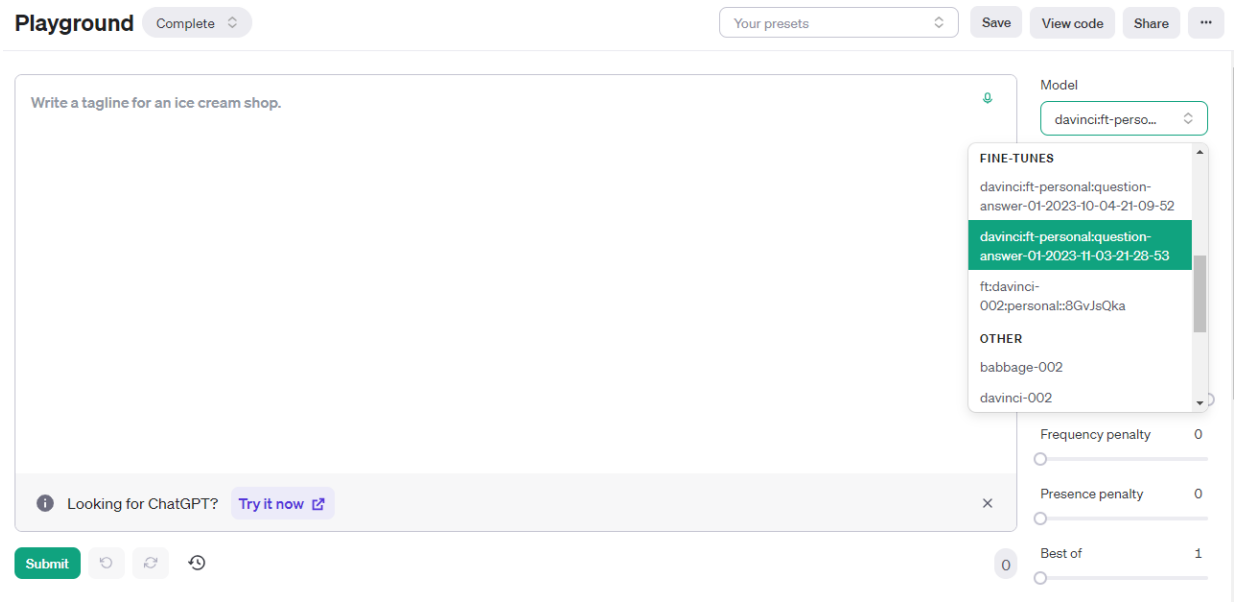
El primer paso fundamental es iniciar sesión en la cuenta de Open IA, que se empleó en el desarrollo de una sección del proyecto. Es crucial verificar que se están utilizando las credenciales adecuadas para el inicio de sesión.

Ilustración 4.1 *Inicio de sesión OpenAI*

Fuente: OpenAI, 2023

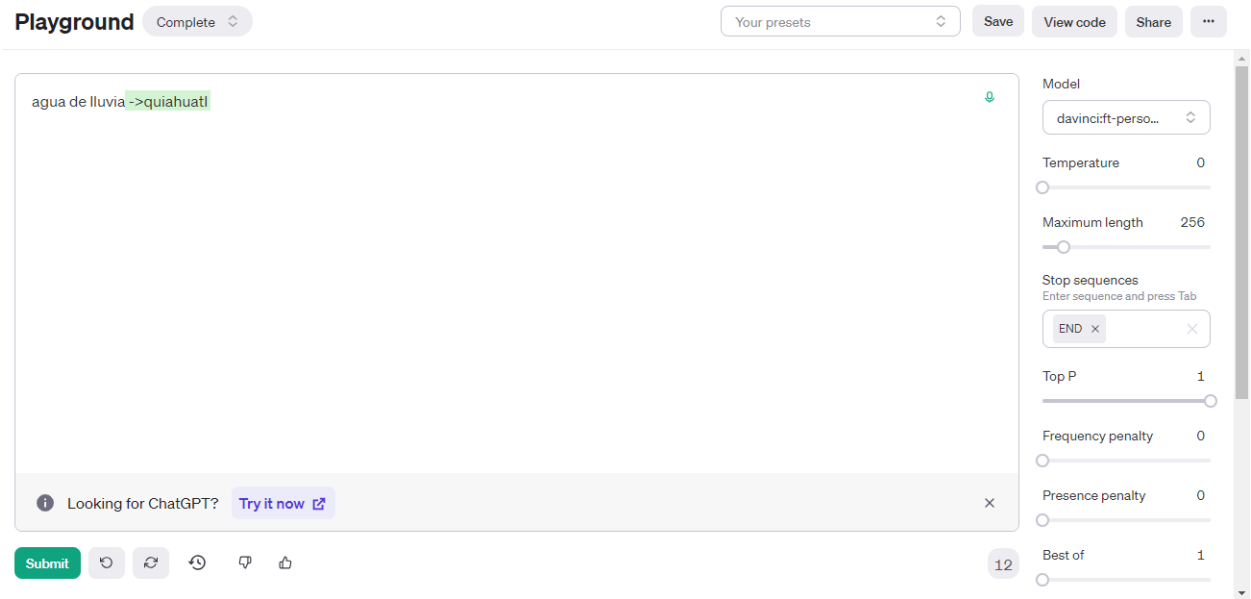
Para hacer uso y verificar el funcionamiento del modelo de traducción, es necesario ingresar a la sección "Playground" (Patio de juegos). Una vez allí se tiene que seleccionar el modelo que se ha entrenado previamente.

Ilustración 4.2 *Playground*

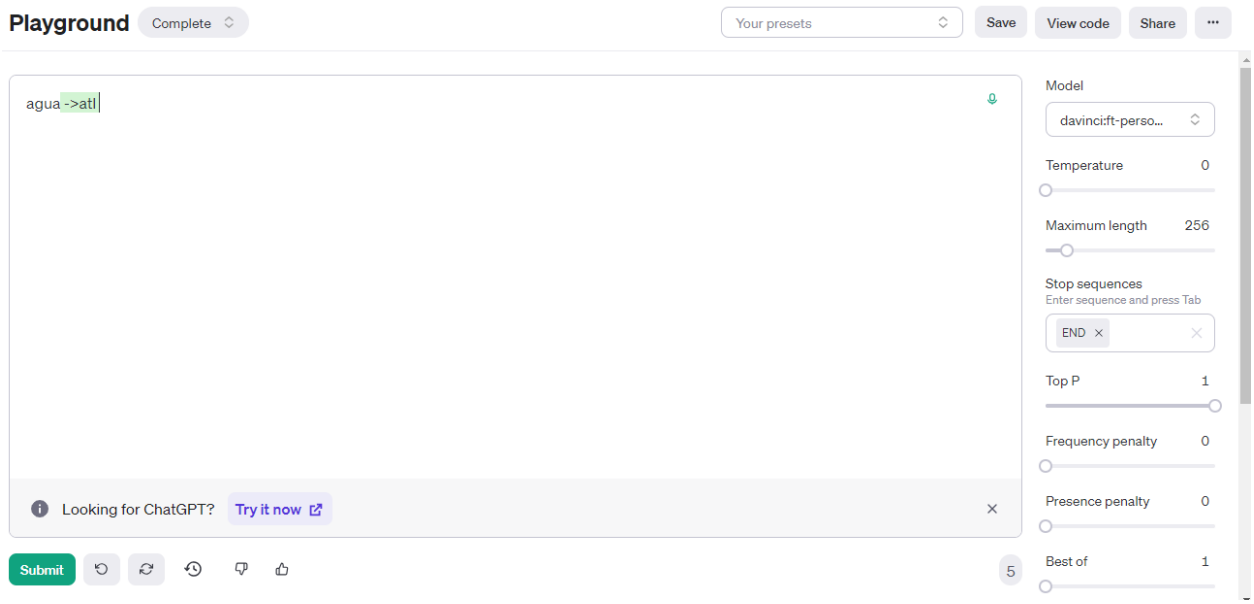


Fuente: OpenAI, 2023

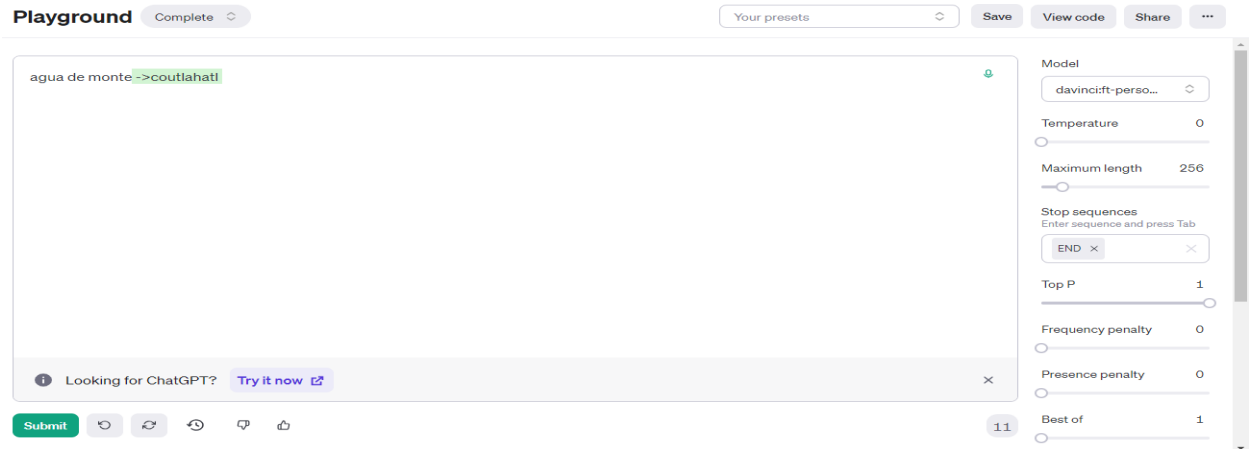
La siguiente imagen presenta un ejemplo de traducción: se ingresa "agua de lluvia" y se recibe "quiahualt", que corresponde a su traducción en Náhuatl.



La siguiente imagen presenta un ejemplo de traducción: se ingresa "agua" y se recibe "alt", que corresponde a su traducción en Náhuatl.



La siguiente imagen presenta un ejemplo de traducción: se ingresa "agua de monte" y se recibe "coutlahatl", que corresponde a su traducción en Náhuatl.



CAPÍTULO V
CONCLUSIONES

5.1 Conclusiones del proyecto

Este proyecto consistió en el desarrollo de un asistente virtual para la traducción de lenguas indígenas, empleando técnicas de Procesamiento de Lenguaje Natural y Machine Learning. El asistente virtual puede traducir textos del español al náhuatl, una de las lenguas indígenas con más hablantes en México. El asistente virtual se va integrar como un servicio de traducción para la plataforma digital Cultura Viva de intercambio económico cultural, que busca fomentar y conservar las lenguas y culturas indígenas.

Los resultados obtenidos en este proyecto demuestran que el asistente virtual tiene un rendimiento adecuado en la traducción. Asimismo, los resultados señalan que el asistente virtual tiene un impacto positivo en el fortalecimiento de las lenguas indígenas, al facilitar la comunicación y el acceso a la información en estas lenguas.

No obstante, el proyecto también presenta algunas limitaciones y desafíos, tales como la escasez y la calidad de los datos disponibles para el entrenamiento del modelo de traducción, la complejidad y la diversidad de las lenguas indígenas, y la falta de evaluación por parte de hablantes nativos de estas lenguas. Estas limitaciones afectan a la validez y la generalización de los resultados, y sugieren la necesidad de realizar más investigaciones y mejoras en el futuro.

5.2 Recomendaciones

Basándose en los resultados y las conclusiones del proyecto, se sugieren las siguientes recomendaciones para optimizar el rendimiento y la calidad del asistente virtual:

Aumentar y perfeccionar la base de datos de textos bilingües en español y náhuatl, incorporando fuentes, diversidad y calidad. Esto permitiría entrenar un modelo de traducción, sólido y exacto, que pueda abarcar más casos y contextos.

Incluir técnicas de aprendizaje activo y adaptativo al modelo de traducción, para que pueda aprender de las retroalimentaciones y las correcciones de los usuarios, y así mejorar su rendimiento con el tiempo.

Agregar más funcionalidades y opciones al asistente virtual, como la posibilidad de elegir entre diferentes variantes o dialectos del náhuatl, la posibilidad de escuchar la pronunciación de las palabras o frases traducidas, o la posibilidad de acceder a recursos culturales o educativos en náhuatl.

Realizar una evaluación completa y participativa del asistente virtual, involucrando a hablantes nativos de náhuatl y a usuarios potenciales de la plataforma digital de intercambio económico cultural. Esto permitiría obtener una retroalimentación válida y confiable sobre la calidad y la utilidad del asistente virtual, así como identificar las necesidades y expectativas de los usuarios.

5.3 Experiencia profesional

Este proyecto ha brindado una valiosa experiencia profesional en el campo del Procesamiento de Lenguaje Natural, el Machine Learning y el fortalecimiento de las lenguas indígenas. A continuación, se detallan algunas de las competencias, habilidades y conocimientos profesionales que se han desarrollado o aplicado en este proyecto:

Competencia en Procesamiento de Lenguaje Natural

He obtenido un conocimiento de los conceptos, técnicas y herramientas del Procesamiento de Lenguaje Natural, una rama de la inteligencia artificial que se

ocupa de la interacción entre las computadoras y el lenguaje humano. He utilizado este conocimiento para entrenar un asistente virtual capaz de traducir textos del español al náhuatl.

Competencia en Machine Learning

He obtenido un conocimiento de los algoritmos, técnicas y herramientas del Machine Learning, una rama de la inteligencia artificial que se ocupa de entrenar a las computadoras para que aprendan de los datos. He utilizado este conocimiento para entrenar un modelo de traducción a partir del aprendizaje supervisado, utilizando un corpus paralelo de textos en español y náhuatl.

Competencia en gestión de datos

He obtenido una habilidad para manejar grandes conjuntos de datos, incluyendo la recolección, limpieza, transformación y análisis de los mismos. He utilizado esta habilidad para crear y depurar el corpus paralelo de textos en español y náhuatl, aplicando técnicas como la alineación de oraciones, la normalización ortográfica, la eliminación de ruido y la extracción de características.

Competencia en programación

He obtenido una habilidad para programar en diferentes lenguajes y entornos, como JavaScript, php, TensorFlow, etc. He utilizado la programación para entrenar el asistente virtual.

Esta experiencia profesional ha sido muy enriquecedora y satisfactoria para mí, ya que me ha permitido aplicar y desarrollar mis conocimientos y habilidades en un campo que me apasiona y que tiene una gran relevancia social. Creo que este proyecto contribuye a la preservación y el fortalecimiento de la lengua y la cultura

náhuatl, así como a la democratización y el acceso al conocimiento para las comunidades indígenas de México.

5.4 Experiencia Personal

Este proyecto ha proporcionado una serie de valiosas experiencias personales, relacionadas con los aspectos emocionales, motivacionales, actitudinales y valorativos del proceso de investigación.

A continuación, se detallan algunos de estos aspectos:

Sentimientos y emociones: Se ha vivido una variedad de sentimientos y emociones durante el desarrollo del proyecto, tanto positivos como negativos. Entre los sentimientos positivos, se resaltan el entusiasmo, la curiosidad, la satisfacción, el orgullo, la alegría, etc. Entre los sentimientos negativos, se resaltan el estrés, la frustración, la ansiedad, el miedo, la tristeza, etc.

Motivación: Se ha conservado una alta motivación durante el desarrollo del proyecto, debido al interés y la pasión por el tema de investigación, así como por el propósito social y cultural del mismo. Se ha sentido que el proyecto tiene un sentido y un valor, tanto personal como profesional, y que contribuye a una causa importante, como es el fortalecimiento de las lenguas indígenas.

Compromiso: Se ha evidenciado un alto compromiso durante el desarrollo del proyecto, al dedicar tiempo, esfuerzo y recursos para llevarlo a cabo. Se ha cumplido con las responsabilidades y los plazos establecidos, y se ha buscado la excelencia y la calidad en el trabajo realizado. Se ha asumido la responsabilidad de los propios errores y se ha buscado aprender de ellos.

5.5 Conclusiones relativas a los objetivos específicos

Se logró cumplir con los objetivos específicos en su mayoría. Se realizó una revisión bibliográfica exhaustiva sobre las tecnologías de Machine Learning aplicadas al problema de la traducción automática, especialmente de lenguas indígenas.

Se identificaron y analizaron los principales algoritmos, técnicas y herramientas utilizados en este campo, así como los desafíos y oportunidades que presenta. Sin embargo, no se logró llevar a cabo en su totalidad el entrenamiento de datasets y la generación de resultados, lo cual limitó el alcance del objetivo general de desarrollar un servicio de traducción para plataforma digital de intercambio económico cultural.

A pesar de esto, los hallazgos y análisis realizados contribuyen significativamente al campo de estudio y sientan las bases para futuras investigaciones y desarrollos.

5.6 Conclusiones relativas al objetivo general

El proyecto indica, a partir de los datos existentes, que el asistente virtual logra una traducción de textos del español al náhuatl con un rendimiento bastante bueno. No obstante, se admite que estos datos son escasos y que se necesita una evaluación más extensa y estricta por parte de hablantes nativos de náhuatl.

Para evaluar el rendimiento del asistente virtual, se han empleado indicadores técnicos como el BLEU y el METEOR, que contrastan las traducciones del asistente virtual con las de un traductor humano o con las de otras herramientas de traducción en línea.

No obstante, el proyecto también tiene algunas limitaciones y desafíos, como la escasez y la calidad de los datos disponibles para el entrenamiento del modelo de traducción, la complejidad y la diversidad de las lenguas indígenas, y la falta de

evaluación por parte de hablantes nativos de estas lenguas. Estas limitaciones afectan a la validez y la generalización de los resultados, y sugieren la necesidad de hacer más investigaciones y mejoras en el futuro.

5.7 Aportaciones originales

Aportación metodológica: Este proyecto contribuye al campo del Procesamiento de Lenguaje Natural y el Machine Learning, al emplear la metodología CRISP-DM como la más adecuada para el desarrollo del proyecto. Esta metodología es un referente de la industria para el análisis de datos y la minería de datos, que consta de seis fases: comprensión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

Esta metodología permite abordar el proyecto de forma sistemática, rigurosa y eficiente, garantizando la calidad y la validez de los resultados. Este proyecto es uno de los primeros en emplear esta metodología al problema de la traducción de lenguas indígenas, que presenta desafíos específicos como la escasez y la calidad de los datos, la complejidad y la diversidad de las lenguas, y la falta de evaluación por parte de hablantes nativos.

Este proyecto contribuye a divulgar y fomentar el uso de esta metodología en el campo del Procesamiento de Lenguaje Natural y el Machine Learning.

Aportación práctica: Este proyecto contribuye al campo del Procesamiento de Lenguaje Natural y el Machine Learning, al desarrollar e implementar un asistente virtual.

5.8 Limitaciones del modelo planteado

Con el entrenamiento y desarrollo del Asistente Virtual se lograron apreciar áreas de oportunidad, las cuales fueron surgiendo en el proceso de evaluación del modelo de traducción.

A continuación, se listan algunas de estas limitantes:

- Complejidad de la lengua
- Recursos limitados
- Calidad y cantidad de datos
- Tiempo y recursos computacionales
- Adaptabilidad y generalización

CAPÍTULO VI
COMPETENCIAS DESARROLLADAS

6.1 Competencias desarrolladas y/o aplicadas

- Trabajo en equipo
- Gestión de datos
- Toma decisiones
- Resolución de problemas
- Impacto social y cultural
- Competencia en Machine Learning

CAPÍTULO VII
FUENTES DE INFORMACIÓN

8.1 Referencias

Alvarenga Alcaraz, N. B., Alvarenga Alcaraz, P. R., Alvarenga Alcaraz, N. B., & Alvarenga Alcaraz, P. R. (2020). Aplicación web de Análisis y Traducción Automática Guaraní-Español / Español—Guaraní. *Revista Científica de la UCSA*, *7*(2), 41–69. https://doi.org/10.18004/ucsa/2409-8752/2020.007.02.041

anandmeg. (2023, octubre 31). *¿Qué es Visual Studio?* https://learn.microsoft.com/es-es/visualstudio/get-started/visual-studio-ide?view=vs-2022

Aragones, X. F., & Graells, J. A. (2019). *Técnicas de clasificación supervised learning*.

Bender, A., Nicolet, S., Folino, P., Lopez, J. J., & Hansen, G. (2023). Generación Automática de Código Fuente a través de Modelos Preentrenados de Lenguaje. *Electronic Journal of SADIO (EJS)*, *22*(1), Article 1.

Elías, A. N. O., & Morán, R. C. D. (2023). Implementación de un asistente virtual para los estudiantes de pregrado de una universidad peruana. *Revista Conrado*, *19*(92), Article 92.

Espinosa-Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería,*

investigación y tecnología, *21*(1). https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Gavilán, I. G. R. (2021, julio 5). Metodología para Machine Learning (III): SEMMA. *Ignacio G.R. Gaviláin*. https://ignaciogavilan.com/metodologia-para-machine-learning-iii-semma/

Hernandez, P. R., & Cruz, D. V. (2022). Los Asistentes virtuales basados en Inteligencia Artificial. *ReCIBE, Revista electrónica de Computación, Informática, Biomédica y Electrónica*, *11*(2), Article 2. https://doi.org/10.32870/recibe.v11i2.251

Jofre, C. M. (2023). *ChatGPT, Inteligencia Artificial y Universidad. Nuevas tensiones, transformaciones y desafíos en la educación superior.*

JSONL: Definition, JSON Lines Format, Use Cases, and More. (2022, septiembre 11). DevOps and Software Engineering Glossary Terms | Atatus. https://www.atatus.com/glossary/jsonl/

Monsiváis, F. P. E. (2019). Contexto del intérprete y traductor de lenguas indígenas en México para el desempeño de su labor. *BInvestigación*, *1*, Article 1. https://revistas.uaa.mx/index.php/bi/article/view/1808

Moreno Padilla, R. D. (2019). La llegada de la inteligencia artificial a la educación. *Revista de Investigación en Tecnologías de la Información: RITI*, *7*(14), 260–270.

Ochoa Toro, A. M. (2023). *Lenguaje de programación para el aprendizaje por refuerzo*. https://repositorio.uniandes.edu.co/entities/publication/3ddca825-21bc-486c-aa2b-1f7b17357935

Pacheco, C. A. V. (2023). ChatGPT: Furores y límites. *Cuadernos Fronterizos*, *58*, Article 58. https://doi.org/10.20983/cuadfront.2023.58.10

Pérez, M. D. S. (2020). Análisis del uso de asistentes virtuales en el aula como recurso complementario en la práctica docente. *Enseñanza y Aprendizaje de Ingeniería de Computadores*. https://doi.org/10.30827/Digibug.64782

Procesamiento del lenguaje natural (PLN) - GPT-3.: Aplicación en la Ingeniería de Software | Tecnología Investigación y Academia. (2021). https://revistas.udistrital.edu.co/index.php/tia/article/view/17323

Sánchez, L. R. A., Sánchez, B. A. B., Ángel, L. D. Á. D., & Aguilar, C. J. R. (2021). Asistente virtual para registro de siniestro vehicular con aplicación en WhatsApp: Virtual assistant for vehicular claim registration with WhatsApp application. *Tecnología Educativa Revista CONAIC*, *8*(3), Article 3. https://doi.org/10.32671/terc.v8i3.220

Sevilla Salcedo, J., Carrasco-Martínez, S., Castillo, J. C., Castro-González, Á., & Salichs, M. Á. (2022). *Modelos de lenguaje natural para robots sociales*. 828–834. https://ruc.udc.es/dspace/handle/2183/31342

Terrones Rodríguez, A. L. (2020). Inteligencia artificial, responsabilidad y compromiso cívico y democrático. *Terrones Rodríguez, Antonio Luis 2020 Inteligencia artificial, responsabilidad y compromiso cívico y democrático Revista CTS 15 44 253 276*. https://roderic.uv.es/handle/10550/86325

CAPÍTULO VIII ANEXOS

8.2 Asignación de Asesor, Comisión Revisora, Entrega de trabajo Profesional y Dictamen

EDUCACIÓN

Instituto Tecnológico Superior de Teziutlán

Asunto: **Asignación de Asesor(a), Comisión Revisora, Entrega de Trabajo Profesional y Dictamen**

Teziutlán, Puebla, 03/enero/2024

Asesor(a): **EDGAR DEGANTE AGUILAR**
Integrante de Comisión Revisora: **PATRICIA OCHOA TRUJILLO**
Integrante de Comisión Revisora: **EMMANUEL VAZQUEZ BENITO**
Presentes

Por este medio me permito informar que ha sido asignado como asesor(a) y comisión revisora del trabajo profesional que se convertirá en Tesis de:

Alumno(a):	**ORTIZ BAUTISTA ADOLFO** (Apellido paterno/materno/nombre (s))		
Número de Control:	**19TE0477**	Licenciatura o Posgrado:	**INGENIERIA INFORMATICA**
Plan:	**2010**	Correo Electrónico:	**L19TE0477@TEZIUTLAN.TECNM.MX**

Cuyo tema es: **DESARROLLO DE UNA ASISTENTE VIRTUAL CON IMPLEMENTACION DE PNL EN PLATAFORMA DE INTERCAMBIO ECONOMICO CULTURAL PARA EL FORTALECIMIENTO DE LENGUAS INDIGENAS**

25 palabras (máximo)

Se ha enviado a su correo institucional el trabajo profesional o de grado, por lo cual la comisión revisora tendrá 5 día há biles para realizar las observaciones al alumnado, el(la) interesado(a) tendrá igualmente **5 días** para corregir y las enviará al correo electrónico institucional de la comisión revisora, agradezco de antemano su valioso apoyo en esta importante actividad para la formación profesional de licenciatura o de posgrado de nuestro alumnado egresado.

Dictamen de Comisión Revisora, Aprobación de Impresión o Grabación y Autorización para subirla al Repositorio del TecNM

Siendo el día: 13 de febrero de 2024 se reunieron los miembros de la comisión para revisar el trabajo asignado y una vez analizado se decidió liberarlo y aprobarlo para su grabación y programación de examen profesional.

EDGAR DEGANTE AGUILAR
Nombre y Firma del(la) Asesor(a)

PATRICIA OCHOA TRUJILLO
Nombre y Firma de integrante de la Comisión Revisora

EMMANUEL VAZQUEZ BENITO
Nombre y Firma de integrante de la Comisión Revisora

MYRIAM SÁNCHEZ PÉREZ
Subdirección Académica

R08/06/2023
Folio:28

S.E.P. GOBIERNO DEL ESTADO

F-SAC-18

INSTITUTO TECNOLÓGICO SUPERIOR DE TEZIUTLÁN SUBDIRECCIÓN ACADÉMICA

TEC de Teziutlán · 100% PLASTIC · PUEBLA

Fracción I y II s/n Aire Libre, Teziutlán, Puebla, C.P. 73960 Tels. 231 311 4000 / 4001 / 4002 / 4003
e-mail: itsteziutlan@hotmail.com | www.teziutlan.tecnm.mx

2024 Felipe Carrillo PUERTO

8.3 Carta de autorización del autor para la consulta y publicación electrónica del trabajo de investigación

Tecnológico Nacional de México
Instituto Tecnológico Superior de Teziutlán

CARTA DE AUTORIZACIÓN DEL(LA) AUTOR(A) PARA LA CONSULTA Y PUBLICACIÓN ELECTRÓNICA DEL TRABAJO DE INVESTIGACIÓN

El que suscribe:

ADOLFO	**ORTIZ**	**BAUTISTA**

Con Número de Control	**19TE0477**
Perteneciente al Programa Educativo	**INGENIERÍA INFORMÁTICA**

Por este conducto me permito informar que he dado mi autorización para la consulta y publicación electrónica del trabajo de investigación en los repositorios académicos.

Registrado con el producto:	**TESIS**

Cuyo Tema es:

ASISTENTE VIRTUAL CON IMPLEMENTACIÓN DE PNL EN PLATAFORMA DE INTERCAMBIO ECONÓMICO CULTURAL PARA EL FORTALECIMIENTO DE LENGUAS INDÍGENAS.

Correspondiente al periodo:

AGOSTO-DICIEMBRE 2023

Y cuyo(a) director(a) de tesis es:

M.S.C. EDGAR DEGANTE AGUILAR

ATENTAMENTE

ADOLFO ORTIZ BAUTISTA

Nombre y firma

Fecha de emisión: **07/02/2024**
c.c.p. Subdirección Académica

8.4 Formato de licencia de uso

LICENCIA DE USO OTORGADA POR **Adolfo Ortiz Bautista**, de nacionalidad **mexicana,** mayor de edad, con domicilio ubicado en la calle **Privada 1ero de Mayo, Tepepan, en el municipio de Chignautla, Puebla**, en mi calidad de titular de los derechos patrimoniales y morales y autor(a) de la tesis denominada **"Desarrollo de una Asistente virtual con implementación de PNL en plataforma de intercambio económico cultural para el fortalecimiento de lenguas indígenas"** en adelante "LA OBRA" quien para todos los fines del presente documento se denominará "EL(LA) AUTOR(A) Y/O EL(LA) TITULAR", a favor del Instituto Tecnológico **Superior de Teziutlán** del Tecnológico Nacional de México, la cual se regirá por las cláusulas siguientes:

PRIMERA –OBJETO: "EL(LA) AUTOR(A) Y/O EL(LA) TITULAR", mediante el presente documento otorga al Instituto Tecnológico **Superior de Teziutlán** del Tecnológico Nacional de México, licencia de uso gratuita e indefinida respecto de "LA OBRA", para almacenar, preservar, publicar, reproducir y/o divulgar la misma, con fines académicos, por cualquier medio en forma física y a través del repositorio institucional y del repositorio nacional, éste último consultable en la página: (https://www.repositorionacionalcti.mx/).

SEGUNDA - TERRITORIO: La presente licencia se otorga, de manera no exclusiva, sin limitación geográfica o territorial alguna, de manera gratuita e indefinida.

TERCERA -ALCANCE: La presente licencia contempla la autorización para formato uso de "LA OBRA" en cualquier formato o soporte material y se extiende a la utilización, de manera enunciativa más no limitativa a los siguientes medios: óptico, magnético, electrónico, virtual (en red), mensaje de datos o similar, conocido o por conocerse.

CUARTA – EXCLUSIVIDAD: La presente licencia de uso aquí establecida no implica exclusividad en favor del Instituto Tecnológico **Superior de Teziutlán**; por lo tanto, "EL(LA) AUTOR(A) Y/O EL(LA) TITULAR" conserva los derechos patrimoniales y morales de "LA OBRA", objeto del presente documento.

QUINTA – CRÉDITOS: El Instituto Tecnológico **Superior de Teziutlán** y/o el Tecnológico Nacional de México reconoce que "EL(LA) AUTOR(A) Y/O EL(LA) TITULAR" es el(la) único(a), primigenio(a) y perpetuo(a) titular de los derechos morales sobre "LA OBRA"; por lo tanto, siempre deberá otorgarle los créditos correspondientes por la autoría de la misma.

SEXTA – AUTORÍA: "EL(LA) AUTOR(A) Y/O EL(LA) TITULAR" manifiesta ser el(la) único(a) titular de los derechos de autor que derivan de "LA OBRA" y declara que el material objeto del presente fue realizado por él(ella), sin violentar o usurpar derechos de propiedad intelectual de terceros; por lo tanto, en caso de controversia sobre los mismos, se obliga a ser el(la) único(a) responsable.
Dado en la Ciudad de **Teziutlán, Puebla**, a los **once** días del mes de **febrero** de dos mil **veinticuatro.**

"EL(LA) AUTOR(A) Y/O EL(LA) TITULAR",	"EL INSTITUTO TECNOLÓGICO SUPERIOR DE TEZIUTLÁN"
Adolfo Ortiz Bautista	**Arminda Juárez Arroyo**
Nombre y Firma	Directora General Nombre, Firma y Sello

Índice de Ilustraciones