MÁSTER EN DATA SCIENCE PARA FINANZAS

Colegio Universitario de Estudios Financieros

2020/21

TRABAJO FIN DE MÁSTER

Utilización de modelos de Transformers en la gestión de respuestas a preguntas en PLN. Aplicación a asistentes conversacionales

Autores: Afán de Ribera Olaso, Isabel; Blanco García, Gabriel;
Díaz Torres, Valentina; Jiménez Zúñiga, Andrea

Tutor/a: Izquierdo Catalán, Francisco

Convocatoria: Julio 2021

ÍNDICE

RESUMEN **4**

ABSTRACT **4**

INTRODUCCIÓN **5**

MOTIVACIÓN DEL TRABAJO **6**

CAPÍTULO I: HISTORIA Y APROXIMACIÓN A LOS ASISTENTES VIRTUALES Y A LOS TRANSFORMERS. **8**

1. ASISTENTES VIRTUALES, EMBEDDINGS Y TRANSFORMERS **8**

1.1. HISTORIA 8

1.2. CONCEPTOS TEÓRICOS 10

1.2.1. Asistentes Virtuales 10

1.2.2. Los Transformers 11

1.3. CASOS DE USO Y CASOS ACTUALES 15

CAPÍTULO II: VISIÓN DE NEGOCIO, CON PLAN DE MARKETING Y PLAN FINANCIERO. **16**

2. VISIÓN DEL NEGOCIO **16**

2.1. APLICACIÓN EN EL SECTOR TURÍSTICO 16

2.2. PLAN DE MARKETING 17

2.2.1. Mercado Objetivo 17

2.2.2. Segmentación 18

2.3. MARCO LEGAL 20

2.4. PLAN FINANCIERO 21

2.4.1. Caso WhatsApp 21

CAPÍTULO III: METODOLOGÍA Y PROPUESTA. ESTRUCTURA DEL ASISTENTE Y VENTAJAS PRINCIPALES. **25**

3. METODOLOGÍA Y PROPUESTA **25**

3.1. METODOLOGÍA: CHATBOT DESIGN CANVAS 25

3.2. NUESTRA PROPUESTA 29

3.2.1. Uso del Transformer 31

CAPÍTULO IV: DESARROLLO TÉCNICO DEL PROYECTO **32**

4. DESARROLLO TÉCNICO **32**

4.1. FUNCIONAMIENTO 32

4.1.1. El asistente 33

4.1.2. Los datos 33
4.1.3. El Transformer 34
4.2. INTEGRACIÓN Y ACCESO: TELEGRAM 35
4.3. DESARROLLO DEL SISTEMA 36
4.3.1. Fases del sistema 37
4.4. PORTABILIDAD, SEGURIDAD Y DESPLIEGUE 40
4.4.1. Conflictos entre dependencias 40
4.4.2. Conflictos entre versiones de una misma librería 41
4.4.3. Conflictos entre sistemas operativos 41
4.4.4. Docker 41
CAPÍTULO V: PROBLEMAS ENCONTRADOS. 43
5. PROBLEMAS ENCONTRADOS 43
5.1. PROBLEMAS RELACIONADOS CON EL TEXTO 43
5.2. PROBLEMAS RELACIONADOS CON LOS MODELOS 45
CONCLUSIONES 46
BIBLIOGRAFÍA 48

RESUMEN

Los modelos de Transformers llegaron al Procesamiento de Lenguaje Natural en 2017 aportando eficiencia, rapidez y mejores resultados a todo lo que se había hecho anteriormente. A partir de la Atención y novedosos mecanismos, han conseguido mayor facilidad para la gestión de conversaciones, proporcionar respuestas concretas y cerradas y una fácil escalabilidad. El objetivo de este estudio será construir una simulación completa, aplicando estos modelos a un asistente virtual orientado al turismo, con el fin de que este proporcione respuestas precisas, extraídas de diferentes textos. Además, se tratará el plan estratégico, financiero, y de arquitectura, entre otras cuestiones que serían necesarias para ser implementado.

Palabras clave: Asistente virtual, Machine Learning, Deep Learning, Transformers, Inteligencia Artificial, Embeddings, atención, turismo.

ABSTRACT

The Transformers models arrived at the Natural Language Processing in 2017 bringing efficiency, speed and better results to everything that had been done before. From the Attention and novel mechanisms, they have achieved greater ease for the management of conversations, providing concrete and closed answers and an easy scalability. The aim of this study will be to construct a complete simulation, applying these models to a virtual assistant oriented to tourism, in order that it provides precise answers, drawn from different texts. In addition, the strategic, financial, and architectural plan will be addressed, among other issues that would be necessary to be implemented.

Key words: Virtual assistant, Machine Learning, Deep Learning, Transformers, Artificial Intelligencce, Embeddings, Attention, Tourism.

INTRODUCCIÓN

Desde hace unas décadas, la investigación sobre el Procesamiento de Lenguaje Natural y la constante mejora en la comunicación entre máquinas y humanos son temas de especial interés en el campo de la Inteligencia Artificial.

Este campo de estudio de la lingüística computacional tiene numerosas aplicaciones como, por ejemplo, la traducción automática, la búsqueda avanzada de información o la clasificación automática de documentos y mensajes, entre otras. Sin embargo, este proyecto se enfoca en el desarrollo del Procesamiento de Lenguaje Natural en lo relativo a respuestas a preguntas mediante asistentes conversacionales.

Un asistente conversacional se trata de un programa que simula mantener una conversación con una persona al proveer respuestas automáticas a entradas hechas por el usuario. Una tecnología que empezó a plantearse en 1950 por Alan Turing.

En un primer momento, los asistentes conversacionales eran desarrollados de una manera muy limitada pues era necesario programar de forma explícita las preguntas y respuestas de tal manera que cuando el usuario se salía de la forma en la que la pregunta había sido configurada el asistente perdía utilidad. Además, hay que añadir las limitaciones por diferencias de idioma y la sensibilidad a la ortografía.

Sin embargo, en 2017 la creación de los modelos de Transformers publicados en el paper "Attention is all you need" supusieron un gran avance en el desarrollo del Procesamiento de Lenguaje Natural permitiendo, entre otras cosas, superar los límites antes mencionados en los asistentes conversacionales. Principalmente, la idea detrás de los Transformers pasa por utilizar encodings posicionales junto con la codificación de las palabras permitiendo el reconocimiento de palabras claves en las preguntas, la identificación de estas en el texto y con ello la extracción de la respuesta exacta.

Ante esta situación, el presente Trabajo Fin de Máster tiene como objetivo principal poner a prueba esta reciente tecnología a través de su aplicación en el desarrollo de un asistente virtual enfocado al turismo. Mediante este asistente los usuarios podrán obtener respuestas exactas a sus preguntas sobre distintos lugares de interés turístico en la ciudad de Madrid como pueden ser museos, monumentos, restaurantes u otros lugares destacados.

En cuanto a su estructura, este Trabajo Fin de Máster comienza con un apartado de motivación, se divide en cinco capítulos y finaliza con un apartado de conclusiones. El Capítulo I pretende proporcionar una base para la comprensión de los pilares en los que se fundamenta esta investigación: los asistentes virtuales y los Transformers. Se comienza con una visión más general hablando sobre la historia y evolución de los asistentes virtuales, el Procesamiento de Lenguaje Natural y los modelos de Transformers. Seguidamente, se incluyen explicaciones teóricas sobre estos conceptos a fin de poner al lector en contexto y finalmente se comentan algunos casos de uso actuales.

El Capítulo II aborda todo lo relativo a la visión de negocio de este proyecto pues con esta investigación no solo se pretende ampliar los conocimientos y poner a prueba las capacidades de la tecnología de los Transformers sino también llegar a implementar el sistema de tal manera que el mundo empresarial pueda beneficiarse de los hallazgos y

avances obtenidos. Este capítulo empieza con una breve explicación del porqué de la aplicación del asistente al sector del turismo. Posteriormente, continúa con el plan de marketing establecido, el cual incluye todo lo relativo a mercado objetivo, precio, producto, punto de venta y promoción. Continúa con un breve marco legal en el que se han planteado las barreras legales a las que podría enfrentarse este asistente virtual. Y, finalmente, incluye un plan financiero en el que se han analizado las fuentes de ingresos y costes para el desarrollo del proyecto.

El Capítulo III trata de dar una visión global del proyecto de creación del asistente virtual enriquecido por modelos de Transformers a través de la metodología *Chatbot Design Canvas* donde se analizan cuestiones claves para el establecimiento del plan estratégico seguido. Además, esto es complementado con una explicación de la propuesta del proyecto en la que se comentan los beneficios que aporta la tecnología de los Transformers frente a los asistentes tradicionales.

En el Capítulo IV se comienza con el desarrollo de la parte técnica. En primer lugar, se explica el funcionamiento del sistema empezando por el asistente, continuando por los datos y finalizando con el propio Transformer. En segundo lugar, se relata el desarrollo del asistente, hecho en su totalidad con el software de programación Python, a través de cinco fases. En tercer lugar, se continúa con todo lo relativo a la integración y acceso del asistente por medio de Telegram. Por último, se incluye un apartado de portabilidad, seguridad y despliegue donde Docker es el principal protagonista y en el que se hace referencia a los problemas que pueden encontrarse al implementar el asistente por conflictos entre dependencias, versiones de una misma librería o sistemas operativos, así como una propuesta de solución.

En el Capítulo V se analizan todos los problemas encontrados durante el desarrollo del asistente virtual empezando por todos aquellos relacionados con los textos sobre los distintos lugares turístico y de ocio de la ciudad de Madrid incluidos en el sistema y mediante los cuales se ha ofrecen respuesta a los usuarios. Y todos los referidos a los modelos de Transformers en sus limitaciones por falta de razonamiento humano.

Finalmente, este Trabajo Fin de Máster incluye una serie de conclusiones donde se analizan los descubrimientos hechos durante esta investigación y las limitaciones y puntos de mejora de la incorporación de la tecnología analizada en los asistentes virtuales.

MOTIVACIÓN DEL TRABAJO

El empleo de asistentes conversacionales se ha convertido en los últimos años en un "trending topic" dentro del marco de la transformación digital de las compañías como consecuencia del cambio experimentado por la sociedad, en lo que a comunicación y hábitos de compra se refiere. Actualmente, las aplicaciones de mensajería y las redes sociales se han convertido en la herramienta de conversación por excelencia y un medio imprescindible en la comunicación directa entre las empresas y los clientes.

Los asistentes virtuales traen consigo numerosas ventajas en lo que respecta a atención al cliente. Primero, es una interfaz que se encuentra operativa 24/7 y proporcionar respuesta de forma automática.

Segundo, se integran fácilmente en aplicaciones de mensajería estando disponibles en todo momento y su uso resulta bastante sencillo pues se trata de un formato conocido por jóvenes y adultos.

Tercero, mejora la experiencia del cliente, quien se siente atendido en todo momento.

Por último, supone un ahorro de costes considerable en las empresas que lo implementan al tener que evitar elevadas contrataciones de personal.

A pesar de ello, el modo en el que hasta hace cuatro años se han venido implementando estos asistentes presenta múltiples limitaciones, las cuales se desarrollaran durante este trabajo. Entre ellas están las distintas formas de expresar una intención o pregunta, la ortografía, el idioma y la escalabilidad.

Ante esta situación, el presente Trabajo Fin de Máster tiene como objetivo principal desarrollar un asistente conversacional incorporando la tecnología de los Transformers, y con ello las numerosas ventajas que esta proporciona, de tal manera que se aprovechen en la mayor medida posible las capacidades del Procesamiento del Lenguaje Natural y de la Inteligencia Artificial.

Esta tecnología es básicamente una arquitectura de red neuronal basada únicamente en el mecanismo de auto-atención y muy paralelizable. Se basan en incrustaciones de entrada y salida formando capas, toman una secuencia de palabras, y la máquina aprende una representación vectorial para cada palabra. Entre los beneficios que aporta, y que se expondrán a lo largo del trabajo está.

Por un lado, evitar una tediosa programación de respuestas para todas las posibles preguntas que pueda hacer el usuario en busca de una misma respuesta. En el caso de asistentes tradicionales es necesario definir intenciones, definir el flujo de la conversación en cada pregunta mediante cajas de diálogo, y en los casos en los que hay muchas preguntas, eso supone mucho trabajo. Sin embargo, los modelos de Transformers se basan en el reconocimiento de palabras clave y su búsqueda en el texto que contiene la información evitando lo anterior. Por otro, superar las barreras del idioma que presentan los asistentes tradicionales ya que los modelos de Transformers son capaces de proveer la respuesta exacta independientemente del idioma en el que se pregunte.

Además, para ello ha sido elegido como ámbito de implementación el sector turístico, pues teniendo en cuenta la gravedad con la que se ha visto afectado este último año debido a los efectos de la pandemia del coronavirus y el enorme potencial que el sistema de los asistentes conversacionales tiene en él por los beneficios en atención al cliente se ha considerado una industria muy acorde.

CAPÍTULO I: HISTORIA Y APROXIMACIÓN A LOS ASISTENTES VIRTUALES Y A LOS TRANSFORMERS.

1. ASISTENTES VIRTUALES, EMBEDDINGS Y TRANSFORMERS

En este apartado se tendrán en cuenta las aproximaciones principales a los asistentes virtuales y a la aplicación de los Transformers en ellos. Para ello se tratarán algunos

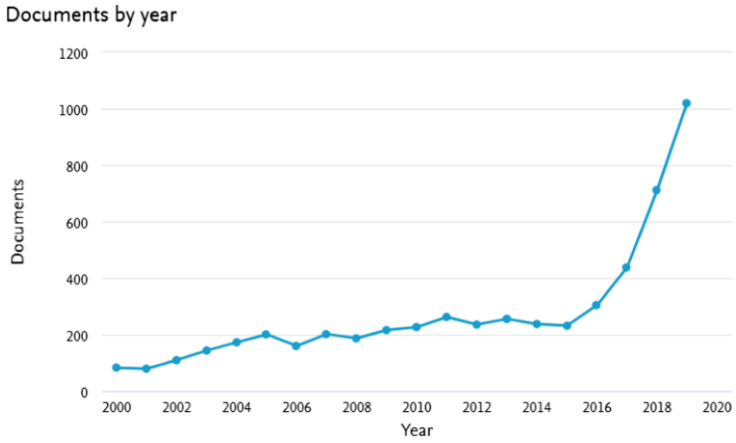
conceptos básicos, necesarios para comprender qué es un asistente virtual, así como otros conceptos relacionados con la materia, modelos y técnicas utilizadas, además de qué es un modelo de Transformer. También, será necesario comentar brevemente la historia de los asistentes virtuales, con el fin de lograr una contextualización para el lector. Por último, se hablará de algunos de los casos de uso que tendrían los Transformers y los asistentes virtuales y se comentarán algunas de las empresas que ya los tienen.

1.1. HISTORIA

Alan Turing fue uno de los propulsores de los asistentes virtuales, en 1950, propuso el Test de Turing, que planteaba el concepto de la posibilidad de que las máquinas pudiesen pensar como humanos, imitando su comportamiento.

En 1966 consta el primer asistente virtual conocido, Eliza. La finalidad de este era cumplir las funciones de un psicoterapeuta, devolviendo al usuario preguntas. Este usaba concordancia de patrones y un mecanismo de respuesta basado en unas plantillas. Su capacidad de respuesta no fue muy precisa, pero sorprendió a la gente del momento, pues era una de las primeras veces que una persona interactuaba con una máquina. La mejora de Eliza vino en 1972, llamado Parry, seguido de Alice, el cual ganó el premio Loebner en los años 2000, 2001 y 2004. Alice utilizaba un algoritmo de concordancia de patrones, pero con inteligencia subyacente basada en el Lenguaje de Marcado de Inteligencia Artificial, lo que permitía a los desarrolladores definir los bloques de construcción del conocimiento del propio asistente virtual. (Adamopoulou & Moussiades, 2020)

De 2000 a 2004 se obtuvo el rango de ordenador más humano. Algunos como SmarterChild, asistentes virtuales sirvieron para dar suporte a aplicaciones de mensajería. A todo esto, le siguió los conocidos: Apple Siri, Microsoft Cortana, Amazon Alexa, Google Assistant o IBM Watson. Fue a partir de 2016 cuando elinterés por los asistentes virtuales se desplegó y creció de forma exponencial, como se puede observar en la siguiente gráfica. (Adamopoulou & Moussiades, 2020)



Gráfica 1. Interés por asistentes virtuales. Fuente: (Adamopoulou & Moussiades, 2020)

Los asistentes virtuales fueron desarrollando cada vez más, no solo por la capacidad de responder a preguntas de forma rápida, sino por el gran beneficio que supone para una empresa, tal como la reducción de costes en atención al cliente y la capacidad de atender a muchos usuarios a la vez. La imagen de los asistentes virtuales se ha consolidado, ya no solo como un mero asistente, sino también, un compañero o colega más. Según algunos estudios los asistentes virtuales tienen un 40% de emotividad, un asistente virtual puede conseguir este carácter gracias al aprendizaje automático, lo que les brinda la posibilidad de detectar sentimientos, y de relacionarse con los clientes, de forma parecida a como lo podría hacer un humano. Esta semejanza humana se puede conseguir mediante señales visuales, nombres de humanos asociados, una identidad definida y señales de conversación que imiten a una conversación humana. Sin embargo, hay que tener en cuenta sus limitaciones, aún quedan avances por hacer en la personificación, ya que estos son incapaces de poseer empatía y entender matices en la conversación. En esta personificación, se revela que podría existir una visión sesgada de género, ya que la mayoría de los asistentes virtuales realizan tareas que históricamente han desempeñado roles femeninos y además son codificados como femeninos, con nombres que también lo son. (Adamopoulou & Moussiades, 2020)

Los modelos de procesamiento del lenguaje natural tradicionales estaban basados en reglas explícitas, lo que los hacía impracticables, ya que se necesitaba el conocimiento de esas reglas y los medios para ejecutarlas. La mayoría de los sistemas actuales de Deep Learning ya no poseen esas reglas y pueden ser ejecutados por todos. Sin embargo, tienen el inconveniente de ser poco interpretables, siendo difícil conocer si la salida ofrecida por el modelo es correcta o no, y de dónde ha salido. Estos se basan en redes neuronales artificiales con aprendizaje automático, partiendo de ejemplos. (Forcada, 2020)

Los Transformers, objeto de este trabajo, tuvieron un especial auge a finales de 2017, cuando Google publicó un paper “Attention is All You Need”, donde presentó la arquitectura de un Transformer, un modelo totalmente nuevo. La diferencia que aportaba este modelo es que sustituía las capas recurrentes por capas de atención. Las Redes Recurrentes eran, hasta ese momento, una de las mejores formas de capturar las

dependencias oportunas en las secuencias. Pero con la aparición de los Transformers, se consiguieron mejores resultados. (Vaca, 2021)

Los Transformers aparecieron para resolver el problema de la paralelización, utilizando redes neuronales junto con modelos de atención, que aumentan la velocidad con la que el modelo puede traducir de una secuencia a otra. Esto fue un gran avance, lo que ha hecho que estos modelos aportaran muchos mejores resultados que todo lo anterior.

Uno de los primeros modelos sería BERT (Bidirectional Encoder Representations from Transformers), un nuevo modelo de los investigadores de Google AI Language, presentado y de código abierto a finales de 2018, y desde entonces ha causado un gran revuelo en la comunidad de PNL. La innovación clave del que aportó el modelo BERT radicó en la aplicación del entrenamiento bidireccional de los modelos Transformer al modelado del lenguaje. Los resultados demostrados por el modelo BERT mostraron que un modelo lingüístico entrenado bidireccionalmente puede tener un sentido más profundo del contexto y el flujo del lenguaje que los modelos lingüísticos de una sola dirección. BERT tiene numerosas aplicaciones como la respuesta a preguntas, análisis de sentimientos o la clasificación de documentos. (koleva, 2020)

1.2. CONCEPTOS TEÓRICOS

1.2.1. Asistentes Virtuales

Aquí se tratarán algunos de los conceptos necesarios de cara al entendimiento y preparación para el resto del trabajo, básicamente centrados en Transformers y en su contexto tecnológico previo.

Un asistente virtual consiste en un programa informático que responde como una entidad inteligente cuando se conversa con él, mediante texto o voz, y entiende uno o más idiomas humanos mediante el Procesamiento del Lenguaje Natural. Es decir, un programa diseñado para simular una conversación con usuarios humanos, especialmente a través de Internet. Estos, también son conocidos como bots inteligentes, agentes interactivos, asistentes digitales o entidades de conversación artificiales. Los asistentes virtuales tienen múltiples funcionalidades, además de imitar una conversación humana puede ser aplicados a negocios, medicina, turismo, entre otros. (Cameron, y otros, 2017)

Entre algunas de las técnicas que son importante destacar encontramos el PatternMatching, esta consiste en introducir un estímulo en la entrada y una respuesta en la salida y se crea una consonancia con lo que introduce el usuario. Eliza y Alice, nombradas anteriormente, fueron los primeros asistentes virtuales en usar esto. La desventaja es que carecen de la personificación de la que se hablaba más arriba. De 1995 a 2000 se creó el Marcado de Inteligencia Artificial (AIML), basado en los conceptos de Reconocimiento de Patrones o Coincidencia de Patrones. Esto se aplica al modelado del lenguaje natural para el diálogo entre humanos y asistentes virtuales que siguen un enfoque estímulo-respuesta. Son unidades básicas de diálogo denominadas categorías, formadas por patrones de entradas del usuario y respuestas el asistente virtual. (Adamopoulou & Moussiades, 2020)

También, es importante comentar el Análisis Semántico Latente (LSA) puede utilizarse junto con AIML para el desarrollo de asistentes virtuales. Se emplea para descubrir similitudes entre palabras como representación vectorial. Su uso radica principalmente en las preguntas sin respuestas en plantillas, que tienen que ser respondidas a partir de LSA, es decir buscando similitud en otras palabras para proporcionar una respuesta a la pregunta, que se pueda acercar a la realidad. (Adamopoulou & Moussiades, 2020)

Chatscript sucedió al AIML, este es un sistema experto, que se compone de reglas que se asocian a temas, encontrando el mejor elemento que coincida con la cadena de consulta del usuario y ejecutando una regla en ese tema. Este también incluye memoria a largo plazo en forma de variables que se pueden utilizar para almacenar información específica del usuario como el nombre o la edad. Además, distingue entre mayúsculas y minúsculas, ampliando las posibles respuestas que pueden dar a la misma entrada del usuario en función de la emoción que se pretenda. (Adamopoulou & Moussiades, 2020)

Todo lo anterior, puede englobarse en uno de los conceptos básicos como es el de Procesamiento del Lenguaje Natural (PLN), un área de la inteligencia artificial que explora la manipulación de textos o discursos en lenguaje natural por parte de los ordenadores. La mayoría de las técnicas de PNL se basan en el aprendizaje automático como motor principal. Además, la comprensión del lenguaje natural (NLU) es el núcleo de cualquier tarea de PNL, es una técnica para implementar interfaces de usuario naturales, como un asistente virtual. Su objetivo es extraer el contexto y los significados de lo que el usuario introduce, que puede estar estructurado o no y una vez que la intención del usuario ha sido identificada, extrae identidades específicas de ese lenguaje, para poder proporcionar una respuesta. (Adamopoulou & Moussiades, 2020)

El Procesamiento de Lenguaje Natural ha evolucionado mucho, hasta hace diez años la forma que existía para solventar un problema de análisis de texto era con variables del texto. Lo que realmente revolucionó el Procesamiento de Lenguaje Natural fue la aparición del Deep Learning y de los Ebeddings. Estos últimos son representaciones matriciales y estáticas de un texto, donde cada palabra aparece codificada por un vector diferente. Hasta 2017, los modelos basados en Redes Neuronales Recurrentes, que usaban embeddings, se convirtieron en los más comunes. Estos, además de tener en cuenta el significado de la palabra, también lo hacían de la posición en la que se encontraba en el texto completo. (Adamopoulou & Moussiades, 2020)

Actualmente, el 99% de los asistentes virtuales que podemos encontrar en el mercado siguen el paradigma de "detección de intención". Esto quiere decir que el asistente virtual trata de captar (o detectar) la intención del usuario a partir del lenguaje natural (texto o voz). Una vez captada la intención del usuario, el asistente virtual puede reaccionar y enviar la conversación al flujo de diálogo adecuado. (Mora, 2020)

1.2.2. Los Transformers

Sin embargo, desde que los Transformers irrumpieron en el PLN se han convertido en los más usados en análisis de texto, siendo los modelos que mejores resultados han obtenido en las distintas aplicaciones de PLN. Entre ellas estarían el resumen y

generación de textos, la identificación de identidades, las respuestas a preguntas, la desambiguación de textos, entre otros.

Los Transformers se basan en búsqueda semántica, lo que permite buscar similitudes entre los textos por su significado, en lugar de usar palabras clave, como lo hacen los asistentes virtuales tradicionales. En pocas palabras, estos nuevos modelos pueden predecir la probabilidad condicional de una palabra (o secuencia de palabras) dado un contexto. (Mora, 2020)

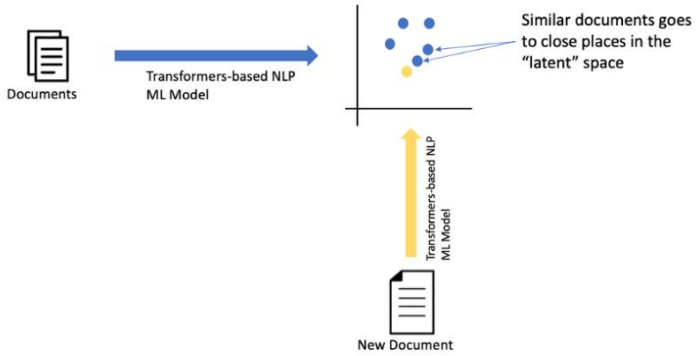


Ilustración 1. How semantic search works. Fuente: Towards Data Science

Para realizar la traducción de frases, el Transformer averigua las dependencias y conexiones. Esto lo hace, mediante las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN). Las RNN, consisten en bucles que permiten que la información persista y pase de un paso del proceso al siguiente. Para entender cómo funciona un bucle se podría atender a la siguiente imagen, donde se puede observar que la red neuronal consiste en varias copias de la misma red, que va pasando el mensaje a la siguiente red:

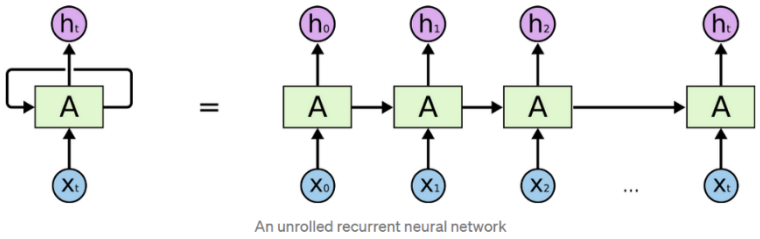


Ilustración 2. An unrolled recurrent neural network

Uno de los problemas de estas redes neuronales radica en que, cuando la frase no es clara y no hay mucha información sobre el contexto, estas pueden no ser tan precisas. La ventaja de las RNN es que pueden aprender a utilizar la información anterior, del contexto, y averiguar cuál es la siguiente palabra de esta frase. Pero, las RNN se vuelven muy ineficaces cuando la distancia entre la información relevante y el punto donde se necesita es muy grande. Esto se debe al hecho de que cuanto más larga es la cadena, más probable es que la información se pierda a lo largo de la misma. (Giacaglia, 2019)

La **Long-Short Term Memory (LSTM**) o memoria a largo plazo, es un tipo de RNN, que resuelve un problema principal. Cuando se va añadiendo nueva información a las cadenas de información previas, las RNN modifican toda la información, sin priorizar entre qué es importante y qué no. Sin embargo, en las LSTM, la información fluye por un mecanismo que se conoce como estados de celda. Ahí, mediante pequeñas modificaciones, se filtra la información relevante de la que no lo es tanto. No obstante, se encuentra el mismo problema encontrado anteriormente, cuando la frase es muy larga, se pierde precisión. Otro problema con RNNs y LSTMs es que es difícil paralelizar el trabajo de procesamiento de la frase, ya que implican que se procese palabra por palabra. Estos son aspectos que se deben tener en cuenta a lo largo del desarrollo del asistente virtual y que en este trabajo han sido tratados cuidadosamente. (Giacaglia, 2019)

Por otro lado, los Transformers usan las **redes neuronales convolucionales (CNN),** para solventar el problema de la paralelización, se emplean junto a los modelos de atención para resolver dicho problema y aumentar la velocidad. El motivo por el que solucionan la paralelización es porque cada nueva palabra que se introduce puede ser procesada al mismo tiempo y no depende de las palabras anteriores. Además, mejora el aspecto de la distancia entre palabras. (Giacaglia, 2019)

Ilustración 3. Funcionamiento transformer

A diferencia de las redes neuronales recurrentes, los Transformers no necesitan que los datos secuenciales se procesen en orden, lo que permite una paralelización mucho más precisa y reduce el tiempo de entrenamiento. Además, la auténtica novedad que compone el Transformer es que prescinde de las capas recurrentes utilizadas tradicionalmente. En su lugar, se basa por completo en la atención. Esta, consiste en prestar atención a palabras específicas. Las RNN pueden lograr este comportamiento, centrándose en parte de un subconjunto de la información que se les da. (kortschak, 2020)

Uno de los conceptos muy relacionados con la **atención** es la arquitectura de codificación y decodificación de los Transformers. Esto permite el procesamiento en paralelo y lo hace mucho más rápido que cualquier otro modelo con el mismo rendimiento. De este modo, prepararon el camino para los modelos lingüísticos modernos (como BERT y GPT) y, recientemente, también para los modelos de generación de imágenes. (kortschak, 2020)

El tipo de atención que usan los Transformers es el de la **autoatención**. Esta, a partir de la entrada que recibe, crea tres vectores: de consulta, de clave y de valor. Por ejemplo, la palabra a la que se le va a prestar atención en este caso es a “kicked”:

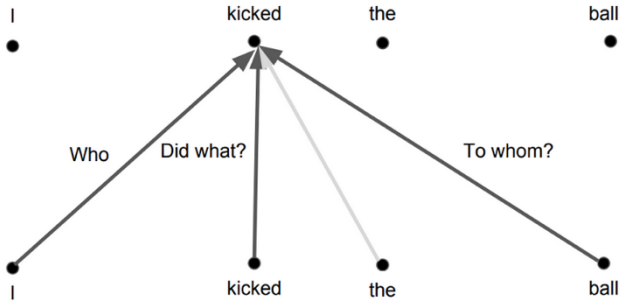


Ilustración 4. Ilustración kicked

En segundo lugar, se calcula una puntuación para cada palabra de la frase, con respecto a la palabra a la que se le quiere prestar atención. Tras ello se dividen las puntuaciones y se opera con ellas, de tal forma que todas sumen 1. Por último, cada puntuación se multiplica por los vectores previamente formados, pero, se intenta disminuir la importancia de aquellas palabras menos relevantes, por lo que se multiplican por números muy cercanos a 0, intentando reducir su importancia. Al sumar todos los valores de los vectores, esto produce una salida, de lo denominado como capa de atención que se utilizará en el análisis. (Giacaglia, 2019)

Una vez explicado de forma genérica cómo funciona la autoatención utilizada por los Transformers, se puede volver a la atención y a el modelo de codificación-decodificación. Las RNN, aportan una gran ventaja que es, que, en vez de codificar una frase entera, cada palabra tiene un "estado oculto" que se pasa por una etapa de decodificación. Esto se hace con la idea de que cada palabra puede esconder información relevante en una frase, por lo que hay que utilizar la atención, para hacer una decodificación precisa de las mismas. (Giacaglia, 2019)

La estructura de cada codificador y decodificador es muy similar entre ellas. A continuación, se pueden observar:

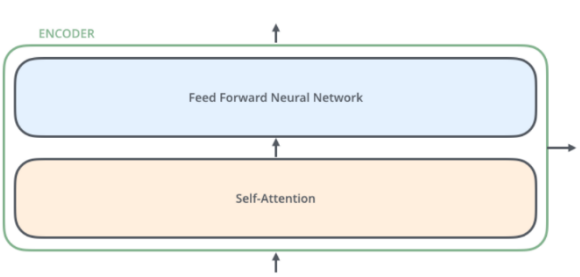


Ilustración 6. Estructura decodificador

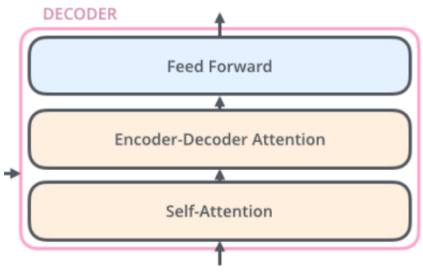


Ilustración 6. Estructura codificador

Tanto el codificador como el decodificador constan de dos y tres subcapas, respectivamente, la de autoatención y luego otras de red neuronal que alimenta a los siguientes pasos, conteniendo el decodificador una intermedia más. Además, trabajan en dos fases diferentes, la de Pre-training, donde el modelo aprende cómo se estructura

el lenguaje en general y la de Pre-tuning donde los modelos ya preentrenados se adaptan a tareas concretas. Estas dos fases hacen que sean modelos muy versátiles y que se puedan mover de una tarea a otra con gran facilidad. (kortschak, 2020)

Como hemos visto anteriormente, un Transformer, consta en su interior de un conjunto de capas de codificación, que procesan la entrada introducida capa tras capa. Además, también cuentan con un conjunto de capas de decodificación, que hacen el procesamiento para el output que proporciona el proceso de codificación. (Vajpayee, 2020)

Por último, la arquitectura secuencia a secuencia (Seq2Seq) en los Transformers hace que sean especialmente buenos en traducción, por traducir secuencias de palabras de un idioma y transformar esa misma en otro idioma. Con solo mecanismos de atención, sin necesidad de usar Redes Neuronales Recurrentes, el modelo de Transformers ha conseguido obtener mejores resultados en varias tareas. (Maxime, 2019)

1.3. CASOS DE USO Y CASOS ACTUALES

Los asistentes virtuales se han hecho muy comunes en numerosas empresas de todos los sectores, seguros, bancos, entre otros. Las compañías incorporan un asistente virtual con el fin de agilizar procesos en tareas que un tradicional asistente telefónico podría desempeñar. Incluso empresas de moda, como H&M, han desarrollado un asistente virtual para ayudar a los compradores con sus prendas. (Faggella, 2019)

Con el desarrollo de Procesamiento de Lenguaje Natural y la llegada de los modelos Transformer, se han obtenido mejores resultados, con un mayor rendimiento y velocidad, lo que ha permitido que surjan nuevos casos de uso, que pronto muchas organizaciones incorporarán.

Uno de estos casos es el de Google, mediante el uso del ya comentado anteriormente BERT. Desde el desarrollo de este modelo, el equipo de investigación de Google lo ha usado para mejorar la comprensión en las consultas en el buscador al poner en contexto cada palabra, lo que ayuda a comprenderlas mejor. Google decidió en 2018 hacer libre el código de BERT y esto hizo que muchas otras empresas como Microsoft o Facebook llevaran a cabo versiones de este modelo, aplicado a sus compañías. En el caso de Facebook, estos desarrollaron una versión llamada RoBERTa, para abordar la moderación de contenidos, intentando que aprendiera varios idiomas al mismo tiempo. Con esto, Facebook ha conseguido supervisar el contenido de forma automática y asegura bloquear aproximadamente el 70% más de contenidos nocivos que lo que hacía antes de este algoritmo. (koleva, 2020)

Pero no solo tienen gran aplicabilidad en las empresas tecnológicas, los Transformers se están desarrollando por todas las empresas, por ejemplo, para aquellas que trabajan a grandes distancias, en campos eólicos, en misiones submarinistas, entre otros escenarios, donde necesitan conseguir una respuesta precisa del funcionamiento de algún elemento. La precisión que alcanzan estos modelos hace que una respuesta exacta pueda ser proporcionada en un momento, y esto es un gran adelanto para muchas empresas.

Es conocido también el caso la Agencia Tributaria del Ministerio de Hacienda. Este implementó un asistente virtual, pero, sin embargo, no ha incluido aún el modelo de

Transformer. Actualmente, ante preguntas, como, por ejemplo, las relacionadas con la declaración de la renta, este responde con un párrafo donde se encuentra o podría encontrarse la respuesta a la pregunta. Sin embargo, si incorporarse este modelo, conseguiría dar la respuesta exacta a los ciudadanos. Esto conlleva la gran responsabilidad legal de proporcionar este dato equívocamente y es lo que hace que grandes organismos, sobre todo oficiales, se lo piensen dos veces antes de usarlo. No obstante, cada día están desarrollándose más y pronto muchas empresas lo adaptarán a su negocio.

CAPÍTULO II: VISIÓN DE NEGOCIO, CON PLAN DE MARKETING Y PLAN FINANCIERO.

2. VISIÓN DEL NEGOCIO

En esta sección se expondrá caso de uso, su aplicación al sector turismo. Se busca exponer el punto de vista económico de la idea, qué beneficios reales generaría a las empresas y a la sociedad, y por qué tiene sentido económico.

2.1. APLICACIÓN EN EL SECTOR TURÍSTICO

La Industria 4.0 está desencadenando cambios sociales y económicos masivos en todo el mundo, y con ello en todos los sectores de actividad, siendo uno de los principales afectados el sector Turismo, pues la forma en la que las personas planifican sus viajes hoy en día ha cambiado. Las empresas de viajes deben adaptarse, encontrar nuevas formas de responder a las necesidades de los viajeros y mejorar la experiencia del cliente con el objetivo de retener clientes, aumentar clientes recurrentes y captar a potenciales clientes.

Los viajeros de hoy en día ya no acuden a su agente de viajes para organizar sus viajes, están cada vez más conectados, poseen conocimientos digitales, y hacen toda su investigación de manera online antes de realizar un viaje. Como se muestra en un Estudio de atribución de viajeros realizado por Expedia Media Solutions, las personas terminan visitando 38 sitios web en promedio mientras planifican sus viajes y buscan cada vez más ofertas y planes de viaje personalizados.

En este nuevo contexto, la industria turística debe cambiar la forma en la que se relaciona con los usuarios y una de las formas de hacerlo es mediante la incorporación de la Inteligencia Artificial conversacional. Se sabe que uno de los principales retos está en la exigencia, cada vez mayor, por parte de los clientes de obtener una respuesta rápida a sus preguntas, siendo fundamental para ello una disponibilidad ilimitada, las 24 horas del día los 7 días de la semana. Como ya hemos comentado, el enfoque digital es el más eficiente para cumplir con estos deseos. Precisamente, en este marco entran en juego los asistentes conversacionales en los que se centra este trabajo de investigación debido a la multitud de beneficios que puede aportar al sector del turismo. Pero ¿cuáles son estos retos?

En primer lugar, mejorar la atención al cliente. Los clientes esperan una comunicación instantánea, especialmente si necesitan consejo o ayuda algo que puede conseguirse fácilmente con un asistente virtual integrado en alguna aplicación de mensajería móvil, en cualquier momento y a cualquier hora. Además, al mismo tiempo esto permite a la empresa de turismo reducir significativamente sus costes sustituyendo o reduciendo las horas de trabajo de los encargados de atención al cliente.

En segundo lugar, optimizar el tiempo y sacar el máximo partido al viaje. El hecho de tener una conversación personalizada con la que obtener respuestas y sugerencias instantáneas hará la experiencia mucho más agradable evitando perder muchas horas investigando en sitios webs o preguntando a terceras personas.

En tercer lugar, recopilación y aprovechamiento de datos por parte de la empresa. Mediante un asistente virtual se puede almacenar y procesar mucha información del usuario: datos cualitativos, experiencias, comentarios, quejas que permiten a las empresas realizar estudios de mercado y así trabajar de forma más eficaz en sus estrategias de marketing.

Como ya se ha expuesto, los asistentes virtuales tienen múltiples beneficios para las empresas turísticas. Disponibilidad 24/7, tiempo de respuesta rápido, experiencia de usuario unificada son algunos de los muchos aportes de los asistentes virtuales. Al ayudar a los viajeros a encontrar la información adecuada en el momento adecuado, evitar una planificación tediosa y simplificar las reservas, brindan una experiencia fluida y sin complicaciones a los usuarios, lo que ayuda a que sean fieles a la marca.

2.2. PLAN DE MARKETING

2.2.1. Mercado Objetivo

El mercado objetivo que se persigue es, en primer lugar, los turistas, es decir la demanda, y la segunda son las empresas que brindan los servicios de turismo, es decir, la oferta. Con esto no se quiere decir que dicho mercado objetivo sean ambos.

El público al que se quiere dirigir es cualquier persona que quiera aprovechar al máximo su tiempo de ocio. Está dirigido principalmente a personas de 15 a 60 años, ya que se ha considerado que estos son los que están más familiarizados con el uso de su Smartphone para poder utilizar esta función.

Los principales clientes son los turistas. Estos no pagan por obtener la información que se les facilita, sino que ayudan a desarrollar la ventaja competitiva, es decir, la base de datos. Una vez que el cliente finaliza su experiencia se le hace una encuesta de satisfacción, de manera positiva o negativa, para así poder tener una información más rica acerca de los lugares que los turistas preguntan en el s.

La estrategia que se persigue es incentivar el uso del asistente virtual por parte de los turistas, e incluso locales que quieran recibir información acerca de entradas, horarios, etc., siendo estos el imán para poder atraer a las principales empresas que brindan estos servicios de turismo, lo cual puede llegar a ser una gran fuente de ingresos. Esto es así ya que a estas empresas se les puede implantar una comisión para que puedan estar dentro

de nuestro asistente virtual. La ganancia para las empresas es clara, ya que el sistema les serviría como una plataforma de comunicación, publicidad y marketing, siendo retribuidos con una posible venta. Será necesario hacerles ver la nueva forma de aprovechar y optimizar el tiempo de ocio para estos turistas ofreciéndole el servicio que da el asistente virtual, ya que, como se ha comentado anteriormente, con estas conversaciones personalizadas que permiten generar sugerencias hacen que la experiencia sea mucho más rápida y amena. Por otro lado, será necesario fidelizar al target, es decir, a las empresas que ofrecen servicios de turismo, para que puedan seguir teniendo la necesidad de estar promocionándose en dicha plataforma.

2.2.2. Segmentación

Este asistente virtual no está limitado a ciertos segmentos del público objetivo, sino que está dirigido a todas las ramas que pueden surgir de este. Como consecuencia, dicho público se podría dividir en distintas categorías.

En primer lugar, tal como se comentó anteriormente, está dirigido tanto a turistas como a locales, resultando en la primera diferenciación que se podría hacer. Las recomendaciones que se les da a cada uno pueden llegar a ser bastante distintas ya que es más probable que un turista busque información acerca de museos, teatros, etc., mientras que un local puede buscar más información acerca de restaurantes a los que ir o planes de ocio para hacer.

Por otro lado, este mercado objetivo se podría segmentar en distintas categorías. Por un lado, se está dirigido a personas de 15 a 60 años, por lo que se puede diferenciar entre jóvenes, adultos y padres, pudiendo viajar solos o acompañados, es decir, en pareja, en grupo de amigos o en familia. En referencia a esto, las familias lo que tienden a buscar son actividades relacionadas con conocer la ciudad, su gastronomía, es decir, lo que se refiere más a planes culturales y para todas las edades. Para aquellos que viajan en grupos de amigos, suelen buscar más diversión, es decir, lo referido más al ocio. Por otro lado, aquellos que viajan en pareja suelen preferir lo relacionado a la gastronomía y a exposiciones que estén de moda en ese momento. Por último, para aquellos que prefieren viajar en solitario, suelen buscar lo más típico de la ciudad, museos, gastronomía, planes culturales, etc.

Para poder analizarlo a grandes rasgos, es necesario tener en cuenta las 4 P’s de Marketing, es decir: Producto, Precio, Punto de Venta y Promoción.

1. Producto

Lo que se está ofreciendo es un asistente virtual dirigido al sector turístico, con la novedad de la implementación de Transformers. Se ofrece todo tipo de información que un turista puede tener, ofreciendo información acerca de horarios, precio de entradas, información general, etc., haciendo más fácil y ameno la búsqueda ya que se brinda una experiencia fluida y sin complicaciones a los usuarios. Esto está dirigido a cualquier persona (turista o local) que quiera optimizar su tiempo y aprovechar al máximo su experiencia.

Se les satisface con la necesidad de tener un asistente que les pueda ayudar a organizar un viaje en la ciudad, o que ayude a descubrir actividades nuevas en la ciudad en la que se es residente, sin dejar de lado el hecho de que es a medida, ya que con el asistente virtual se almacena y recopila información de dicho usuario, acerca de la satisfacción de estos mismos en experiencias, comentarios, quejas, etc., haciéndose esto de forma gratuita.

Por otro lado, de cara al target que se persigue, es decir, las empresas que brindan los servicios de turismo son los que van a pagar una comisión por utilizar este servicio. A estos se les ofrece una manera de promocionarse ante una gran cantidad de consumidores potenciales, lo cual satisface una necesidad latente en este tipo de negocios.

Es decir, se ofrecen dos cuestiones a dos partes distintas del mercado. El asistente virtual actuaría como un intermediario entre los demandantes (personas) y ofertantes (empresas ofertantes de servicios de turismo), ofreciendo un servicio innovador a aquellas personas que quieran optimizar el tiempo de búsqueda y aprovechar al máximo su experiencia.

2. Precio

El servicio que se ofrece es gratuito, de manera que se atrae a aquellos clientes o consumidores de este asistente y para así poder atraer a los otros clientes, a aquellas empresas mencionadas anteriormente, las cuales pagarían una comisión para poder promocionarse, resultando en una de la fuente de ingresos.

3. Punto de Venta

En este caso no se tiene un punto de venta físico, ya que este servicio se ofrece de forma online a los consumidores y empresas dirigidas al turismo, a través de Telegram.

Al no disponer de punto de venta físico resulta muy importante el marketing y darse a conocer de una manera adecuada, siendo necesario hacer saber al público objetivo esta nueva forma de aprovechar el ocio y optimizar el tiempo de búsqueda de experiencias.

4. Promoción

Cuando se habla de promoción resulta ser uno de los puntos más importantes dentro de los desarrollados anteriormente, con el cual se va a tener que dedicar una gran parte de los recursos.

En primer lugar, es importante darse a conocer a aquellos que van a ser los consumidores de este servicio, es decir, al público. Es necesario que conozcan esta nueva forma que se ofrece para exprimir al máximo las experiencias, siendo muy probable que cuando estos comiencen a utilizarlo las empresas vean que es una gran oportunidad para aumentar sus ventas y así sus ingresos, queriendo entrar en la base de datos.

Hoy en día la mayor plataforma utilizada para publicidad son las redes sociales, pudiendo hacer uso de Facebook Ads, Instagram Ads, Twitter Ads, Youtube Ads, LinkedIn Ads, entre otros. Estos presentan una amplia interfaz publicitaria, donde los se puede disponer d de todo tipo de opciones para poder promocionar y presentar el servicio dirigido al turismo que se ofrece.

Centrando la publicidad en las redes sociales va a permitir incrementar la visibilidad de una manera más rápida, al igual que va a poder llegar a una audiencia potencial mayor y

que estén situados en cualquier parte del mundo. Según statista los usuarios de Facebook en 2021 llegan a 2.700 millones, viéndose aumentado este número cada año.

En lo relacionado a costes, solo se pagará por los clics de los usuarios, siendo el coste de esto muy económico. Pagando sólo por los clics que los usuarios den a los anuncios va a reducir que las inversiones que se hagan sean desperdiciadas. Se puede comenzar a hacer publicidad en redes sociales sin necesidad de un gran presupuesto, siendo una gran opción también para pequeñas y medianas empresas.

2.3. MARCO LEGAL

Como se ha visto hasta el momento el uso de los asistentes conversacionales presentan numerosas aplicaciones prácticas y trae consigo muchos beneficios en la transformación de la comunicación tradicional. No obstante, es necesario tener en cuenta que en este tipo de comunicación uno de los interlocutores es un software, pudiendo, su incorrecta utilización, derivar en responsabilidades legales. Por esta razón, resulta imprescindible analizar brevemente la multitud de disposiciones legales que implicaría el uso de este sistema.

En primer lugar, debe tenerse en cuenta que la finalidad última de un asistente conversacional es proporcionar respuesta a solicitudes, reclamaciones y consultas lo que los lleva a recopilar información de los usuarios para dar las respuestas más óptimas posibles, siendo en muchas ocasiones datos de carácter personal. En estos supuestos entra en juego la privacidad de los usuarios regulada por el Reglamento General de Protección de Datos.

En este sentido, es esencial que se cumplan la normativa. Por ejemplo, asegurarse de que se cuenta con el consentimiento libre, específico, informado e inequívoco del usuario, informar antes de utilizar la información con fines que puedan quebrar la privacidad o asegurarse del correcto tratamiento por parte de terceros. Si bien, es cierto, que en el caso que ocupa esta investigación no se requiere del tratamiento de datos personales de los usuarios sin embargo conviene tenerlo en cuenta para futuras ampliaciones en el desarrollo.

En segundo lugar, cabe recordar que un asistente virtual se desarrolla mediante el entrenamiento de algoritmos de tal manera que sean capaces de comprender y después responder con la misma precisión que un humano, algo que puede ser realizado por un equipo capacitado de manera que se programen las respuestas y pueda controlarse la imagen de marca.

Sin embargo, el modelo también podría entrenarse de forma automática y no supervisada y ello podría llevar a crear conflictos por respuestas con todo ofensivo o intrusivo aprendidas de interacciones anteriores con otros usuarios afectando al derecho al honor, a la intimidad o a la propia imagen. Como ocurrió en 2016 con Tay, el asistente virtual de Microsoft que llegó a proporcionar respuestas racistas.

Por último, otro aspecto a tener en cuenta es el web scraping llevado a cabo con el objeto de recopilar recursos de Internet con los que generar las bases de datos que serán la fuente de información del asistente virtual. Aquí, entra en juego el riesgo de incurrir en la

vulneración de los derechos de propiedad de los titulares de una página web, la posibilidad de ser considerado un hecho de competencia desleal por el uso de información en la comercialización de servicios similares o una violación de los términos legales y condiciones de uso. En el caso objeto de estudio, se ha establecido que con el objeto de minimizar al máximo el riesgo que esto supone se van a descartar todas aquellas empresas que se encuentren fuera de los límites de la comunidad económica europea para evitar diferencias regulatorias. Este último punto solo sería necesario en el caso de que la empresa que desplegase el sistema optase por la extracción de los datos mediante web scrapping. Éste es el caso de la propuesta, motivo por el cual se decide mencionar este punto.

2.4. PLAN FINANCIERO

Tal como se ha comentado, este asistente virtual que implanta la técnica de Transformers se va a utilizar a través de Telegram. El uso de este programa resulta gratuito a diferencia de otros. Existen distintas plataformas que se pueden utilizar en cuanto a costes se refiere.

Por un lado, están los frameworks de IA Conversacional de IBM, cuya plataforma se llama Watson, Microsoft con LUIS y Oracle con Conversation. Las licencias de dichas plataformas tienen un coste elevado y requieren de un desarrollador certificado para realizar el desarrollo. En el caso de querer hacer uso de alguna de estas plataformas es necesario poder disponer de suficiente capital para poder cubrir costes relacionados con la implementación y también de mantenimiento.

A continuación, se encuentran los plug&play, los cuales se caracterizan por tener una cuota mensual baja, cuyos softwares se descargan directamente por internet, sin embargo, estas plataformas son limitadas en cuanto a procesos complejos se refiere. Para ello, se podría utilizar Google Dialog Flow, sin embargo, el coste de mantenimiento resulta muy elevado.

Si por el contrario se quisiese implementar en un futuro este asistente en WhatsApp es necesario tener en cuenta los costes que supondría utilizar dicha plataforma.

En este caso se va a crear un plan financiero para el supuesto que se implante en la API de WhatsApp Business, aunque nuestro proyecto utilice Telegram.

2.4.1. Caso WhatsApp

Siguiendo con el supuesto de querer implantar el asistente virtual en esta línea de mensajería, es necesario conocer el coste que tiene acceder a la API de Whatsapp y el coste en función de unos parámetros ya establecidos por esta compañía de mensajería.

Los costes suelen ser mensuales en función del número de usuarios únicos al mes, y de media suele estar en torno a 0.20 céntimos de euro por usuario distinto al mes.

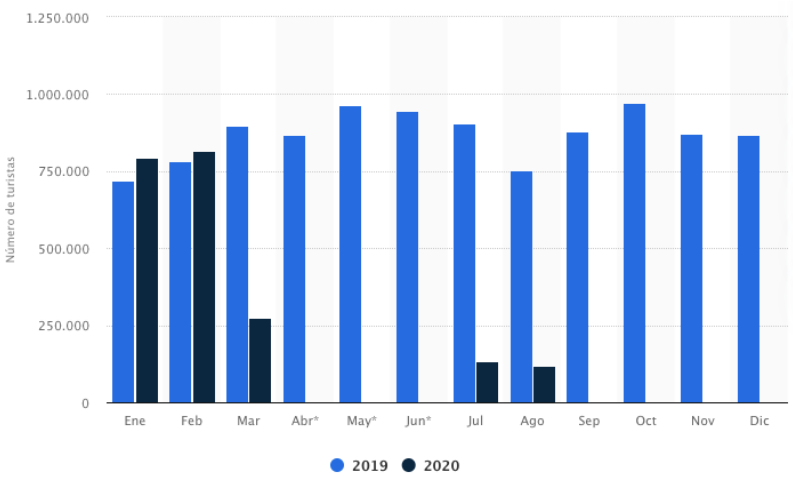
Por otro lado, es importante resaltar que el objetivo principal del presente asistente virtual no es sacar dinero, sino ofrecerlo más bien como un servicio.

Hay distintos costes a los que hay que hacer frente cuando se va a utilizar este servicio.

1. Las tarifas por mensaje: Esta tarifa es de 3 céntimos por plantilla de mensaje. Las plantillas por mensaje son mensajes predeterminados que se utilizan para enviar un mensaje en este servicio a aquellos clientes nuevos o los que no hayan respondido a un chat ya existente en las últimas 24 horas.[1] Es necesario clarificar que los mensajes del cliente son siempre gratuitos. Este coste va a depender de la atención al cliente, es decir, si se le pone al cliente en contacto con la ventana de atención al cliente se dispone únicamente de 24 horas para responder al usuario, si transcurridas esas 24 horas no se le ha contestado, se le pone en contacto directamente con el cliente a través de la plantilla de mensajes y el coste de este servicio es de 3 céntimos por plantilla de mensaje.
2. Por último, se encuentran las tarifas por llamada al mes. Esta tarifa tiene presencia cuando se le deriva al cliente a una llamada de atención al cliente. La tarifa media es de 0.20 céntimos de euro por llamada.

Teniendo en cuenta lo anterior, se va a realizar una estimación con los costes ya definidos.

En primer lugar, es necesario conocer el número medio de turistas que vienen a España, teniendo en cuenta que dependiendo del mes hay más o menos turistas. En este caso, se va a tener en cuenta los datos publicados correspondientes al turismo de 2019, ya que a causa de la pandemia de COVID-19 no se disponen de datos disponibles para los meses de abril, mayo y junio de 2020. A continuación, se puede observar la escasez de datos sobre la cifra mensual de turistas en Madrid en 2020.



Gráfica 2. Cifra mensual de turistas en Madrid 2019-2020. Fuente: Statista

El número medio de turistas que visitaron la ciudad de Madrid en 2019 es de 868.309[2] personas al mes, siendo un total de 10.419.708 de turistas anuales los que visitan Madrid.

Por otro lado, es necesario tener en cuenta a su vez a los locales, ya que este asistente virtual no sólo está destinado a turistas, sino que también a aquellas personas locales que necesiten información o ayuda a la hora de elegir qué hacer en Madrid. Según la revisión

[1] Definición dada por *Callbell* https://callbellsupport.zendesk.com/hc/es/
[2] Media realizada a partir de los datos ofrecidos por Statista.

del Padrón Municipal de Habitantes (PMH) de Madrid, referida al 1 de enero de 2020, muestra una cifra de un total de 3.334.730 habitantes.[3]

Teniendo esto en consideración, si, por ejemplo, bajo el supuesto de que este servicio conversacional sea utilizado por un 15% de la población de Madrid y por un 30% de los turistas que visitan la ciudad de Madrid, se estaría hablando del uso de este mismo por un total aproximado de 500.210 personas locales y 3.125.912 turistas anuales aproximadamente. Sumando ambas cifras, se estaría utilizando por un total de 3.626.122 personas aproximadamente. Por otro lado, tal como se desarrollará cuando se defina el modelo de ingresos, el porcentaje de captura de clientes es del 10%, por lo que se estaría hablando de un total de 362.612 clientes anuales.

Utilizando esta cifra se podría hacer una estimación de los costes que supone utilizar la API de WhatsApp para el asistente:

1. Tarifas por mensaje: Si de las 362.612 personas que utilizan este servicio no se les contesta a las 24 horas se atribuye una tarifa de 3 céntimos por plantilla de mensaje, es decir, habría que pagar un total de 10.878,37 euros, sin embargo, la idea de este asistente conversacional es de contestar inmediatamente, por lo que esa cifra no sería representativa para este caso. Se podría establecer un porcentaje pequeño en caso de que por algún motivo falle el asistente conversacional y se exceda de las 24 horas de las que se dispone para contestar al usuario. En este caso, si por ejemplo se atribuye un 5% a esta cifra se estaría hablando de un total de 21.361 personas a las que por algún motivo el asistente ha fallado y no se les ha podido contestar, lo cual tendría un coste total de 543,92 euros.
2. Tarifas por llamada: Si por ejemplo de ese 5% se destina a una llamada de atención al cliente, el coste sería de 4.351 euros aproximadamente.

Para poder hacer frente a estos costes es necesario establecer cuál va a ser la fuente de ingresos. Teniendo en cuenta el número de clientes que se ha estimado que van a utilizar este asistente conversacional, se puede definir como la primera fuente de ingresos la captura de clientes:

- A partir de los datos otorgados por la encuesta Frontur publicada por el Instituto Nacional de Estadística (INE), se ha calculado la tasa de crecimiento anual de los turistas que visitan Madrid, la cual se sitúa en un 9,4%. Por otro lado, el INE ha publicado que la población inscrita en la Comunidad de Madrid tiene un incremento del 1%. Si se suman ambos incrementos se tiene que hay un incremento anual del 10% de captura de clientes.

Otra fuente de ingresos es la llamada Coste por Click (CPC):

- Esta fuente de ingresos hace referencia al número de visitas que los clientes hagan a hoteles, restaurantes, museos, actividades, etc. Con ello se recibirá una cantidad por cada 1000 clics que reciba cada uno de los servicios mencionados. El precio medio CPC se ha fijado en 7 euros por cada 1000 clics. Dicho valor se ha establecido en base a una comparación realizada a otras empresas y precios del mercado. El CPC irá aumentando anualmente en función del IPC (2%). Cabe destacar que es necesario fijar

[3] Dato obtenido del Padrón Municipal de Habitantes del Ayuntamiento de Madrid.

el número de clics por cliente, el cual se ha fijado en 3 basándose en que el cliente a la hora de buscar información suele buscarla sobre algo en concreto, como por ejemplo el precio de entradas de museos. Si un cliente quiere obtener información sobre un museo y quiere comparar precios con otro se les redirige a los links de estos mismos. Se basa en el hipotético caso de que dicho usuario no suele comparar en más de 3 sitios.

Por último, es necesario tener en cuenta que puede ser usando este asistente conversacional para publicidad:

- Publicidad en el asistente conversacional: Se puede aprovechar que el servicio es a medida del cliente para ofrecer a los negocios la posibilidad de acceder a este público objetivo. Se puede incluir publicidad de distintas empresas como puede ser El Corte Inglés, Booking, El Tenedor, etc. Esto va a hacer posible que se le ofrezca al cliente una publicidad más precisa, al igual que ofertas personalizadas. El modelo de ingresos por esta vía es el que se denomina Coste por Mil (CPM). Con este sistema los negocios pagan una cantidad específica para que aparezcan sus ofertas y publicidad 1000 veces en la plataforma. Se supone que a lo largo del tiempo a través de este canal la demanda por publicarse en él aumentará, por lo que el precio medio por 1000 no solo aumenta al IPC sino que a su vez a una tasa de crecimiento. El sistema de precios se hará a partir de subasta, habiendo un número limitado de anuncios para así no abrumar al cliente. El precio medio por 1000, es decir PM por 1000, se ha fijado en 8 euros por cada 1000 apariciones.

Un modelo de ingresos bajo estas condiciones sería el siguiente:

Modelo de Ingresos					
Turistas + locales Anuales actuales	13.754.438				
	2021	2022	2023	2024	2025
Crecimiento clientes		3,00%	3,20%	3,30%	3,50%
Estimación clientes	3.626.122	3.734.906	3.854.423	3.981.618	4.120.975
% Captura clientes	10%	10%	10%	10%	10%
Numero de clientes	362.612	373.491	385.442	398.162	412.098
Numero clicks / cliente	3	3	3	3	3
Precio medio CPC (x1000)	7 €	7,14 €	7,28 €	7,43 €	7,58 €
Incremento PM CPC (x1000)	2%	2%	2%	2%	2%
Ingresos por CPC	7.614,86 €	8.000,17 €	8.421,30 €	8.873,18 €	9.367,42 €
Precio medio publicidad (x1000)	8 €	8,40 €	8,82 €	9,26 €	9,72 €
Incremento PM publicidad	5%	5%	5%	5%	5%
Numero de publicidades por cliente	2	2	2	2	2
Ingreos por publicidad (CPM)	5.801,80 €	6.274,64 €	6.799,20 €	7.374,75 €	8.014,51 €
Ingresos Anuales Totales	**13.416,65 €**	**14.274,81 €**	**15.220,50 €**	**16.247,94 €**	**17.381,93 €**

Tabla 1. Modelo de Ingresos. Fuente: Elaboración propia

A continuación, se muestra una tabla donde se muestran los ingresos y los costes que se tendrían que hacer frente a lo largo de un año, reflejando como resultado el beneficio total.

Modelo de Ingresos (2021)	
Turistas + locales Anuales actuales	13.754.438
Estimación clientes	3.626.122
% Captura clientes	10%
Numero de clientes	362.612
Numero clicks / cliente	3
Precio medio CPC (x1000)	7 €
Incremento PM CPC (x1000)	2%
Ingresos por CPC	7.614,85 €
Precio medio publicidad (x1000)	8 €
Incremento PM publicidad	5%
Numero de publicidades por cliente	2
Ingresos por publicidad (CPM)	5.801,79 €
Ingresos Anuales Totales	**13.416,64 €**
Modelo de Costes (2021)	
Tarifas por mensaje	0,03 €
Tarifas por llamada al mes	0,02 €
Tarifas por llamada annual	0,24 €
% No respuesta	5%
Coste por mensaje	543,92 €
Coste por llamada	4.351 €
Costes Anuales Totales	**4.895,26 €**
BENEFICIO TOTAL	**8.521,38 €**

Tabla 2. Modelo Ingresos y Costes. Fuente: Elaboración propia.

Como se puede observar, la fuente de ingresos se estima que vaya incrementando a lo largo de los años, pudiendo hacer frente a los costes definidos anteriormente.

CAPÍTULO III: METODOLOGÍA Y PROPUESTA. ESTRUCTURA DEL ASISTENTE Y VENTAJAS PRINCIPALES.

3. METODOLOGÍA Y PROPUESTA

3.1. METODOLOGÍA: CHATBOT DESIGN CANVAS

Durante la realización de este trabajo de investigación y puesta en marcha de un asistente conversacional enriquecido por modelos de Transformers, ha resultado fundamental el uso de una herramienta denominada Canvas. Esta herramienta compuesta por 13 cuestiones permite formar un plan estratégico que ayuda a tener una idea general del proyecto y avanzar más rápidamente durante todo el proceso de desarrollo de un asistente virtual.

A continuación, se irán desarrollando cada uno de estos puntos estratégicos lo que permitirá al lector conocer de forma global los puntos clave de este proyecto.

<table>
<tr><td>Barriers

Incurrir en competencia desleal o vulneración de la propiedad de los dueños de las páginas web por scrapeo de datos.</td><td>Discovery

Instagram Ads
Facebook Ads
Instagram Ads</td><td rowspan="2">Value proposition

Fomentar el turismo en la ciudad de Madrid:

1) Contestar dudas frecuentes
2) Asistencia ininterrumpida 24/7
3) Aprovechamiento máximo del tiempo de viaje con asistencia inmediata
4) El uso de la tecnología de los Transformes ampliará el abanico de información a preguntar por el usuario al no limitarse a lo programado previamente, haciendo la experiencia más rica</td><td>Users

Viajeros de edades comprendidas entre 15 y 60 años.</td><td>Current solutions

Reto: acercarse lo máximo posible al comportamiento y precisión de un humano en las respuestas.</td></tr>
<tr><td>Fallback

En caso de fallos se recopila el historial de conversación para enviarlo a un asistente humano.</td><td>Development and deployment
1) Python como software de programación
2) Librerías de NLP: Spacy y NLTK
3) Librerías de Telegram para su integración
4) Librería Hugginface para modelos de Transformers</td><td>Devices/Modalities

Smartphones y ordenadores.</td><td>Channels

Actualmente, este asistente es ofrecido en Telegram. En un futuro se implementará en plataformas más populares como WhatsApp y aplicación Web.</td></tr>
<tr><td>Background tasks

Por un lado, para dar la información solicitada por el usuario el asistente requiere del texto procesado en formato JSON sobre cada uno de los lugares turísticos incluidos, previo scrapeo web.

Por otro lado, la extracción y devolución de la respuesta exacta se consigue por medio de modelos de Transformer, en concreto, distill BETO, BETO, Electra y RuPERTa,</td><td>Relationship

Se trata de una relación de tipo esporádico. El asistente no establece una relación con el usuario, es decir, la conversación no es recordada por el sistema.

En un futuro se plantea la posibilidad de almacenar la información del usuario para una relación más cercana.</td><td>Personality

Elsa

Carismática, amable y positiva, con un estilo informal.</td><td colspan="2">Conversational tasks

Además de saludos, agradecimientos y explicación de su función. El objetivo conversacional es proporcionar al usuario toda la información que necesite sobre una atracción turística/ocio de Madrid.
1) Conocer la intención turística del usuario
2) Extraer y enviar la respuesta correcta al usuario a partir de la documentación de restaurantes, museos, monumentos y landmarks</td></tr>
</table>

Ilustración 11. Chatbot Design Canvas. Fuente: elaboración propia

1. **Value proposition:** cual es el propósito del asistente virtual y que valor aporta al cliente.

La razón de ser de este asistente conversacional es fomentar el turismo en la ciudad de Madrid pues es un sector de actividad que ha sido gravemente afectado durante el último año debido a los efectos del COVID-19.

A pesar de que son muchas las empresas turísticas que ofrecen actualmente soluciones de este tipo para facilitar las estancias de sus clientes se ha detectado una clara posibilidad de mejorar y explotar aún más las capacidades de un asistente virtual en este sector. Por esta razón, se ha decidido poner a prueba la tecnología de los Transformers en este contexto, de tal manera que pueda sacarse el mayor provecho al Procesamiento del Lenguaje Natural superando las barreras que aparecían en la configuración tradicional de un asistente conversacional, entre las que están la identificación de idiomas, la forma de expresarse a la hora de hacer una pregunta o los problemas ortográficos.

En cuanto al valor en sí que este asistente virtual aportará al cliente puede resumirse en dos bloques. Por un lado, el ahorro y aprovechamiento del tiempo disponible al máximo al dar respuestas inmediatas a las preguntas del viajero, evitando que este tenga que estar explorando decenas de páginas web para encontrar información. Por otro lado, la asistencia ininterrumpida 24/7 proporcionará un alto valor añadido al usuario que podrá sentirse en todo momento atendido.

2. **Users:** identificar para que grupo de personas está pensado el asistente virtual.

Principalmente, enfocado a viajeros de edades comprendidas entre 15 y 60 años. Aunque puede ser utilizado por personas de todas las edades se ha decidido enfocar en este rango de edad pues se considera que son los más familiarizados con herramientas tecnológicas.

3. **Current solutions:** que características de las soluciones actuales son irremplazables con el asistente virtual.

Hay ciertas características de las soluciones actuales alternativas a un asistente virtual que no pueden superarse como el cara a cara en una agencia de viajes o en un punto de información. Sin embargo, con el rápido avance de las tecnologías y el alto dominio que hoy en día los jóvenes tienen sobre Internet y las plataformas de comunicación online esta cercanía con el cliente puede acabar siendo reemplazada.

Además, con la incorporación de la tecnología de los Transformers en el asistente, punto principal de este proyecto, se conseguirá una aproximación mucho más precisa entre las preguntas y las respuestas permitiendo una simulación de conversación prácticamente humana.

4. **Devices and modalities:** a través de que dispositivos se va acceso del asistente al target.

La modalidad de uso de este asistente virtual optimizado mediante Transformers es a través de dispositivos móviles pues es la modalidad más cómoda teniendo en cuenta el sector de actividad en el que se sitúa este proyecto. Al tratarse de turismo los clientes estarán la gran mayoría de tiempo en la calle y moviéndose de un lugar a otro.

5. **Channels:** plataforma de mensajería.

La plataforma en la que será distribuido este servicio será Telegram, aplicación de mensajería instantánea alojada en la nube que permite a los usuarios enviar mensajes de chat y realizar llamadas de audio y vídeo. El usuario simplemente deberá tener descargada esta aplicación en su móvil y buscar el bot Elsa. Cabe mencionar que de cara a un futuro cercano se pretende implantar este asistente en plataformas más populares como WhatsApp e incluso la propia web.

6. **Conversational tasks:** identificar según el tema elegido las preguntas que puede hacer el usuario y los problemas que querrá resolver con el asistente virtual.

En primer lugar, el asistente virtual incluirá aspectos básicos de una conversación como son los saludos, despedidas o agradecimientos. Además, de una presentación y explicación de su función.

En segundo lugar, el asistente preguntará por la intención del usuario, es decir, a dónde quiere ir o de qué lugar quiere conocer información para poder devolverle la información deseada. Para ello el asistente en un inicio pondrá a disposición distintas opciones de lugares de tipo turístico y ocio de la ciudad de Madrid. mediante el uso de botones, entre las opciones están: museos, monumentos, restaurantes o landmarks (calles, plazas, parques). Estas localizaciones adoptarán la función de entidades, es decir, de expresiones claves que será detectadas por el asistente virtual y a partir de las cuales se activarán las acciones correspondientes.

Una vez seleccionada la categoría se preguntará por el lugar exacto del que se quiere obtener información y a partir de ese momento el usuario podrá realizar distintas preguntas del tipo: direcciones, precios, horarios, cómo llegar a un sitio u otra información concreta de cada lugar.

7. **Personality:** configurar una personalidad al asistente virtual.

El asistente virtual cobra personalidad propia con nombre Elsa y un estilo informal y amable durante la conversación.

8. **Relationship:** pensar si el asistente virtual va a establecer una relación con el usuario o será útil de forma esporádica.

Actualmente, el asistente virtual no establece ningún tipo de relación con el usuario, es decir, cumple una función esporádica. El usuario abre el asistente, pregunta la información que necesita y la conversación queda registrada pudiendo el usuario leerla en cualquier momento. Sin embargo, el sistema no almacena dicha información para utilizarla en futuras conversaciones.

Para un futuro desarrollo del asistente si se plantea el establecimiento de una relación con el usuario durante su estancia en la ciudad de tal manera que cada día pueda saludar al usuario por su nombre o sepa donde estuvo el día anterior para preguntarle qué tal su experiencia. Una vez finalizada la estancia el asistente virtual podría olvidar toda esa información.

9. **Background tasks:** el asistente virtual no solo sirve para responder preguntas sino también para realizar tareas.

En este caso, el asistente virtual tiene como objetivo proporcionar información al usuario, pero no mediante respuestas prediseñadas sino mediante la detección de la entidad, la búsqueda entre los documentos incorporados y recopilados mediante web scrapping, y la recuperación de la respuesta exacta mediante encodings posicionales todo ello mediante modelos de Transformers.

A parte de esta forma de respuesta incorporada mediante los Transformers, y que hace diferenciar este asistente de los generales del mercado, el asistente virtual podría incorporar funciones como realizar una reserva en un restaurante, reenviar al usuario a una web o comprar una entrada.

10. **Fallback:** si el asistente virtual falla cómo se va a resolver.

En caso de que el asistente falle por alguna pregunta inesperada puede pasarse el historial de información a un agente o simplemente responder educadamente diciendo que no se tiene conocimiento para dar esa información y preguntar si puede ayudarle con otra consulta. El sistema usa excepciones para mantener al cliente informado, y que sepa en todo momento lo que está sucediendo, y evitar que los errores bloqueen el asistente.

11. **Development:** cómo se va a crear el asistente virtual, es decir, plataforma, librerías y código empleado para su desarrollo.

El software de programación empleado para la implementación de este asistente ha sido Python, lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. En cuanto a los datos, tomados como fuente de las respuestas, serán recopilados mediante scrapeado web y estructurados y almacenados en formato json siguiendo el esquema clave-valor.

Las librerías utilizadas serán principalmente las relativas a Procesamiento de Lenguaje Natural como Spacy o NLTK, librerías de Telegram para su integración y de Transformers para la aplicación de los modelos que evite la programación de las

respuestas y permita la búsqueda de respuesta en el texto, en concreto, la librería Hugginface de la cual se han empleado los cuatro modelos de español disponibles, distill BETO, BETO, Electra y RuPERTa,

Por último, cabe añadir la inclusión de la plataforma Docker en todo este proceso con la intención de que en caso de una futura incorporación del asistente en páginas web de empresas turísticas se solventen los problemas que generan los conflictos entre dependencias, versiones de librerías instaladas en el sistema y entre sistemas operativos, como se detalla más adelante.

12. Barriers: barreras o limitaciones que pueden encontrarse en el desarrollo.

Principalmente, este asistente podría verse limitado en su desarrollo por causas de tipo legal. En primer lugar, será necesario prestar especial atención a la recopilación de datos de los usuarios donde en todo momento deberá respetarse lo regulado en el Reglamento General de Protección de datos para evitar problemas en el tratamiento de datos de carácter personal.

En segundo lugar, deberá controlarse el entrenamiento no supervisado del asistente virtual para evitar un estilo o tono de respuesta inapropiado fruto de lo aprendido en anteriores conversaciones y recopilado en la base de datos del sistema. Por último, prestarse elevada diligencia en el scrapeo de datos de páginas web con los que recopilar información para otorgar respuesta al usuario de forma que se evite incurrir competencia desleal o vulneración de la propiedad de los dueños de las páginas web.

13. Discovery (marketing): como el asistente virtual y sus servicios serán descubiertos por el target

Teniendo en cuenta el contexto tecnológico y la revolución que las redes sociales están suponiendo en la publicidad se ha considerado como opción más recomendable para dar a conocer este servicio el marketing online. En concreto, publicaciones en Instagram, Facebook y Twitter serán fundamentales para dar a conocer al bot pues permitirán llegar de forma masiva e internacional al público.

Además, también podría ser muy efectivo involucrar a los propios usuarios en las etapas finales de su diseño pues las personas suelen responden de manera más enfática a aquellas cosas en las que participan. Por ejemplo, colaborando en la asignación de un nombre o en el diseño de la interfaz.

3.2. NUESTRA PROPUESTA

Como se viene comentando, el desarrollo de asistentes virtuales, esto es, de máquinas con la capacidad de mantener una conversación fluida con un ser humano no es algo nuevo. El desarrollo de este campo comienza a mediados de los años 60, con ELIZA, y desde entonces, no ha parado de crecer. El uso de dichos asistentes puede mejorar la productividad en ciertas tareas mecánicas, como la gestión de consultas telefónicas, o la guía turística. Algunas de las ventajas que ofrecen estos sistemas son las siguientes:

- Coste: suponen una reducción en los costes empresariales, puesto que el uso de estos asistentes permite la reducción de la plantilla de empleados que antes ocupaban esta labor, reduciendo los costes salariales.
- Disponibilidad: los asistentes pueden estar disponibles de manera prácticamente ininterrumpida.
- Fiabilidad: al no tener emociones, sentimientos o cansancio, el comportamiento es siempre el mismo, independientemente de las horas que lleve activo, el día de la semana que sea, o la situación económica del país.

La manera tradicional en la que se han venido elaborando los asistentes podría describirse con el siguiente código:

```
# Entidades genéricas
saludos = ['Hola', 'Buenos días', 'Buenas']
despedidas = ['Hasta pronto', 'Adiós', 'Hasta luego']
# Contexto de la conversación
solicitar_credito = ['Quiero un crédito', 'Necesito un préstamo']
solicitar_seguro = ['Vengo a hacer  un seguro', 'Necesito un seguro']

# Asistente
user_input = input()  # lo que dice el cliente

if user_input in saludos:
  return 'Buenos días, ¿en qué puedo servirle?'

elif user_input in despedidas:
  return 'Espero haberle ayudado, hasta otra'

elif user_input in solicitar_credito:
  return 'De acuerdo, te haré unas preguntas'

elif user_input in solicitar_seguro:
  return 'Vale, ¿qué tipo de seguro quiere?'
else:
  return 'Lo siento, no le he entendido, ¿podría repetirlo?'
```

Este sistema presenta múltiples debilidades.

En primer lugar, se basa en que lo que el usuario dirá, estará previamente definido dentro de algún objeto, como los saludos o la solicitud de seguro. Esto implica que, a la hora de programar el asistente, haya que incluir todas las formas posibles en que un cliente pueda solicitar algo. El riesgo está en que, naturalmente, existen infinidad de maneras de preguntar una misma cosa. Todas ellas varían en función de la persona, cultura, edad, nivel educativo o situación. Es prácticamente imposible conseguir cubrir todos los casos, y cubrir una amplia gama implica programar a mano una cantidad ingente de código.

En segundo lugar, es altamente sensible a la ortografía. Existen métodos para paliar este problema, como puede ser el fuzzy matching (añadir pie de página) o la conversación a través de un sistema de botones, que limite los inputs del cliente. Nuevamente, esto requiere programación manual adicional.

En tercer lugar, el idioma. Para un asistente cuyos clientes hablan distintos idiomas, por ejemplo, inglés y español, esto representa un impedimento adicional. Las soluciones al problema podrían ser:

a) **Introducir una capa de traducción al lenguaje del asistente:** problemático, puesto que añade dependencia con servicios externos, y puede ralentizar el sistema. Además, según qué idioma, el resultado puede ser pobre.
b) **Tener un asistente en cada idioma:** implica repetir el mismo código cambiando el idioma tantas veces como idiomas se espere que reciba el asistente.

En cuarto lugar, con los sistemas tradicionales es muy complejo conseguir una respuesta corta y breve. Estos sistemas se basan en reglas lógicas, y devuelven trozos de texto grandes, lo que empeora la experiencia del cliente, quien quiere una respuesta breve y concisa.

Finalmente, la escalabilidad. ¿Qué sucede si es necesario que el mismo asistente resuelva problemas específicos de cada cliente, para una multitud de clientes? ¿Y si el problema es similar pero solo cambia el contexto? Para el caso de uso que nos ocupa, la asistencia es muy similar entre las distintas actividades turísticas. Las preguntas que recibe no varían mucho al cambiar el destino de la visita. Sin embargo, todas ellas requieren ser reprogramadas si se utiliza el enfoque tradicional. Una vez más, esto añade complejidad al sistema, puesto que habría que pensar qué partes han de ser constantes y cuales paramétricas o dinámicas.

Hay que tener en cuenta que todos estos inconvenientes no afectan solo a la fase de creación del asistente, afectan a todo su ciclo de vida. Más código y conexiones implica un mayor coste de mantenimiento, una mayor dificultad para localizar y solventar las anomalías, y un mayor riesgo general para el sistema, puesto que depende de un mayor número de piezas individuales susceptibles de error.

3.2.1. Uso del Transformer

Como se ha venido comentando a lo largo del trabajo, los Transformers son una tecnología nueva, con aproximadamente 4 años de edad. En pocas palabras, podrían describirse como modelos similares a las redes neuronales cuya finalidad es trabajar con texto. Éstos pueden llevar a cabo múltiples tareas de NLP, entre ellas, responder preguntas. Han sido entrenados de manera previa con una enorme gama de ejemplos conversacionales, llegando a los 150.000, en algunos casos. Si el caso de uso es muy específico, también es posible reentrenarlos.

La forma en que estos modelos son capaces de generar valor es similar a la de cualquier modelo de Machine Learning: no necesitan programación explícita. Éstos son de gran ayuda en aquellas situaciones en que existen demasiados escenarios posibles para ser predefinidos a mano. Este es el caso de una conversación humana.

Más concretamente, la aplicación de dichos modelos requiere de aproximadamente 5 líneas de código. Permiten responder preguntas de manera rápida, con alta precisión y de manera concisa, pese al poco esfuerzo que requiere su puesta en uso. A modo de ejemplo, la aplicación de este TFM se compone de 300 líneas de código, de las cuales 290 corresponden al asistente virtual, y 10 al Transformer, que es el gestor final de la conversación.

El modelo se compone de dos únicas piezas: el contexto, que es el texto sobre el que se va a preguntar, y las preguntas. Dadas estas características, es posible gestionar distintas conversaciones con facilidad, simplemente cambiando el parámetro del contexto, y dejando al modelo hacer el resto del trabajo.

Así, la programación explícita y el asistente virtual pasan a jugar un papel auxiliar, secundario, cuyo único objetivo es determinar el contexto. En el caso de uso que nos ocupa, determinar a dónde quiere ir el turista, para así poder cargar el contexto de ese lugar, y dejar al modelo hacer el resto del trabajo. Así, la empresa ahorra recursos, y puede invertirlos en mejorar todavía más la experiencia del cliente.

```
{
  # ---------------------------------------------------------
  # Código tradicional, para averiguar que quiere el cliente |
  # y cargar el contexto                                      |
  # ---------------------------------------------------------
}

user_input = input()  # lo que dice el cliente

modelo_transformer(
  {'context': texto,  # los datos sobre los que se va a preguntar
   'question': user_input}

  )['answer']  # Devuelve distintos outputs, como la probabilidad
               # que asigna a su respuesta. En este caso lo que
               # interesa es la respuesta.
```

Las ventajas de este enfoque son numerosas, pero las más directas son:

1. El modelo es capaz de devolver respuestas concretas y sencillas.
2. No es necesario programar la conversación de manera explícita.
3. El sistema es robusto a faltas de ortografía y estilos de preguntas.
4. En muchos casos, no dependen del idioma, pero existen modelos específicos.
5. Escalabilidad, se puede usar en distintos casos, cambiando el contexto.
6. Mejora de los modelos con la actualización de los mismos, nuevas versiones.
7. Posibilidad de entrenar el modelo para un caso de uso específico.

CAPÍTULO IV: DESARROLLO TÉCNICO DEL PROYECTO.

4. DESARROLLO TÉCNICO

4.1. FUNCIONAMIENTO

El sistema tres componentes claramente diferenciados: asistente, datos y transformer. El siguiente esquema resume la estructura de dichos componentes, y la relación entre ellos.

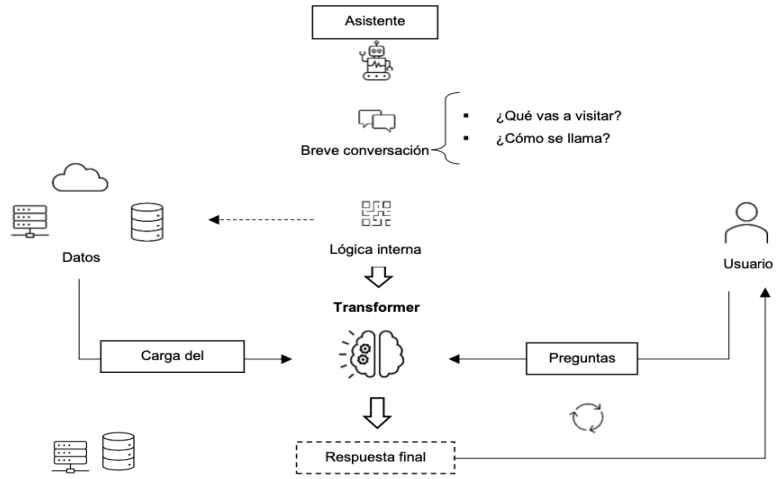


Ilustración 12. Funcionamiento. Fuente: Elaboración propia

4.1.1. El asistente

Es el primer contacto con el usuario. Saluda, explica su función y comienza las preguntas. Su función es auxiliar, puesto que el valor aquí lo genera el uso del Transformer. El asistente no presenta grandes diferencias respecto a los asistentes habituales.

Lo primero que pregunta es qué tiene pensado ver el usuario, a dónde va a ir. Por simplicidad, se ofrece en pantalla un teclado con distintas opciones. El objetivo en este punto es saber el de sitio al que va a ir el usuario, la "categoría". Las opciones son:

- Museos
- Landmarks
- Restaurantes
- Monumentos

Una vez que ya se conoce el tipo de sitio al que irá el usuario, se le pide que escriba el nombre concreto del lugar. Se pretende construir una estructura key value, donde la key es la categoría del sitio, por ejemplo, "museo", y el value es el lugar concreto, por ejemplo "museo del prado". Este formato es el que se considera conveniente para el presente caso de uso, puesto que permite un mayor orden y control de los datos, que están organizados por carpetas. No obstante, no es la única manera de hacerlo.

4.1.2. Los datos

Los datos se almacenan en ficheros de formato json[4]. Se organizan dentro del directorio datos siguiendo una estructura key value, idéntica a la descrita en el punto anterior. De este modo, es posible generar las rutas hacia los datos utilizando el siguiente procedimiento:

```
# Preguntas iniciales
respuesta_1 = 'landmarks'
respuesta_2 = 'Catedral de la Almudena'
```

[4] JSON son las siglas de JavaScript Object Notation. Es un formato de texto sencillo de tipo *key-value*, similar a los diccionarios de Python.

```
# Paso intermedio para confirmar el destino
destino_usuario = {respuesta_1: respuesta_2}

# Generación de la ruta
categoria = destino_usuario.keys()
lugar = categoria = destino_usuario.values()

ruta_hacia_fichero = f'data/{categoria}/{lugar}.json'
```

Para minimizar el riesgo, se añade una capa previa de tratado del mensaje, donde los espacios se sustituyen por "_", se eliminan los acentos y la capitalización. Después, se verifica si existe el archivo en la base de datos. Se utilizan excepciones en el código para evitar problemas en este paso, y que el sistema no quede vulnerable ante la búsqueda del archivo. Si no se localiza, se comunica al usuario, y se le ofrece la posibilidad de volver a intentarlo.

Una vez que se ha localizado el archivo, se crea una variable global con él, haciendo que esté disponible en cualquier parte del entorno. Es importante mencionar que, pese a estar almacenado en formato json, el texto se trasforma a formato plano antes de llegar al modelo, quien recibe los datos como un solo bloque de texto. El formato json se emplea por ser más ordenado a la hora de estructurar medianamente los datos procedentes de internet.

El almacenamiento de datos se ha planteado como local, puesto que lo que se está mostrando es un ejemplo, y no es necesario un espacio grande. No obstante, otra buena opción sería hacer uso de bases de datos documentales. Una alternativa interesante sería mongoDB con mongoDB Atlas. Este sistema ofrece hasta 5GB de almacenamiento gratuito en clúster.

4.1.3. El Transformer

Tras este último punto, la programación ya no es explícita. No existen preguntas ni respuestas programadas. La conversación se gestiona a través del modelo.

No obstante, antes de dar entrada al Trasformer, se da la posibilidad de elegir el modelo que se quiere emplear, de entre 3 posibilidades: el modelo estándar (BETO destilado), BETO (BERT entrenado sobre texto español) y Electra. En primeras versiones del sistema también se incluía Ruperta, pero se descartó por no ofrecer buenos resultados para nuestro caso.

Este punto no se incluiría en una aplicación real, donde el modelo puesto en producción sería el que mayor valor generase. Sin embargo, debido al carácter experimental de esta investigación, decidimos incluir esta opción, con el fin de poder probar distintos modelos fácilmente, y comparar el comportamiento de cada uno de ellos.

Las características de cada modelo se incluyen a continuación.

Modelo	Entramiento	Épocas	Tasa de aprendizaje	Batch Size
Normal	130k	5	3 x 10^-5	12
BETO	130k	2	3 x 10^-5	12
Electra	130k	10	3 x 10^-5	16

Tabla 6. Características del modelo. Fuente: Elaboración propia.

Los tres han sido entrenados con una Tesla P100 GPU y 25GB de RAM. Toda la información está disponible en la página de Huggingface, en el apartado "modelos", filtrando por modelos en español y question answering.

Una vez alcanzada esta fase de la conversación, el flujo de diálogo permanece aquí, permitiendo que el usuario realice tantas preguntas como desee al modelo, hasta que decida terminar la conversación usando el comando /stop. En las siguientes secciones se incluye un ejemplo de conversación, explicando todo el proceso mediante capturas de pantalla.

4.2. INTEGRACIÓN Y ACCESO: TELEGRAM

El sistema se despliega en forma de bot de Telegram. Para ello, se utiliza la API de Telegram, a través de su paquete en Python, que actúa a modo de wrapper. La decisión de la implementación en Telegram se debe a que esta plataforma cuenta con su propio sistema de gestión de bots, por lo que se asegura la legitimidad del proceso, y se facilitan algunas gestiones.

El asistente se crea desde el usuario BotFather, que es a su vez un bot, cuya labor es dar soporte y ayudar a gestionar el resto de bots que conviven en Telegram.

La conexión se hace mediante un token secreto y único, que permite controlar la conversación desde un servicio externo, en este caso, desde un contenedor Docker, donde se alojará el servicio.

Otra de las ventajas de optar por Telegram es la posibilidad de incluir comandos en el asistente. Los comandos son acciones específicas, que se disparan utilizando el carácter "/" seguido de la palabra del comando. Permiten llevar a cabo determinadas acciones dentro de la conversación.

Se han incluido los siguientes comandos:

- /start: inicia la conversación
- /help: muestra ayuda sobre el asistente, los comandos disponibles
- /destinos: muestra la lista de destinos disponibles
- /cambio_destino: permite cambiar el destino al que el usuario quiere ir
- /cambio_modelo: para cambiar de modelo y facilitar los experimentos, no se incluiría en una versión comercial
- /stop: termina la conversación

El sistema es gratuito y constituye una opción perfecta para este TFM. El inconveniente que presenta es que el asistente solo puede mantener una conversación a la vez. Si se intenta ejecutar el servicio simultáneamente en dos sistemas diferentes, devuelve un error. Desde el punto de vista en una aplicación real, esto sería algo a solventar, puesto que cabría esperar que el asistente atendiese a múltiples peticiones a la vez. No obstante, para la demostración del uso de Transfomers en asistentes virtuales, es más que suficiente.

Telegram permite múltiples opciones, de cara al asistente, si se quiere complicar el sistema. Es posible pedir contendido al usuario, como imágenes o su ubicación. También permite personalizar la apariencia del asistente, cambiando su nombre, foto de perfil y descripción.

Finalmente, existe la posibilidad de añadir el asistente a un grupo de Telegram, lo cual según el negocio podría constituir alguna ventaja. Para nuestro ejemplo, no ha sido el caso.

En síntesis, Telegram constituye una buena opción para el proyecto, pero en fases posteriores, el sistema se implementaría en otras aplicaciones más populares, como es el caso de WhatsApp, o directamente en páginas web.

4.3. DESARROLLO DEL SISTEMA

El desarrollo del sistema se hace al completo utilizando código Python. En un primer lugar, se valoró y testó la opción del uso de una herramienta de mercado, concretamente, Watson Assistant, de IBM. Sin embargo, por una serie de motivos, se decide optar por Python. Algunos de los motivos que justifican la decisión son:

1. **Integración con los modelos:** la existencia de librerías de Python con modelos de Transformers hace que sea relativamente fácil integrarlos en el asistente.
2. **Extracción de datos:** mediante web scraping[5]. Los scripts para conseguir el texto se desarrollan también en Python.
3. **Control del sistema:** un mayor control sobre toda la infraestructura, derivado de mantener las tres fases en el mismo lenguaje (asistente, datos y transformer)
4. **Portabilidad:** posibilidad de crear un contenedor de Docker que incluye el servicio y todas las librerías necesarias, evitando problemas de compatibilidad entre versiones, sistemas operativos, y facilitando así su puesta en producción.

A continuación, se presenta un ejemplo de las fases del sistema, y de lo que sería una conversación completa.

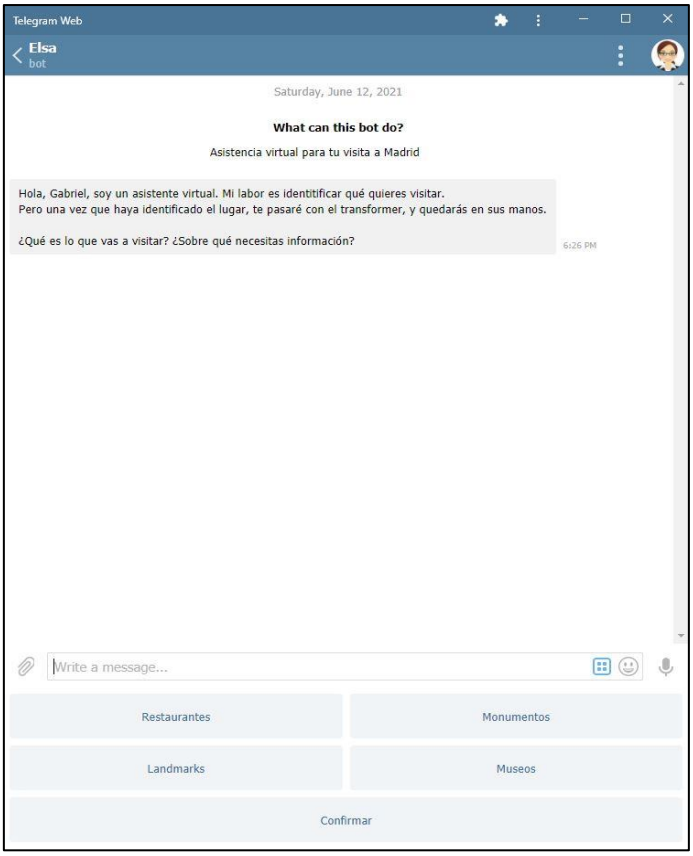
[5] Se conoce como *web scraping* al uso de herramientas digitales y de programación para automatizar la extracción y el parseado de información procedente de la web. Lo utilizamos para poder extraer los textos con comodidad y eficiencia, y no tener que copiarlos a mano.

4.3.1. Fases del sistema

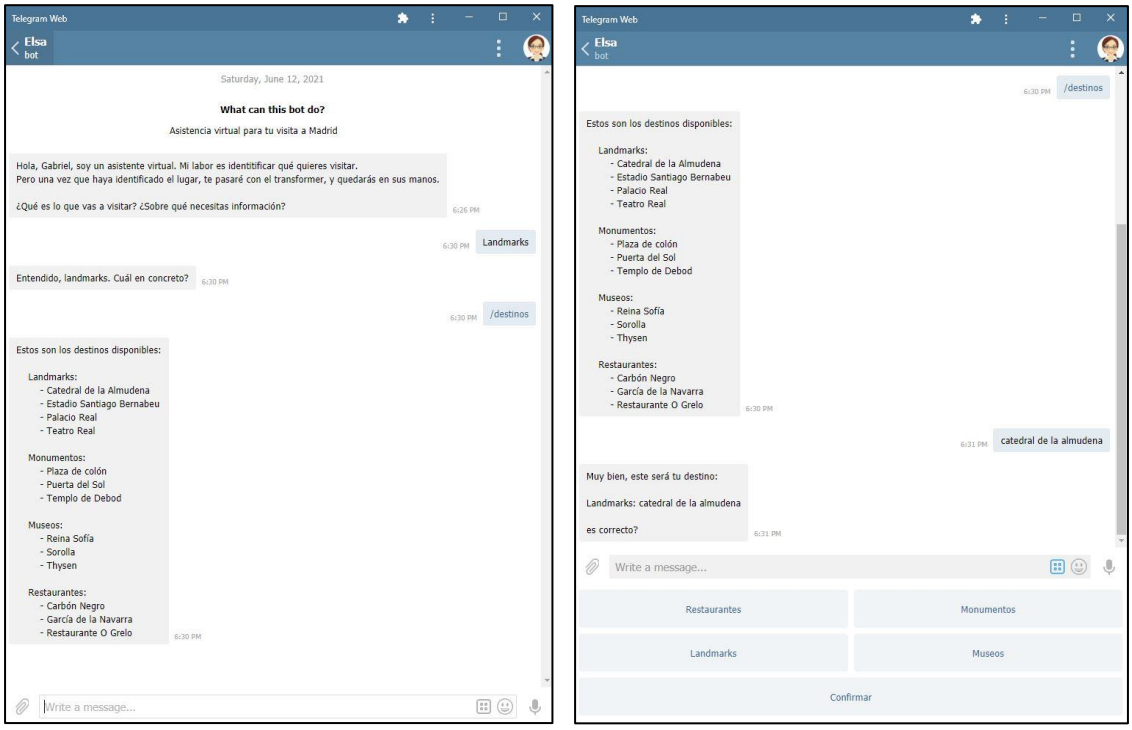
Se compone de 5 fases o nodos a través de los cuales va fluyendo el proceso a medida que el usuario interactúa

La primera de ellas consiste en saludar al usuario, y preguntarle qué va a visitar. El usuario recibe un tablero con 4 opciones para seleccionar la categoría.

Después de haberla seleccionado, se le pedirá que escriba el nombre del destino que quiere visitar. Es posible mostrar a lista de destinos disponibles con el comando "destinos", que se activa escribiendo "/destinos"



Captura 1. Muestra asistente en Telegram



Captura 2. Muestra asistente en Telegram

Tras haber escrito el destino, se pedirá la confirmación del usuario. En este momento, lo que está sucediendo por dentro, en Python, es un "try – except". Se ha construido la ruta hacia el archivo que contiene los datos del destino, y se está comprobando si existe. De lo contrario, la excepción envía un mensaje al usuario informando de que no ha encontrado lo que busca, y le invita a introducirlo de nuevo. El uso de la excepción es importante, puesto que, si no se usase, el error haría caer todo el sistema, y el turista no sabría qué está pasando.

Para mejorar la experiencia y minimizar los errores, el sistema es robusto a capitalización y acentuación. Aunque el usuario escriba "PlAZA de CóLON", el sistema será capaz de encontrar los datos. No obstante, convine mencionar que el sistema es sensible a errores, puesto que esta parte todavía corresponde al asistente, no al modelo.

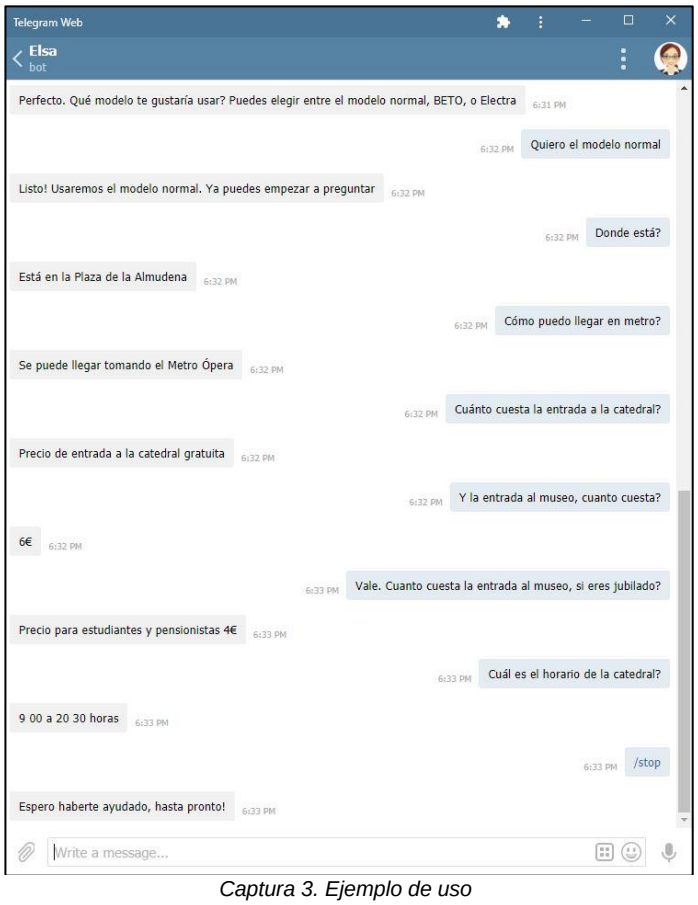
En un primer momento, se planteó e implementó el fuzzy matching, o "matcheo difuso". Bajo este enfoque, el sistema no busca coincidencias exactas, sino que asigna una puntuación a cada palabra de tener "match" con otra. Así, si el usuario escribe "plza de clon", el destino con una mayor puntuación será "plaza de colon", y será seleccionado.

Sin embargo, este enfoque se descartó. El motivo es que, dada una lista de destinos:

- Plaza de Santa Ana
- Prado
- Parque Retiro

y dada una petición "plaza de santa maría", el sistema seleccionaría "Plaza de Santa Ana". Es decir, siempre devuelve el destino cuya "distancia" en términos semánticos es la menor con respecto a la petición del usuario. Esto generaría problemas, puesto que se estaría cargando un texto totalmente ajeno al destino real del usuario. Este es uno de los problemas a los que son sensibles los asistentes virtuales tradicionales. Los Transformers, como ya se ha comentado, no presentan este inconveniente.

Finalmente, una vez que se han cargado los datos correctamente, se pasa al Transformer, donde la programación ya no es explícita. Dado que este TFM tiene un carácter científico y experimental, se ha incluido una capa más que permite seleccionar entre tres tipos de modelos distintos. Esto no se incluiría en la aplicación real del negocio.

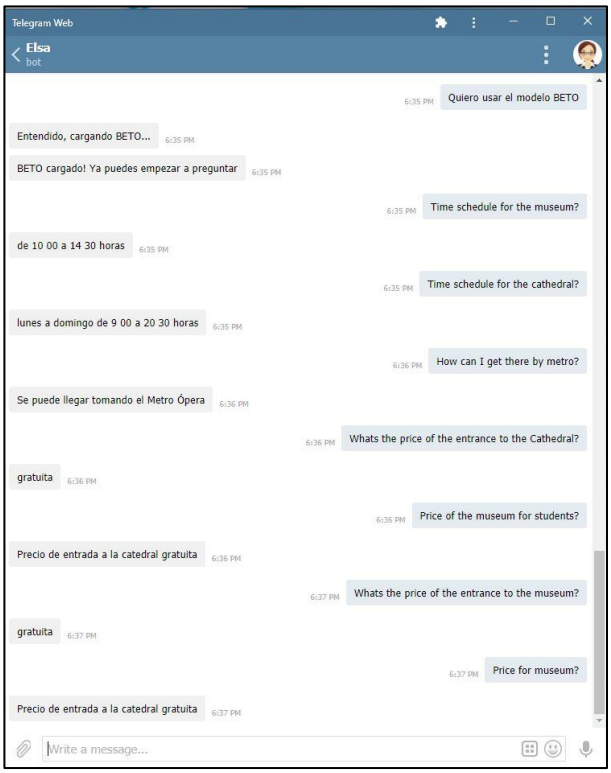


Captura 3. Ejemplo de uso

Se puede observar cómo funciona el sistema de atención, haciendo que el modelo ponga el foco en partes concretas del texto. Esto hace que el Trasnformer sea capaz de distinguir entre cambios sutiles en el texto, como el horario de la catedral y el horario del museo, el precio de cada lugar, e incluso el precio para distintas personas. Además, las respuestas son concretas y precisas, no fragmentos de texto largos.

Esta fase de la conversación es cíclica. El sistema permanece aquí para que el usuario pueda hacer tantas preguntas como desee, y la conversación no termina hasta que se usa el comando /stop. El diálogo está gestionado al completo por el modelo Transformer. Pese a estar trabajando con texto en español, el modelo es capaz de gestionar las preguntas en inglés sin problema. El modelo también es capaz de gestionar la conversación en inglés.

Al conversar en inglés, el modelo comete algunos fallos. La prueba en inglés se ha hecho utilizando el modelo BETO, que ofrecía mejores resultados. No obstante, cualquiera de los tres modelos es capaz de mantener la conversación recibiendo las preguntas en inglés.



Captura 4. Ejemplo de uso en inglés

Pese a tener algunos fallos, los resultados son verdaderamente prometedores. Los modelos han sido entrenados al 100% en español, no saben inglés. Y el texto con el que están trabajando está en español. No existe ninguna capa de traducción intermedia, que pase las preguntas a español, y después al modelo. Recibe las preguntas directamente en inglés.

De cara al usuario, una alternativa interesante sería introducir otro modelo, en las salidas del primer Transfomrer, que tradujese el texto al idioma del cliente. Podría usarse TextBlob, librería de NLP que permite traducir texto. Otra opción sería utilizar un Transformer adicional, encargado de traducir el texto.

4.4. PORTABILIDAD, SEGURIDAD Y DESPLIEGUE

Uno de los inconvenientes más habituales que puede surgir al trabajar con lenguajes de código abierto es el conflicto entre librerías y dependencias. Las librerías, que son colecciones de código, reciben actualizaciones frecuentemente. Normalmente, las librerías no suelen funcionar por sí mismas, sino que requieren a su vez de otras librerías. Este fenómeno se conoce como "dependencias". Estos dos fenómenos, las actualizaciones y las dependencias, suelen causar dos tipos de problemas.

4.4.1. Conflictos entre dependencias

Supóngase el siguiente caso. El código empleado en este proyecto utiliza la librería Transformers. Ésta, a su vez, necesita la librería pandas, en versión 3.5.0. Hasta aquí no hay ningún inconveniente.

Tras terminar el proyecto, se decide venderlo a una agencia de viajes, a la que le ha gustado la idea. Resulta que esta agencia tiene en su ordenador la librería scikit-learn, librería que requiere, a su vez, pandas. Pero la versión de pandas que emplea scikit-learn no es la 3.5.0, sino la 4.0.0. Puesto que, en un mismo sistema operativo, dentro de un mismo entorno virtual, solo puede haber una y solo una versión de una librería (no es posible que convivan pandas 4.0.0 y pandas 3.5.0) el código quedaría inservible, y todo el trabajo habría sido en valde. El cliente quedaría insatisfecho, puesto que ha comprado algo que no funciona como debería.

4.4.2. Conflictos entre versiones de una misma librería

El otro caso problemático no tiene por qué involucrar dependencias. Podría ser, y de hecho sucede, que el código de una determinada librería, por ejemplo, scikit-learn, haya cambiado el nombre de determinados parámetros al actualizarse.

Ejemplificando, la función HashingVectorizer() utiliza un parámetro para controlar el signo de las variables. Este parámetro, en primeras versiones de la librería, se llamaba non_negative. Con el paso del tiempo, la librería se ha ido actualizando, y el parámetro ha pasado a denominarse alternate_sign.

Dado este supuesto, si un usuario desarrollase código en la primera versión, utilizando el primer nombre de parámetro, y actualizase la librería, el código quedaría inservible, puesto que el parámetro ha cambiado de nombre. El parámetro antiguo ya no existe, la función está llamando a un parámetro que ahora es desconocido para ella. El código ya no funciona.

Puesto que es fácil actualizar librerías sin tener demasiado control, debido en parte también a las dependencias de unas de otras, esto representa otro grave riesgo.

4.4.3. Conflictos entre sistemas operativos

Finalmente, cabría destacar también un riesgo adicional asociado al proyecto, y es la compatibilidad entre sistemas operativos. El código desarrollado en Windows puede no funcionar correctamente en Linux o MacOS.

Para paliar estos problemas, y por otros motivos que luego se explican, se toma la decisión de usar Docker.

4.4.4. Docker

Docker nace en 2013 con una idea que revoluciona la industria del desarrollo. Ésta consiste en generar y gestionar compartimentos estancos de software. Dicho en otras palabras, permite utilizar dentro de un mismo ordenador, fragmentos de software

totalmente ajenos a él, aislados e independientes. Estos compartimentos son los contenedores.

Docker tiene tres pilares básicos:

- **Dockerfile:** es un archivo que recoge las instrucciones paso a paso para construir una imagen.
- **Imagen:** es una réplica, una foto, de un fragmento de software en un momento concreto del tiempo.
- **Contenedor:** es la realización de la imagen, el compartimento estanco donde vive el software, en idénticas condiciones a las de la imagen. Es independiente del resto del ordenador.

La idea es la siguiente: se genera un Dockerfile que recoge todos los pasos necesarios para construir todo el sistema, desde el sistema operativo, pasando por las librerías y versiones, código, todo. Con ese Dockerfile, se construye una imagen. Y a partir de esa imagen, se pueden construir tantos contenedores como se quiera. Todos ellos son una realización exacta de la imagen.

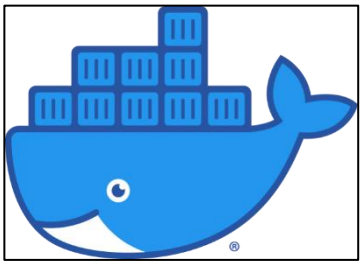


Ilustración 13. Docker. Fuente: docker.com

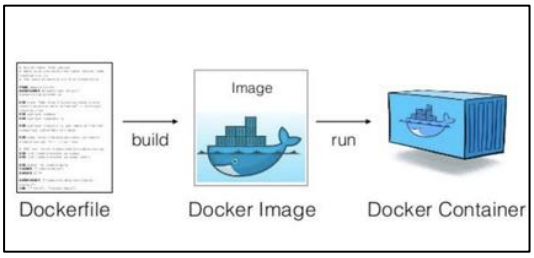


Ilustración 17. Docker. Fuente: xataka.com

Esto brinda tres beneficios claros:

1. Solventa los problemas mencionados: el propio contenedor actúa como un ordenador aparte, con su sistema operativo propio, librerías con sus versiones, etc. Las librerías dentro del contenedor no dependen de las de fuera.
2. Facilita el despliegue del servicio y aumenta su portabilidad: la persona que vaya a utilizar el servicio en su ordenador solo tiene que preocuparse de tener Docker instalado y configurado y hacer docker pull de la imagen colgada en Docker Hub. No necesita tener Python instalado ni las librerías. Otra opción consiste en levantar la imagen directamente en su sistema a partir del Dockerfile. Si se quiere hacer uso de máquinas virtuales en servicios web como Azure o AWS, el empleo del contenedor facilita la tarea.
3. Aumenta la robustez de todo el servicio: al gestionar los contenedores como unidades informáticas aisladas, se tiene la garantía de estar utilizando un entrono controlado, sin interferencias del exterior. Además, en el caso de que un servicio tuviese conflictos dentro de sí mismo (por ejemplo, la interfaz de una web y su servidor), bastaría con generar contenedores separados para cada una de las partes, y establecer la comunicación entre ellos.

Por todos estos motivos, se ha empleado Docker para la construcción de una imagen y contenedor que albergue todo el servicio del asistente virtual, para facilitar su uso y

prueba en otros entornos, garantizando en todo momento el funcionamiento del asistente. El sistema operativo elegido para dicho desarrollo es Linux, con Ubuntu. La imagen está disponible en el siguiente enlace.

Finalmente, cabe mencionar que la dockerización del servicio facilita la integración en un sistema más complejo. Si la empresa que implementase la solución ya tuviese sus sistemas en Docker, sería sencillo añadir uno más. Sería necesario el uso de un orquestador, como Kubernetes, encargado de dirigir todo el ecosistema.

CAPÍTULO V: PROBLEMAS ENCONTRADOS.

5. PROBLEMAS ENCONTRADOS

En la actualidad, se está viviendo un fenómeno que podría catalogarse como una ola de hype de la Inteligencia Artificial. Esta no es la primera vez que sucede algo así, de hecho, podría considerarse la tercera. En 1960, con la IA simbólica, la IA también vivió un periodo de expectativas exageradas, inversión desenfrenada y auge. A este periodo le siguió un tiempo de abandono y decadencia. Años más tarde, en 1980, la historia se repetía, con los sistemas expertos en el centro de mira. Los medios se hacen eco de los avances de la IA, los ciudadanos se dejan llevar por la emoción, y la sociedad como conjunto fija unas expectativas injustificadas. Pasan los años, y cuando desaparece el hype, se seca la inversión y la IA vuelve a caer en un rincón.

La etapa de los días que corren es similar. Términos como Machine Learning, Deep Learning o Neural Network están rodeados de un misticismo casi esotérico, que lleva a pensar a los menos formados que la IA es capaz de todo lo imaginable e inimaginable. Se venden como la solución mágica a todos los problemas del ser humano.

Consideramos que es nuestro deber como científicos de datos mantener la conciencia fría, y no dejarnos llevar por el hype cortoplacista. Por este motivo, esta sección se dedica a comentar algunos de los problemas encontrados en la aplicación de los Transformers a la asistencia virtual, con el fin de ilustrar esa otra cara de la IA, tan importante y olvidada a la vez.

La sección se divide en dos pequeños apartados: problemas relacionados con el texto, y problemas relacionados con el modelo.

5.1. PROBLEMAS RELACIONADOS CON EL TEXTO

El modelo se alimenta del texto. Son sus datos. Como sucede con todos los modelos de Machine Learning, aquí aplica el principio GIGO, garbage in, garbage out. Lo que quiere decir es que los resultados que ofrezca el modelo dependen en enorme medida de la calidad de los datos. Si el texto es malo, el modelo ofrecerá un output malo. Hasta aquí nada nuevo que no pueda solucionarse con una capa intermedia de tratamiento. De hecho, puesto que los datos de este proyecto han sido extraídos de fuentes web directamente con

web scraping, ha sido necesario tratarlos. Por tanto, los problemas graves relacionados con el texto no son exactamente del tipo que se acaba de comentar (puesto que tiene solución) sino con el tipo y forma del texto.

Uno de estos problemas es la longitud de las frases. Los Transformers funcionan mediante el ya mencionado sistema de atención, y las frases muy largas dificultan la tarea de focalización, puesto que tienen un vector más largo que recorrer. El tiempo de respuesta se ve afectado, y puede que la contestación no sea la adecuada.

El contenido y la complejidad de la frase también impactan de manera notable en la respuesta. Si una frase contiene varias ideas, una pregunta simple no va a conseguir que la respuesta sea la adecuada, sería necesario concretar mucho la pregunta.

Otro problema es el estilo del texto, y la riqueza del vocabulario. Si el texto tiene palabras bien diferenciadas, variadas, con sinónimos, y evita la homonimia y la repetición de ideas, el modelo trabaja mucho mejor. Dado el siguiente texto

"Ayer vino el paquete que estabas esperando. Contenía el vino que compraste por internet"

Si se le pregunta "¿Qué vino?", el modelo tendrá serios problemas para focalizar su atención, puesto que cuando vea la palabra homónima (vino), no sabrá donde ir, a no ser que la pregunta sea más explícita.

Otro ejemplo de texto problemático sería:

"La catedral se construyó en el S. XVI. El museo se construyó en el S. XIX"

En este caso parece claro que si se le pregunta al modelo "¿cuándo se construyó?" se está generando una confusión, de igual manera que sucedería con un humano.

Estos pequeños detalles hacen que sea necesario someter el Transformer a numerosas pruebas y conversaciones, con el fin de verificar la adecuación del texto. El proceso de pruebas es como sigue:

1. Iniciar el asistente y mantener la conversación inicial.
2. Elegir un texto para el cual se van a hacer las pruebas y mantenerlo durante la sesión de preguntas.
3. Hacer una serie de preguntas al Trasnformer, con el texto a nuestro alcance.
4. Evaluar dónde está fallando y por qué.
5. Si el fallo es de la pregunta, mejorar el planteamiento.
6. Si el fallo es del texto (lo habitual), acceder al texto y modificarlo.
7. Detener el proceso, realizar los cambios en el texto y guardarlos.
8. Volver al paso 1 y repetir el proceso hasta que el modelo devuelva unos resultados aceptables para todo el texto y diferentes preguntas.

Este proceso es laborioso, y aunque es infinitamente más sencillo que programar toda la conversación de manera explícita, requiere tiempo. A modo de ejemplo, en una sesión normal de "examen del modelo", la conversación puede registrar entre 1.000 y 1.500 mensajes. Esta calibración del texto tiene un carácter artesanal, que dificulta su escalado. Con 5, 10, 15 textos, es laborioso, pero posible. Sin embargo, si se quieren gestionar 100 textos, la cosa se complica. Además, los cambios son sutiles, no tan evidentes como los

ejemplos aquí expuestos. El cambio en una sola palabra puede hacer que el modelo pase de acertar sólo la mitad de las preguntas, a acertar todas.

Finalmente, para la calibración del texto, hay que tener presente la arquitectura del sistema, y si es posible intervenir los datos o no. En un primer momento, se planteó la idea de que los datos del asistente se extrajesen en tiempo real de la web, sin almacenarse en ningún lugar. Este enfoque impediría dicha manipulación, o al menos la dificultaría notablemente, forzando a que fuese programada. Al final, se trata de un trade off entre la calidad de las respuestas y la escalabilidad del sistema. Los modelos funcionan de manera aceptable con los datos sin tratar, pero su verdadero potencial se alcanza cuando el texto ha sido procesado. Es una labor de cada empresa encontrar el punto de equilibrio que mejor se ajuste a su estrategia de negocio.

5.2. PROBLEMAS RELACIONADOS CON LOS MODELOS

Los modelos han presentado una serie de inconvenientes que se han ido descubriendo a lo largo de la investigación.

Uno de los problemas lo constituyen los puntos. Por cómo están entrenados y construidos, los modelos utilizan los puntos para localizar las frases en el texto. Esto hace que tengan problemas si el texto contiene abreviaturas, por ejemplo, “S. XIX”, puesto que los confunde, y piensan que la frase termina después de la S, cuando en realidad es una abreviatura.

Otro inconveniente de los puntos son las páginas web. Si el texto contiene un enlace, por ejemplo, cunef.com, el modelo no es capaz de devolverlo como un todo, haciendo que sea “clicable”. De nuevo, interrumpe la frase en el punto, y devuelve el primer trozo. Este comportamiento no se ha producido siempre, pero resta fiabilidad a los modelos en este aspecto.

Otro problema sería la falta de memoria. Si los Transformers necesitasen recordar datos del usuario, sería necesario auxiliarlos con programación tradicional, que guardase dichos datos. Estos modelos, al menos por el momento, no presentan ningún tipo de memoria, y tampoco pueden adquirir experiencia a medida que avanza la conversación.

Finalmente, y en relación con el inicio de este apartado, el razonamiento. Los modelos no poseen una inteligencia humana. Esto parece evidente, pero es necesario recordarlo, puesto que la precisión que tienen puede llegar a cegarnos, sobre todo cuando se pasan varias horas trabajando con ellos.

Lo que se quiere enfatizar es que los modelos no pueden razonar. Si el texto contiene “el horario de lunes a jueves es de 14:00 a 20:00. Los viernes está cerrado”, el modelo, naturalmente, es capaz de responder a “¿Cuál es el horario de lunes a jueves?”. O “¿Qué día está cerrado?”. Pero no es capaz de responder a preguntas del estilo “¿Está cerrado los viernes?”, y contestar “sí” o “no”. Esto, que puede parecer trivial, es un aspecto crucial que los aleja, al menos por el momento, de la capacidad cognitiva humana. Todo lo que saben es lo que está en el texto, y tal y cómo está en él, no de otra forma. No son capaces de construir razonamientos lógicos sobre ese texto, como sí lo haría un humano.

Estos hechos, que así expuestos parecen evidentes, son los mismos que generan la inflación de expectativas en torno a la IA, si no se tienen presentes. También hay que mencionar que los Transformers están en una fase temprana de desarrollo, y que para el poco tiempo que tienen, ofrecen unos resultados asombrosos. Pero dejarse llevar por el hype solo contribuirá a dificultar el progreso de estas tecnologías, puesto que se fijarán expectativas imposibles de satisfacer. Es necesario mantener plena conciencia de qué son y qué no son los Transformers, y la Inteligencia Artificial en general.

CONCLUSIONES

De la evidencia empírica de esta investigación, es posible llegar a una serie de conclusiones, las cuales se presentan en esta sección.

Los asistentes virtuales son una herramienta crucial en la actualidad digitalizada, que pueden multiplicar la productividad de las empresas. Sin embargo, dichos sistemas se quedan obsoletos al tratar de cubrir la infinidad de preguntas que un usuario puede formular. Por su naturaleza, son sistemas de difícil escalabilidad, y cuya complejidad programática crece de manera exponencial a medida que se intenta dar respuesta a más necesidades de los clientes, poniendo en peligro los recursos de la empresa, y generando gastos.

Para poder solucionar estos inconvenientes, es posible emplear, tal como se ha demostrado en este trabajo, los avances de la ciencia de datos, Machine Learning y en especial Deep Learning. El aporte concreto son los modelos Transformers de NLP. Éstos proponen una solución a la gestión de preguntas y respuestas, al manejo de una conversación, y a la escalabilidad.

Tienen la capacidad suficiente para contestar a una amplia gama de preguntas, tantas como albergue el texto del que se nutren. Las ventajas de usar estos modelos son entre otras, la respuesta cerrada y concreta, con el fragmento exacto del texto, la escalabilidad del sistema, la independencia, en muchos casos, del idioma, y, sobre todo, el reemplazo de la programación explícita. Hacen que no sea necesario codificar las respuestas, siendo estas generadas desde el propio modelo. De este modo, generan valor tanto para las empresas que los emplean como para los clientes y usuarios finales.

En concreto, este enfoque de asistencia virtual apalancada por modelos de aprendizaje automático constituye una buena opción para el sector turismo. Las empresas de viajes han de ofrecer respuesta a una amplia gama de preguntas, para diferentes clientes, y distintos idiomas. En muchas ocasiones, lo único que varía es el contexto, el lugar del destino. Por este motivo, la adopción de estos modelos dentro de este sector puede suponer grandes avances, aumentando la productividad y salvando algunos problemas irresolubles con el enfoque de asistencia virtual tradicional. El uso de estos modelos constituye sin duda una ventaja competitiva para estas empresas.

La materialización de esta idea, y su puesta en producción, se hace mediante el software de código libre Python. Otra alternativa posible es la adopción de una aplicación de mercado. Concretamente, este trabajo se desarrolló inicialmente utilizando Watson

Assistant de IBM, pero por una serie de motivos explicados a lo largo del informe, el desarrollo final tiene lugar a través de Python, y concretamente, utilizando la librería Hugginface. Esta librería cuenta con numerosos modelos de Transformers. En este proyecto, se han utilizado los cuatro modelos de español disponibles, distill BETO, BETO, Electra y RuPERTa, aunque el último de ellos no se incluye en la versión final, por no ofrecer buenos resultados.

No obstante, el uso de Python presenta diversos problemas que ponen en peligro el éxito del sistema. Algunos de estos son la gestión de versiones de librerías, los conflictos entre dependencias, y problemas derivados del sistema operativo. Estos inconvenientes son críticos, y amenazan la puesta en producción del sistema, y la generación de valor para el negocio.

La solución aquí propuesta pasa por Docker, una tecnología de creación de contenedores de software que permite crear compartimentos informáticos aislados. Así, se evitan los conflictos entre sistemas operativos, y versiones de librerías, puesto que éstas permanecen fijas y controladas, a través del dockerfile. Este enfoque permite portabilizar el sistema, haciendo que cualquier agencia de viajes que quiera adoptar la idea tenga un fácil acceso al software, y pueda desplegarlo de manera limpia, segura y controlada.

Con todo, los modelos no son perfectos. A lo largo de la investigación se han descubierto una serie de puntos débiles que convienen controlar y mencionar. En relación con el texto, los problemas vienen causados por palabras homónimas, frases con ideas poco diferenciadas, y el estado general del texto. Una depuración de este mejora los resultados de manera significativa. El resto de los problemas tienen que ver con la capacidad cognitiva de los Transformers. Por el momento, estos modelos carecen de una inteligencia que les permita razonar, y construir pensamientos lógicos sobre el texto con el que trabajan.

También es necesario mencionar la visión del negocio. Contar con un enfoque económico, centrado en las necesidades del mercado, es calve. Los modelos y la idea pueden ser brillantes, pero sin una buena propuesta de valor, perecerán. Por este motivo, se ha elaborado un plan de marketing y un plan financiero, pudiéndose demostrar la rentabilidad que supone este tipo de asistente virtual, pudiendo hacer frente a los costes con los que se puede encontrar. También se hace referencia a los aspectos legales a tener en cuenta. La mejora del sistema pasaría por extender su implementación a plataformas más populares, como WhatsApp, o incluso la web.

Finalmente, cabe hacer una mención a la novedad de los modelos, un hecho que desde el punto de vista teórico representa un verdadero reto. Siendo los Transformers una familia de modelos tan novedosa, recabar información teórica, para comprender su funcionamiento, es una tarea sumamente compleja. La historia y literatura que existe al respecto es mínima o nula, y las explicaciones no siempre son comprensibles. Este hecho puede suponer una dificultad para aquellas empresas que quieran adoptar esta idea desde un punto de vista interpretable y explicativo, tal y como ha supuesto para este trabajo. No obstante, con un equipo de científicos de datos de profesionales bien formados, y un compromiso por la ciencia y el negocio, es posible superar el bache, y guiar el uso de Transformers en asistentes virtuales hacia el éxito empresarial.

BIBLIOGRAFÍA

Adamopoulou, E., & Moussiades, L. (2020). *An Overview of Chatbot Technology.* académico, International Hellenic University, Department of Computer Science, Kavala, Grece.

Aunoa. (s. f.). ¿Cuál es el precio de un chatbot con IA para tu empresa? aunoa. https://aunoa.ai/cual-es-el-precio-de-un-chatbot/

Ayto Madrid. (2020, 7 de julio). Padrón Municipal de Habitantes (explotación estadística) - Ayuntamiento de Madrid. Inicio - Ayuntamiento de Madrid. https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Demografia-y-poblacion/Cifras-de-poblacion/Padron-Municipal-de-Habitantes-explotacion-estadistica-/?vgnextfmt=default&vgnextoid=e5613f8b73639210VgnVCM1000000b205a0aRCRD&vgnextchannel=a4eba53620e1a210VgnVCM1000000b205a0aRCRD

Burkov, A. (2020) *Machine Learning Engineering.*

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., . . . McTear, M. (2017). Towards a chatbot for digital counselling. *Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017) (HCI).*

Chollet, F (2017) *Deep Learning With Python.* Manning.

ECIJA (2018) Guía Legal Chatbots: aspectos jurídicos y de mercado. https://ecija.com/wp-content/uploads/2018/10/ECIJA_Chatbot-Chocolate_Paper_Aspectos-jur%C3%ADdicos-y-de-mercado_Chatbots-compressed.pdf

ECIJA (2017) Web Scraping: ¿legal o ilegal? https://ecija.com/web-scraping-legal-ilegal/ artículo de Sonia Vázquez

Facebook for Developers. (2019, 1 de junio). Precios - API de WhatsApp Business – Documentación. https://developers.facebook.com/docs/whatsapp/pricing?locale=es_ES

Faggella, D. (13 de diciembre de 2019). *7 Chatbot Use Cases That Actually Work*. Obtenido de Emerj: https://emerj.com/ai-sector-overviews/7-chatbot-use-cases-that-actually-work/

Forcada, M. (25 de mayo de 2020). *Torre Juana*. Obtenido de Avances en el Procesamiento del Lenguaje Natural y la IA, sesión con M. Forcada: https://ost.torrejuana.es/avances-en-el-procesamiento-del-lenguaje-natural-y-la-ia/

INE. Instituto Nacional de Estadística. (s. f.). INE. https://www.ine.es

INE. (2020, 21 de abril). Avance de la estadística del padrón continuo. (2020). INE. Instituto Nacional de Estadística. https://www.ine.es/prensa/pad_2020_p.pdf

koleva, N. (10 de abril de 2020). *Dataiku*. Obtenido de https://blog.dataiku.com/whats-new-in-nlp-transformers-bert-and-new-use-cases

Kore.ai A Handy Guide on How to Drive Tourism with Chatbots https://blog.kore.ai/a-handy-guide-on-how-to-drive-tourism-with-chatbots [12 mayo 2021]

kortschak, H. (16 de noviembre de 2020). *Towards Data Science.* Obtenido de https://towardsdatascience.com/attention-and-transformer-models-fe667f958378

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N. Gomez, A., . . . Polosukhin, I. (2017). *Attention Is All You Need.*

Lokman, A. S., & Mohamed , A. A. (2019). *Modern Chatbot Systems: A Technical Review.* Universiti Malaysia Pahang,, IBM Centre of Excellence, Faculty of Computer Systems and Software Engineering. Pekan, Malaysia: Springer Science and Business Media LLC.

Maxime. (4 de junio de 2019). *Medium.* (I. Analytics, Ed.) Obtenido de Inside Machine Learning: https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04

Medium (2019) Conversational AI: the Growing Potential of Chatbots and Intelligent Personal Assistants for Businesses https://medium.com/@infopulseglobal_9037/conversational-ai-the-growing-potential-of-chatbots-and-intelligent-personal-assistants-for-27d6be734924

Medium (2019) Are Chatbots Revolutionizing The Travel Industry? https://medium.com/swlh/are-chatbots-revolutionizing-the-travel-industry-12c195b63fb5 [12 mayo 2021]

Mora, E. (26 de noviembre de 2020). *Towards Data Science.* Obtenido de https://towardsdatascience.com/improving-virtual-assistants-performance-using-semantic-search-and-sentence-transformers-9d654b0bb9e6

Pisu, F. (2020, 9 de octubre). WhatsApp Business: Todos los costes de un vistazo. Userlike Live Chat. https://www.userlike.com/es/blog/whatsapp-business-costes#:~:text=Aunque%20la%20app%20de%20WhatsApp,según%20la%20gama%20de%20funciones.

Towards Data Science (2020) What are transformers and how can you use them? https://towardsdatascience.com/what-are-transformers-and-how-can-you-use-them-f7ccd546071a [27 mayo 2021]

Vaca, A. (24 de marzo de 2021). *Instituto de Ingeniería del Conocimiento*. Obtenido de https://www.iic.uam.es/innovacion/transformers-en-procesamiento-del-lenguaje-natural/

Vajpayee, S. (6 de agosto de 2020). *Towards Data Science.*