

**XIKUA Boletín Científico de la Escuela Superior de Tlahuelilpan**  
**13° Congreso Internacional de Computación**  
**Inteligencia artificial: Presente y futuro**  
**Red Iberoamericana de Computación**  
*Vol. 12, Número Especial (2024) 151-160*

**Procesamiento de Lenguaje Natural, los Transformers y los Bots Conversacionales:  
algunas generalidades**

**Natural Language Processing, Trasnformers and Chatbots: some overview**

*Héctor Julio Fúquene Ardila*

---

**Abstract:**

In the present investigation, a description of the Transformer architecture is carried out, emphasizing the most important functional blocks with which these models went from using serial processing and training of recurrent neural networks to parallel processing, thus improving response times and increased processing capacity; that is to say, the Transformers meant a notable evolution in the PLN Natural Language Processing. These achievements were obtained through the use of attention mechanisms and referencing the positions of the words of the inputs, an aspect that facilitated the models to manage a long-term 'memory'; feature, which was a limitation of the old models. In addition, the uses that are currently being given to this architecture and that are not linked to natural language processing (Chatbots) are highlighted, as is the case in the area of cybersecurity with the use of the SecGPT Bot, among others. development and application areas.

**Keywords:**

Transformer, attention networks, ChatBots, machine learning, NLP

---

**Resumen:**

En el presente análisis se realiza una descripción de la arquitectura Transformer, haciendo énfasis en los bloques funcionales más importantes con los cuales estos modelos pasaron de utilizar un procesamiento y entrenamiento en serie de las redes neuronales recurrentes a un procesamiento en paralelo, con lo que mejoraron notoriamente los tiempos de respuesta y aumentaron la capacidad de procesamiento; es decir los Transformer significaron una evolución notable en el Procesamiento de Lenguaje Natural PLN. Estos logros fueron obtenidos a través del uso de mecanismos de atención y referenciando las posiciones de las palabras de las entradas, aspecto que facilitó que los modelos manejaran una 'memoria' a largo plazo; característica, que era una limitante de los antiguos modelos. Se resaltan, además, los usos que en la actualidad se le está dando a esta arquitectura y que no están ligados al procesamiento del lenguaje natural (Chatbots), como es el caso del área de la ciberseguridad con el uso del Bot SecGPT, ente otras áreas de desarrollo y aplicación.

**Palabras Clave:**

Transformer, redes de atención, ChatBots, machine learning, NLP

---

**Introducción**

Y se puso de moda la Inteligencia Artificial (AI- IA, depende del idioma). Es tal el boom de la IA, que resultó convirtiéndose en un tema de dominio público, esto debido,

principalmente, al uso de los ya muy famosos Chatbots como ChatGPT (Generative Pre-trained Transformer o en español Transformador pre- entrenado generativo); el cual no es el único Chatbot, pero si es el más popular en la actualidad. En el presente trabajo se aborda el tema de estos modelos de

---

<sup>3</sup> Autor de Correspondencia, Universidad Distrital FJDC, <https://orcid.org/0009-0005-3946-8190>, Email: [hfuquene@udistrital.edu.co](mailto:hfuquene@udistrital.edu.co)

inteligencia artificial, haciendo un recuento histórico de su invención, del uso inicial que se les dio; de su importancia en el desarrollo del Procesamiento del Lenguaje Natural (NLP) y de su popularización gracias a la creación e implementación de una arquitectura con la capacidad de procesamiento en paralelo y con la posibilidad de ser aplicada no solo en el ámbito de NLP, sino a varias áreas en las cuales su impacto ya es notorio. El uso de los Transformers y la utilización de GPUs (Graphic Processing Unit), que posibilitaron el procesamiento de datos en paralelo, fue lo que hizo que esta área de la inteligencia artificial sea una de la de mayor proyección y una de las más usadas en la actualidad; esto debido a su gran capacidad y rapidez de procesamiento y la facilidad con la que a cualquier usuario se le permite interactuar.

Los modelos Transformer, quienes son los que en últimas hacen la ‘magia’, al ser capaces de procesar infinidad de parámetros y gracias a su entrenamiento previ, son hoy por hoy, una de las fuentes de información más consultadas en temáticas especializadas.

Para entender un poco más como opera un Transformer se abordarán de manera somera algunos temas relacionados, como son: el Procesamiento del Lenguaje Natural PLN, las Redes Neuronales, el Machine Learning, las redes de autoatención, las Redes Neuronales Recurrentes RNR, entre otros temas relacionados.

Finalmente se describen algunas características de SecGPT, que es un Chatbot creado y entrenado para apoyar a los profesionales de la ciberseguridad; cabe destacar, que en la actualidad existen Chatbots especializados en varias temáticas, como es la medicina, el comercio, la visión artificial, entre otros.

## OBJETIVO

Describir la arquitectura Transformer resaltando las principales características y el uso en la generación de Chatbots.

### METODOLOGÍA

El propósito de este artículo es conocer los soportes técnicos y metodológicos utilizados por los Transformers, conocer su funcionamiento, los algoritmos utilizados y entender su funcionamiento y como estos son utilizados en los modelos de Chatbots. Para la investigación se revisó la bibliografía e infografía a partir del año 2017 a la fecha; en su mayoría es información referente a congresos y a artículos científicos publicados en revistas especializadas, es de anotar que su acceso en la gran mayoría de casos se logró en línea. La búsqueda de información se realizó en torno a las siguientes palabras clave: procesamiento de lenguaje natural, redes neuronales, redes neuronales recurrentes, Transformers, Chatbots, redes de atención, procesamiento en paralelo y GPUs. En la metodología utilizada se estructuraron algunas preguntas orientadoras con las cuales se pretendió delimitar el alcance de la investigación. Las preguntas usadas fueron: ¿qué es un Chatbot o qué es un bot conversacional?, ¿Qué tipos de bots existen?, ¿qué es un Transformer?, ¿En qué consiste el procesamiento de lenguaje natural?, ¿Qué es red

seq2seq? ¿Qué es machine learning?, ¿Qué es un mecanismo y red de atención? ¿Qué es un tensor?, ¿Cuáles son los Chatsbots más utilizados? ¿Qué arquitectura tiene un Tranformer? ¿Qué casos de uso tienen los Transformer? ¿Qué aplicaciones existen para ciberseguridad?

## FUNDAMENTOS TEÓRICOS

*Procesamiento de lenguaje natural PLN-NLP:* el procesamiento de lenguaje natural, abreviado PLN o en inglés, Natural Language Processing NLP, es un campo de las ciencias de la computación, de la inteligencia artificial (Machine Learning) y de la lingüística que estudia las interacciones entre los computadores y el lenguaje humano [1].

*Red Neuronal Artificial:* es un método de la inteligencia artificial que enseña a los computadores a procesar datos de una manera que está inspirada en la forma en que lo hace el cerebro humano [2]. Un RN consta de un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales, está constituida por capas (capa de entrada, capa(s) oculta(s) y capa de salida). Ver figura No. 1.

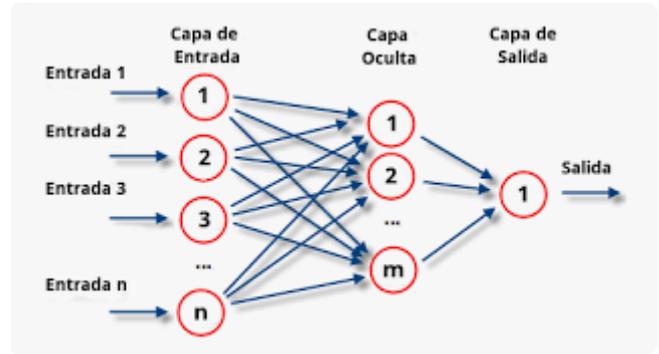


Figura No. 1. Red neuronal

Fuente:

[https://es.wikipedia.org/wiki/Perceptr%C3%B3n\\_multicapa](https://es.wikipedia.org/wiki/Perceptr%C3%B3n_multicapa)

*Chatbot (bot conversacional):* son aplicaciones software que surgen en los años 60, y que simulan mantener una conversación con una persona al proveer respuestas automáticas, las cuales, son previamente establecidas por un conjunto de expertos a entradas realizadas por el usuario. Utilizan redes neuronales para procesar y ‘aprender’ o generar conocimiento, se pueden clasificar en los orientados a tareas y los basados en datos predictivos, que son conocidos como asistentes digitales [3][4]. Ver figura No. 2.

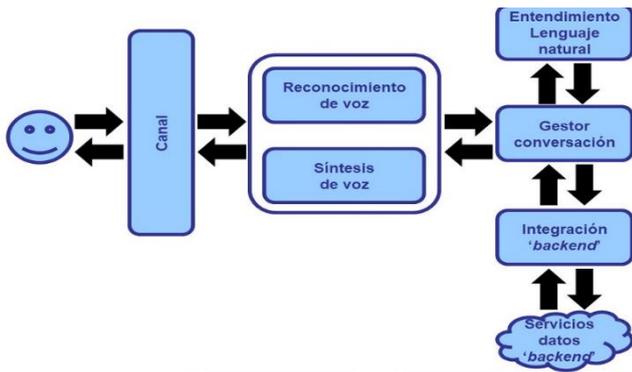


Figura No. 2. Esquema de un Chatbot  
Fuente: <https://ignaciogavilan.com/la-arquitectura-basica-de-una-interfaz-conversacional-chatbot/>

**Transformer:** son redes neuronales que aprenden contexto y, por lo tanto, significado mediante el seguimiento de relaciones en datos secuenciales como las palabras de una oración. Fueron inicialmente utilizados para la traducción de un lenguaje a otro [5]. Los Transformers y su mecanismo de atención posibilitaron la aparición de los grandes modelos generadores de texto GPT2, GPT3 y BERT que ahora podían ser entrenados aprovechando el paralelismo que se alcanza mediante el uso de GPUs. Ver figura No 3. [6].

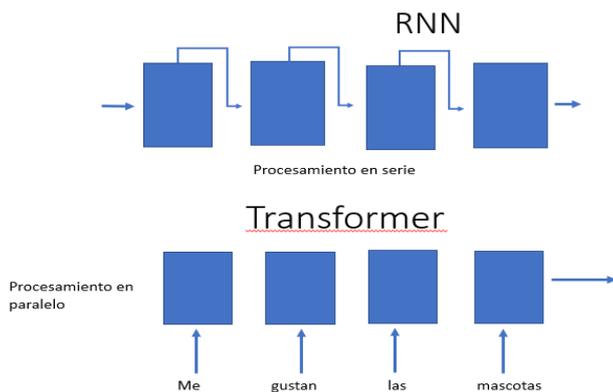


Figura No. 3. Comparación Redes Neuronales Recurrentes y un Transformer  
Fuente: Elaboración propia

**Computación paralela:** es una forma de cómputo en la que muchas instrucciones se ejecutan simultáneamente, operando sobre el principio de que problemas grandes, a menudo se pueden dividir en unos más pequeños, que luego son resueltos simultáneamente, permitiendo con esto un mayor rendimiento en las aplicaciones [7].

**GPU (Graphic Process Unit):** es un coprocesador dedicado al procesamiento de gráficos u operaciones de coma flotante con lo que se logra aligerar la carga de trabajo del procesador C.P.U. central, son usados principalmente en aplicaciones como los videojuegos o aplicaciones 3D interactivas.

La capacidad de las GPU para procesar grandes bloques de datos en paralelo las hace óptimas para entrenar a la inteligencia artificial (IA) y los modelos de aprendizaje profundo que precisan un procesamiento paralelo intenso de cientos de miles de redes neuronales al mismo tiempo [8].

**Red de atención:** son una clase reciente de redes neuronales para secuencias, basadas en la autoatención, que han demostrado estar bien adaptadas al texto y actualmente están impulsando importantes avances en el procesamiento del lenguaje natural. Se encargan de generar tokens a partir de un texto de entrada, jerarquizarlo y referenciarlo de acuerdo a la posición de cada palabra en la frase, esto con el fin generar una salida en un lenguaje diferente al entrado. Su procesamiento difiere de una red neuronal recurrente en que el procesamiento se hace en paralelo [9].

**Tensor:** son objetos matemáticos que almacenan valores numéricos y que pueden tener distintas dimensiones. Así, por ejemplo, un tensor de 1D es un vector, de 2D una matriz, de 3D un cubo etc, [10]. Los tensores son usados en Frameworks de Deep Learning, estos utilizan los tensores como su unidad computacional básica. Como resultado, pueden utilizar los aceleradores de hardware (por ejemplo, GPU), acelerando así el entrenamiento y la inferencia del modelo [11].

**Machine Learning (Aprendizaje Automático):** El aprendizaje automático es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que los computadores aprendan. Esta rama de la inteligencia artificial utiliza algoritmos que podemos clasificar en tres clases a saber: aprendizaje supervisado, aprendizaje no supervisado, y el aprendizaje por refuerzo [12].

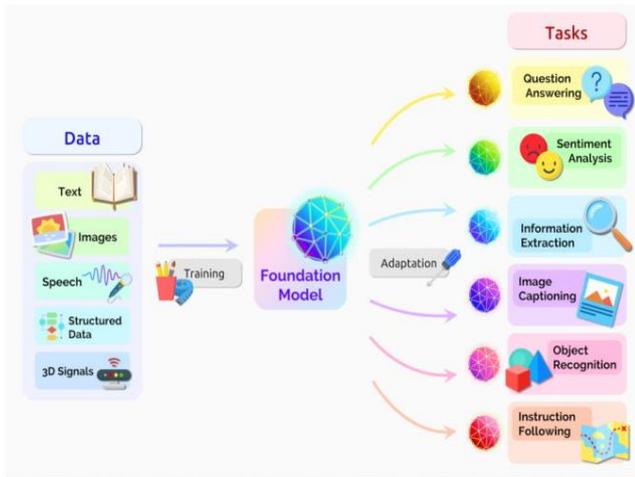
**Redes seq2seq (Sequence-to-Sequence):** es un modelo de Deep Learning que se utiliza para procesar secuencias de entrada y generar secuencias de salida en aplicaciones de procesamiento del lenguaje natural, como traducción automática, resumen de texto o respuesta a preguntas, entre otras [13]. Estas son reemplazadas por los chatbots y sus redes de atención.

**Mecanismo de atención:** en el aprendizaje profundo es una técnica utilizada para mejorar el rendimiento de una red neuronal al permitir que el modelo se centre en los datos de entrada más importantes mientras genera predicciones [14].

## ARQUITECTURA TRANSFORMER

La arquitectura Transformer fue presentada en 2017 en el documento titulado "Attention is all you need" [15], como una herramienta de Deep Learning para el procesamiento del lenguaje natural (PLN), y la cual adoptaba métodos innovadores para la traducción de un idioma a otro de forma más eficiente que las redes seq2seq (Sequence-to-Sequence) usadas hasta ese entonces. Una de las novedades de la arquitectura fue el poder realizar el procesamiento en paralelo y la forma innovadora de utilizar el mecanismo de atención, con lo que se optimizaban considerablemente los tiempos de respuesta. La arquitectura constituida por una serie de codificadores y decodificadores no solo tiene aplicación para la traducción automática del lenguaje natural, sino varias áreas; como se ilustra en la figura No. 4; en donde se resalta el hecho de que los datos de entrada, no se limitan únicamente a la voz y el texto; sino que incluye imágenes, bases de datos y señales multimedia. Se resalta además los casos de uso, en donde se encuentra la respuesta

a interrogantes, el análisis de sentimientos, la extracción de información o sumariación, reconocimiento de objetos y subtítulos y el seguimiento a instrucciones.



Los transformers, a veces llamados modelos de base, ya se están utilizando con muchas fuentes de datos para una gran cantidad de aplicaciones.

Figura No. 4. Aplicaciones de los Transformer  
Fuente: <https://la.blogs.nvidia.com/2022/04/19/que-es-un-modelo-transformer/>

A continuación, en la figura No. 5. se presenta la arquitectura Transformer consignada el documento original de 2017, y se hará un desglose de los principales componentes y su funcionamiento.

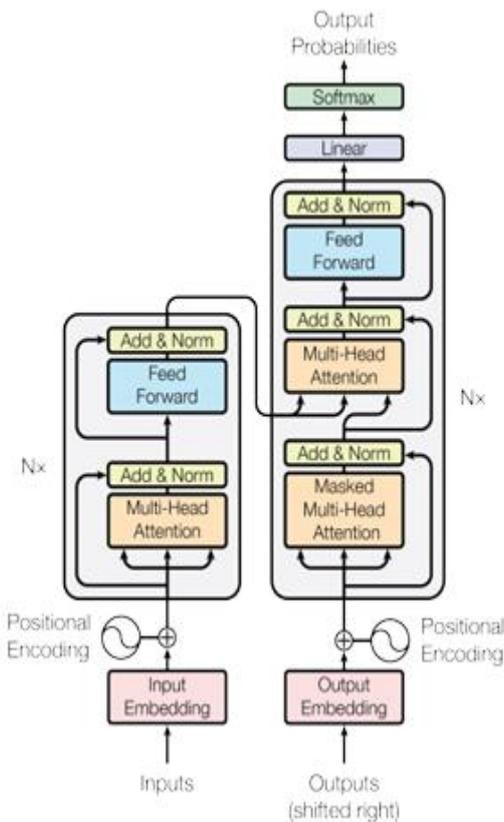


Figura No. 5. Arquitectura Transformer  
Fuente: Attention Is All You Need. Google 2017

En esta arquitectura se resaltan los bloques funcionales de Embedding (entrada y posición), y la serie de codificadores en bloque de entrada (Nx) y los decodificadores en el bloque de salida, los cuales se explicarán a continuación

De una manera sencilla se puede concebir un Transformer como un sistema de forma Entrada --> Tranaformer --> Salida, como se aprecia en la figura No. 6. Como ya se anotó previamente la entrada no necesariamente es texto, pues depende del uso que se le dé al Transformer.

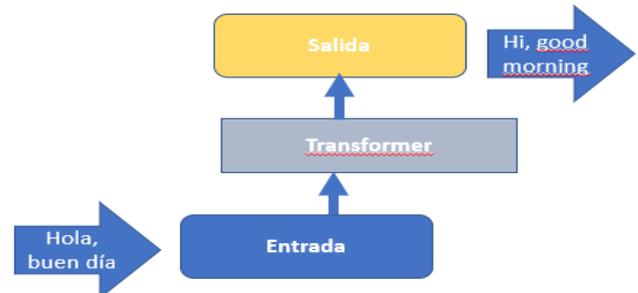


Figura No. 6. Esquema general de un Transformer  
Fuente: Elaboración propia

Haciendo un desglose del esquema anterior podemos identificar nuevos componentes como se muestra en la figura No. 7. A continuación se hace una breve descripción de los elementos nuevos.

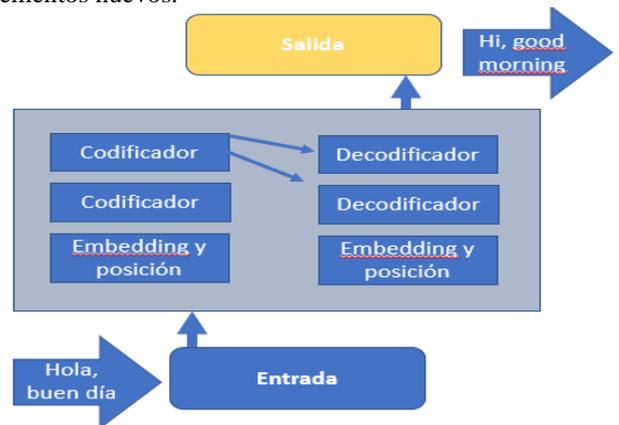


Figura No. 7. Esquema Transformer con entrada, codificadores, decodificadores y salida  
Fuente: Elaboración propia

*Los Embeddings (incrustación de palabras):* Estos usan una técnica de procesamiento de lenguaje natural (usado desde antes de los Transformers), que con base en un texto de entrada lo convierte en unos equivalentes numéricos (tokens) que son almacenados en vectores (de longitud 512), los cuales representan las palabras del texto entrado. Adicionalmente, se generan vectores que referencian las posiciones de las palabras en la oración; estos dos elementos se suman (token + posición) y son utilizados por el mecanismo de atención [16][17][18].

*Codificador posicional:* Es una de las innovaciones del modelo Transformer el cual permite con el procesamiento en paralelo y contribuir en la construcción de la memoria del modelo. Este bloque de codificación agrega un vector de

longitud 512 para codificar las posiciones de las palabras a través de funciones seno y coseno; es utilizado por los mecanismos de atención [17].

**Los Codificadores:** La arquitectura original presentada en 2017 [15], contemplaba un total de 6 codificadores y 6 decodificadores, los cuales a través del procesamiento de matrices y de algoritmos etiquetan los elementos entrados asociándolos a su posición. En la figura No. 8. Se muestran algunos detalles de la estructura de un codificador.

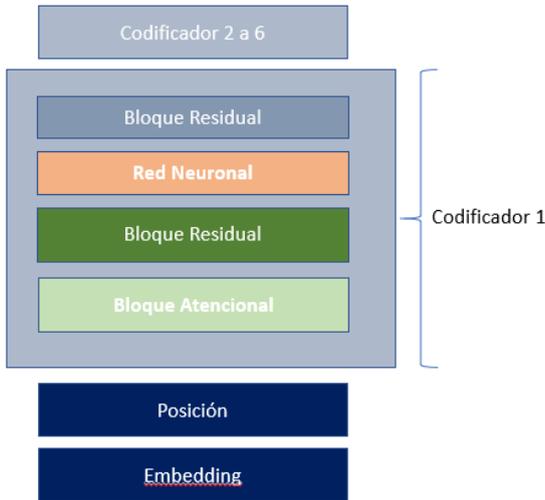


Figura No. 8. El codificador Transformer  
Fuente: Elaboración propia

**Bloque atencional:** es el encargado del análisis de la secuencia de entrada y de establecer relaciones entre sus componentes (entre palabras y entre subfrases de la secuencia de entrada), ver figura No. 9. Recordemos que el procesamiento es en paralelo es una de las principales características y que el bloque atencional se puede considerar como el principal componente del decodificador y del Transformer. Como su nombre lo indica, este bloque es el que establece las relaciones para indicar o jerarquizar cuales son los componentes (tokens) que demandan mayor atención y con los cuales el decodificador estructura la salida de acuerdo a la atención señalada por el mecanismo. En la figura No. 10, se describen parte de los procesos que se han desarrollado en esta instancia. En la gráfica a) se muestra como ejemplo una frase inicial de entrada, la cual es convertida en valores numéricos como se aprecia en la gráfica b); esto teniendo en cuenta la relación e importancia de las palabras de la entrada. Luego en la figura c) se muestra una matriz con valores ponderados en donde el valor corresponde al grado de importancia de una palabra en la frase total; a mayor valor, mayor importancia y por ende mayor atención. Estos valores corresponderían al producto de los vectores Query con los Key, que con la aplicación de una función de activación softmax genera valores entre cero y uno, como se describe más adelante [16][17][18].



Figura No. 9. Relaciones de una frase entrada al Transformer  
Fuente: Elaboración propia

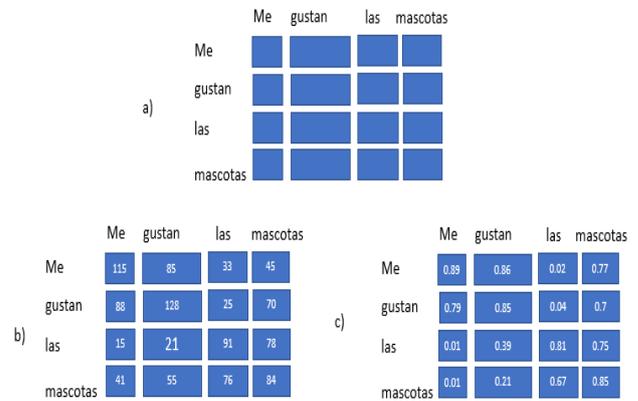


Figura No. 10. Funciones del bloque atencional  
Fuente: Elaboración propia

En esta instancia, el procesamiento está a cargo de tres redes neuronales entrenadas, las cuales generan tres vectores que son: Query(ies), Key(s) y el Value(s), la cantidad de vectores depende del número de palabras obtenidas; siendo estos, representaciones de los datos originales. Ver figura No. 11.

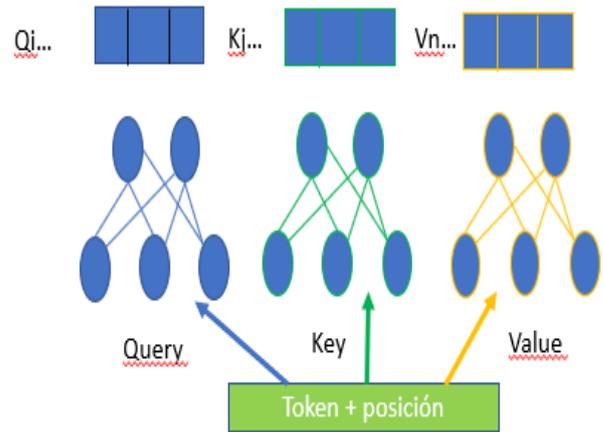


Figura No. 11. Bloque de atención  
Fuente: Elaboración propia

Mediante la comparación (multiplicación) de los de los vectores  $Q_i$  con los  $K_j$  se obtiene unos puntajes que miden el grado de relación existentes (atención) entre palabras o sub -palabras (tokens). La aplicación de la función softmax calcula la distribución de probabilidades del evento sobre 'n' eventos diferentes; es decir, se calculará las probabilidades de cada vector Query con cada vector Key obtenido; valores que como se mencionó previamente están entre cero y uno [18][21]. Ver figura No. 12 y 10 C.

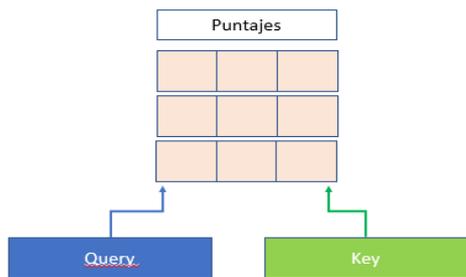


Figura No. 12. Producto de Query y Key  
Fuente: Elaboración propia

Con estos valores obtenidos (matriz de puntajes), se genera un solo vector por cada token, y se halla posteriormente el producto de este con los datos del vector Values y obteniendo nuevos valores que representan la codificación de la información más relevante de contexto para cada token de la secuencia [16][17]. En la figura No. 13, se ilustra el proceso.

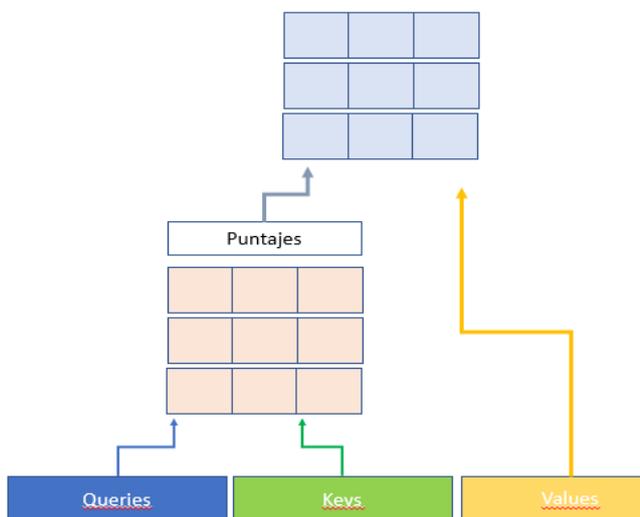


Figura No. 13. Producto de la matriz de puntajes y la de Values  
Fuente: Elaboración propia

La salida obtenida en todos los anteriores procesos se pasa por otra red neuronal que genera un único vector para cada token, luego pasa al bloque de decodificación; es importante resaltar, que se utilizan múltiples bloques atencionales que se encargan de encontrar las relaciones entre palabras y sub frases de la entrada.

**Bloque residual:** recibe tanto la entrada como la salida del bloque atencional, ver figura No. 14, esto con el fin de que los datos no se degraden en el proceso; luego los suma y los normaliza garantizando un estado adecuado para el siguiente proceso, en el cual los datos son procesados por una red neuronal y otro bloque residual [17]. Estos vectores poseen la información atencional requerida en el proceso de decodificación, para estructurar una salida equivalente, pero en un segundo idioma.

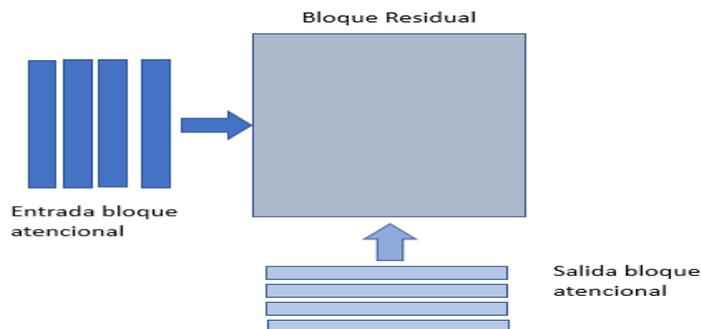


Figura No. 14. Bloque residual  
Fuente: Elaboración propia

**Los Decodificadores:** la arquitectura y funcionamiento de un decodificador es muy similar al de un codificador, por tanto, solo se hará una descripción breve de algunos elementos y conceptos propios de este bloque funcional. Es de anotar, que esta estructura se encarga de la traducción. En la figura No. 15, se muestra el diagrama de bloques de un decodificador, que igual que en el codificador está constituido por 6 decodificadores.

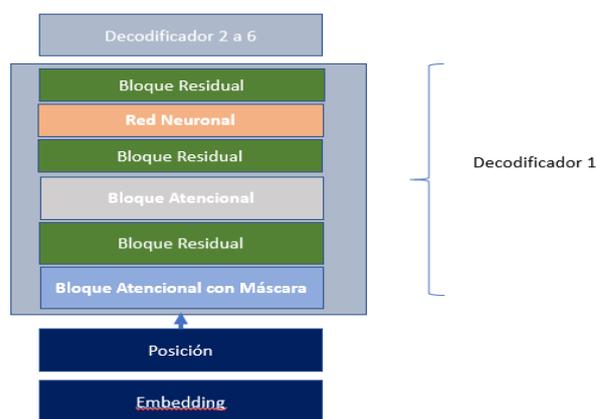


Figura No. 15. Esquema genérico de un decodificador  
Fuente: Elaboración propia

Como la entrada toma los Embeddings y las posiciones con las funciones descritas en el codificador, luego encontramos los 6 decodificadores que están conectados a cada codificador de los cuales toman la información atencional en el idioma original con la cual se hace la traducción [15][17].

La principal diferencia entre un codificador y un decodificador radica en la inclusión de un bloque atencional y uno residual adicionales. Un bloque atencional con máscara que funciona de forma secuencial; el cual, realiza un análisis comparando únicamente hasta las palabras en donde va procesando, obviando las futuras; es decir enmascara parte de la entrada en el proceso de traducción, el cual valiéndose de una capa lineal (red neuronal) y de la aplicación de una función de activación softmax realizan la generación palabra por palabra de la secuencia de salida [15][17], como se mostró en la figura No. 7.

**El bloque atencional del decodificador:** difiere del codificador en que centra su atención en la secuencia de entrada como en la que va a producir; para esto, toma las

salidas del codificador y las pasa a las redes Queries y Keys; el bloque Values toma como entrada la salida del bloque residual anterior como se ilustra en la figura No. 16.

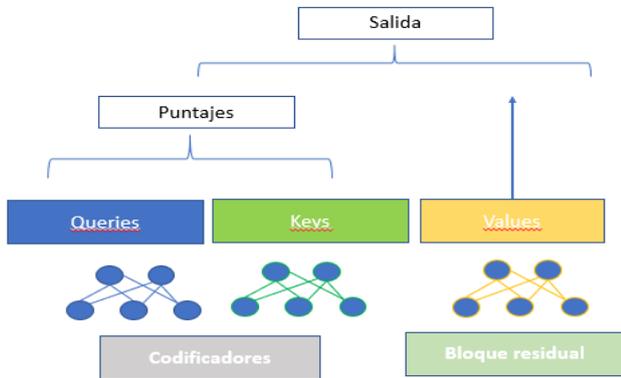


Figura No. 16. Bloque atencional de un decodificador  
Fuente: Elaboración propia

En el proceso de decodificación la capa softmax toma cada elemento del vector producido por la capa lineal y lo convierte en valores en el rango cero a uno [0-1], la posición con el valor más alto contendrá la palabra asociada a dicha posición la cual producirá la salida de forma secuencial hasta obtener la secuencia completa <sup>[17][24]</sup>.

## CARACTERÍSTICAS DE LA ARQUITECTURA

Hay elementos que hacen que esta arquitectura represente una gran revolución en el mundo del procesamiento del lenguaje natural; área de la inteligencia artificial y lingüística, que desde sus inicios por allá en 1960, tuvo como objetivo la generación y comprensión del lenguaje natural y que a través de los años hizo uso de las redes neuronales y de las redes neuronales recurrentes; la cuales manejaban una ‘memoria’ a corto plazo, aspecto que impedía ser eficaces, esto con entradas de gran tamaño, lo anterior como consecuencia de su procesamiento secuencial como se mostró en la figura número tres. La arquitectura Transformer está generando una revolución principalmente porque posee las siguientes características:

Son arquitecturas paralelizables, procesan las n-entradas de forma simultánea

Son de aprendizaje profundo

Usan algoritmos de Machine Learning

Poseen memoria a largo plazo

Presentan menor tiempo de entrenamiento

Son más eficientes que las redes LSTM (Long Short-Term Memory networks)

Se usan con Datasets masivos

Al no ser seriales o secuenciales, requieren de un vector de posición.

Utilizan mecanismos de atención con lo que priorizan las secuencias de entrada

Aprovechan el potencial de las GPUs

## CASOS DE USO

Aunque la arquitectura Transformer fue utilizada originalmente para ser utilizada en NLP, poco a poco se han encontrado nuevas aplicaciones como las presentadas a continuación:

Se utilizan en el diseño de fármacos, pues facilitan la comprensión del ADN, los aminoácidos y las proteínas. Se han generado aplicaciones como AlphaFold2, MegaMolBART, GatorTron.

Adicionalmente su aplicación es exitosa en áreas como la Visión Artificial, las Redes Generativas (GANs), Aprendizaje por Refuerzo, Time Series, Machine Learning, Sumarización, traducción, en autos autónomos (Autopilot), responder preguntas, análisis de sentimientos, extracción de información, procesamiento de imágenes, reconocimiento de objetos y la ciberseguridad.

## TRANSFORMER Y CIBERSEGURIDAD

La seguridad informática es un área que demanda de profesionales con amplios conocimientos en herramientas, técnicas y modelos que garanticen el aseguramiento de la información en las compañías, por otra parte, la inteligencia artificial está revolucionando muchas áreas en donde la ciberseguridad no podía ser la excepción, pues con herramientas como el bot SecGPT, AICert, Hacker AI y CyberRiskAI, entre otras <sup>[26][27][32]</sup>; esta área entra de lleno a contribuir con esta causa. A continuación, un esbozo genérico de esta herramienta.

SecGPT: es un bot creado por Alter-AI <sup>[26]</sup> usado como un asistente virtual de inteligencia artificial diseñado específicamente para la ciberseguridad. Su objetivo es analizar grandes cantidades de datos y detectar patrones y anomalías que podrían indicar una posible amenaza; además, puede contribuir con tareas como la identificación de vulnerabilidades, la realización de pruebas de penetración y la gestión de incidentes de seguridad <sup>[28]</sup>. Se utiliza para:

- Buscar nuevas tendencias y patrones en estrategias de seguridad ofensiva
- Desarrollar técnicas de explotación enfocadas e identificar vulnerabilidades potenciales.
- Obtener una comprensión más profunda de las nuevas estrategias ofensivas y vías de ataque.
- Perfeccionar sus habilidades y técnicas en pruebas de penetración.
- Mejora la detección y análisis de amenazas
- Automatizar respuestas
- Fortalece los sistemas de detección y prevención
- Agiliza la gestión de incidentes

## ALGUNOS BOTS CONVERSACIONALES

Los Bots conversacionales más potentes y que son más populares y conocidos en la actualidad ver son <sup>[29]</sup>: figura No 17.

Megatron-Turing de Microsoft y Nvidia con 530.000 millones de parámetros <sup>[30][31]</sup>.

<https://www.genbeta.com/actualidad/microsoft-nvidia-lanzan-megatron-turing-potente-modelo-lenguaje-mundo-triple-parametros-que-gpt-3#:~:text=Es%20un%20modelo%20monol%C3%ADtico%20de,transformado%20entrenado%20hasta%20la%20fecha%22.>

GPT3 (Generative Pre-trained Transformer): de OpenAI con 175 .000 millones de parámetros

GPT4 (Generative Pre-trained Transformer): de OpenAI con 100 billones de parámetros

BERT (Bidirectional Encoder Representations from Transformers) de Google con 345.000 millones de parámetros

BARD (Bayesian Augmented Retrieval Device) de Google con 137.000 millones de parámetros

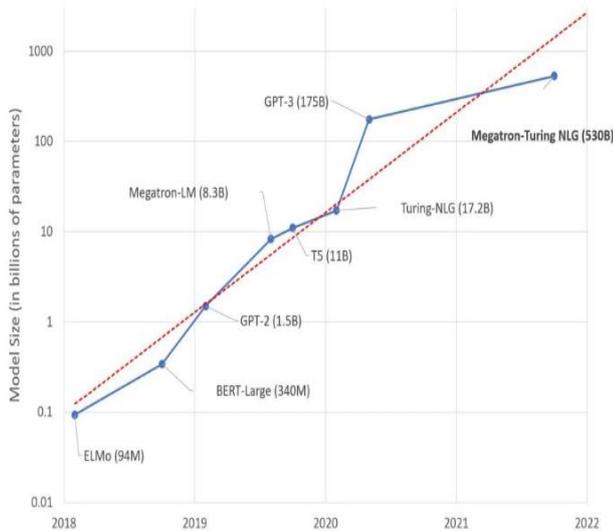


Figura No. 7. Los Transformer y su evolución

Fuente: <https://bloo.media/blog/mejores-herramientas-chatbot/>

## LENGUAJES Y BIBLIOTECAS DE PROGRAMACIÓN PARA LOS TRANSFORMER

Pytorch: es una biblioteca de código abierto, diseñada con Python y creada para proyectos de aprendizaje automático. Se especializa en diferenciación automática, cálculos de tensores y aceleración de GPU; su uso principalmente es en visión artificial y NLP.

Python: es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el Machine Learning.

TensorFlow: es la herramienta desarrollada por Google que ayuda a gestionar e implementar los procesos de aprendizaje automático.

Pandas: es una librería de Python especializada en la manipulación y el análisis de datos. Ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales

NumPy: es una biblioteca para el lenguaje de programación Python que da soporte para crear vectores y matrices grandes

multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas.

Keras: es una biblioteca de Redes Neuronales de Código Abierto escrita en Python. Es capaz de ejecutarse sobre TensorFlow, Microsoft Cognitive Toolkit o Theano. Está especialmente diseñada para posibilitar la experimentación en más o menos poco tiempo con redes de Aprendizaje Profundo.

Scikit-learn: es una biblioteca de aprendizaje automático de software gratuito para el lenguaje de programación Python.

## CÓMO CREAR UN TRANSFORMER EN PYTORCH

Como bien es sabido, los ChatBots de última generación, son multifuncionales; pues no solo son utilizados en NLP, sino que permiten generar aplicaciones software de una forma muy ágil y en el lenguaje de programación que se seleccione. En el presente trabajo se elaboró un 'Prompt' (instrucción o instrucciones que se dan a la IA para realizar una tarea) sencillo, con el cual se genera un esquema general de programa en donde se implementan los algoritmos y estructura presentados aquí previamente, veamos:

Bueno y ahora a programar...el Prompt para pedir a ChatGPT<sup>[25]</sup> me genere el código de un Transformer en el lenguaje Pytorch. En la actualidad hay un variado número de tutoriales en español para la creación de Transformers, la mayoría de herramientas está relacionada con Python, como es el caso de la herramienta elegida y se codifica utilizando Google Colab.

El Prompt fue:

Hola, me podría generar el código para crear un transformers sencillo en pytorch

Al cual me respondió:

¡Hola! Claro, puedo proporcionarte un ejemplo básico de cómo crear un modelo Transformers sencillo utilizando PyTorch. Aquí tienes un esqueleto de código que puedes seguir:

Ver código generado a continuación.

*/\* por espacio no se incluyó \*/*

Una vez generó el código, hizo las siguientes observaciones técnicas:

Este es un ejemplo básico de cómo construir un modelo Transformer en PyTorch. Ten en cuenta que este código es solo un esqueleto y puede requerir ajustes según tus necesidades y el problema que estés abordando. También es importante implementar la parte de entrenamiento y ajustar los hiperparámetros según tu tarea específica. ¡Espero que este código te ayude a comenzar!

Adicionalmente se le preguntó acerca de los hiperparámetros de los cuales hizo una lista enumerando aquellos parámetros que deben ser parametrizados de forma previa al entrenamiento del modelo.

Le agradecemos:

Muchas gracias

Y nos da el siguiente mensaje:

¡De nada! Si tienes más preguntas o necesitas más ayuda en el futuro, no dudes en preguntar. ¡Buena suerte con tu proyecto y tus experimentos con modelos Transformers en PyTorch!

De acuerdo a lo anterior se evidenció cómo los bots producen en cuestión de segundos información estructurada y coherente; que permite agilizar procesos, como en este caso del proceso de desarrollo de software; se deja a criterio del lector el comprobar la funcionalidad del código generado por ChatGPT.

## CONCLUSIONES

De acuerdo con lo planteado por Noam Chomsky<sup>[33]</sup> acerca de las prevenciones que ChatGPT ha generado alrededor del mundo y sobre el cual apuntó “Sospecho que el genio está fuera de la botella”, sugiriendo la dificultad para controlar un avance tecnológico de esta naturaleza, el cual su nivel de popularidad aumenta vertiginosamente y sus áreas de aplicación son cada día mayores. La situación actual nos lleva a adoptar estrategias que permitan generar herramientas y modelos con los cuales contrarrestar cualquier uso indebido de este tipo de tecnología, un ejemplo es lo que está sucediendo en el área de la ciberseguridad en donde rápidamente se está generando herramientas con las cuales mitigar las posibles amenazas provocadas por los ciberdelincuentes con la inteligencia artificial.

## Referencias

- [1] Procesamiento de lenguajes naturales [Internet]. Wikipedia 2023 [Consultada en abril de 2023] Disponible en: [https://es.wikipedia.org/wiki/Procesamiento\\_de\\_lenguajes\\_naturales](https://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales).
- [2] ¿Qué es una red neuronal? [Internet]. AWS 2023 [Consultada en enero 2023] Disponible en: <https://aws.amazon.com/es/what-is/neural-network/>
- [3] Bot conversacional [Internet]. Wikipedia 2023 [Consultada en enero 2023] Disponible en: [https://es.wikipedia.org/wiki/Bot\\_conversacional](https://es.wikipedia.org/wiki/Bot_conversacional)
- [4] ¿Qué es un chatbot? [Internet]. OCI Oracle Cloud 2023 [Consultada en febrero 2023] Disponible en: <https://www.oracle.com/co/chatbots/what-is-a-chatbot/>
- [5] Aprende Machine Learning [Internet]. Utiliza el poder de los LLMs 2023 [Consultada en enero 2023] Disponible en: <https://www.aprendemachinelearning.com/como-funcionan-los-transformers-espanol-nlp-gpt-bert/>
- [6] ¿Qué es un Modelo Transformer? [Internet]. NVIDIA 2023 [Consultada en enero 2023] Disponible en: <https://la.blogs.nvidia.com/2022/04/19/que-es-un-modelo-transformer/>
- [7] Computación paralela [Internet]. Wikipedia 2023 [Consultada en enero 2023] Disponible en: [https://es.wikipedia.org/wiki/Computaci%C3%B3n\\_paralela](https://es.wikipedia.org/wiki/Computaci%C3%B3n_paralela)
- [8] GPU's el futuro de la computación [Internet]. Janus Henderson 2023 [Consultada en enero 2023] Disponible en:

<https://www.janushenderson.com/es-es/investor/article/gpus-ware-driving-the-future-of-computing/>

- [9] Wolfram Language es un lenguaje simbólico [Internet]. Wolfram Computation Meet knowledge 2023 Consultada en febrero 2023] Disponible en: <https://www.wolfram.com/language/>
- [10] Telefónica Tech [Internet]. Telefónica 2023 [Consultada en marzo 2023] Disponible en: <https://telefonicatech.com/blog/>
- [11] Bots conversacionales [Internet]. Planeta Chatbot 2023 [Consultada en febrero 2023] Disponible en: <https://planetachatbot.com/tensores-son-todo-lo-que-necesitas/>
- [12] Aprendizaje automático [Internet]. Wikipedia 2023 [Consultada en enero 2023] Disponible en: [https://es.wikipedia.org/wiki/Aprendizaje\\_automatiko](https://es.wikipedia.org/wiki/Aprendizaje_automatiko)
- [13] Blog de Programación y Tecnología [Internet]. Keepcoding 2023 [Consultada en marzo 2023] Disponible en: <https://keepcoding.io/blog/>
- [14] Mecanismo de atención en aprendizaje profundo [Internet]. HashDork 2023 [Consultada en abril 2023] Disponible en: <https://hashdork.com/es/attention-mechanism-in-deep-learning/>
- [15] Vaswani, Ashish. Shazeer, Noam, et al. Attention Is All You Need. Google 2017.
- [16] ¿Cómo funcionan los Transformers? [Internet]. Aprende Machine Learning 2023 [Consultada en enero 2023] Disponible en: <https://www.aprendemachinelearning.com/como-funcionan-los-transformers-espanol-nlp-gpt-bert/>
- [17] Redes Transformer (... o el fin de las Redes Recurrentes) [Internet]. Codificandobits 2023 [Consultada en abril 2023] Disponible en: <https://www.codificandobits.com/blog/redes-transformer/>
- [18] Redes Transformer [Internet]. Codificandobits 2023 [Consultada en enero 2023] Disponible en: [https://www.youtube.com/watch?v=Wp8NocXW\\_C4](https://www.youtube.com/watch?v=Wp8NocXW_C4)
- [19] Artificial Intelligence [Internet]. Cornell University 2023 [Consultada en enero 2023] Disponible en: <https://arxiv.org/>
- [20] Modelos Transformers [Internet]. NVIDIA 2023 [Consultada en enero 2023] Disponible en: <https://la.blogs.nvidia.com/2022/04/19/que-es-un-modelo-transformer/>
- [21] Mecanismo de atención en aprendizaje profundo [Internet]. HashDork 2023 [Consultada en enero 2023] Disponible en: <https://hashdork.com/es/attention-mechanism-in-deep-learning/>
- [22] Traducción automática [Internet]. Ibidengroup 2023 [Consultada en enero 2023] Disponible en: <https://www.ibidemgroup.com/edu/traduccion-automatiga-gpt3-bert-t5/#:~:text=Un%20Transformer%20es%20un%20tipo,para%20diferentes%20tipos%20de%20datos.>
- [23] ¿Qué Es El Mecanismo De Atención? [Internet]. Hash Dork 2023 [Consultada en enero 2023] Disponible en: <https://hashdork.com/es/attention-mechanism-in-deep-learning/>
- [24] Las matemáticas del machine learning [Internet]. Telefónica 2023 [Consultada en Mayo 2023] Disponible en: <https://telefonicatech.com/blog/las-matematicas-del-machine-learning-funciones-de-activacion>

[25] ChatGPT [Internet]. OpenIA 2023 [Consultada en enero 2023]  
Disponible en: <https://chat.openai.com/chat>

[26] SegGPT [Internet]. Huggin Face 2023 [Consultada en mayo 2023]  
Disponible en: <https://hayunaiaparaeso.com/ia/alter-ai-secgpt/>

[27] SecGPT transforms cybersecurity through AI-driven insights.  
[Internet]. Medium 2023 [Consultada en enero 2023] Disponible en:  
<https://infosecwriteups.com/secgpt-transforms-cybersecurity-through-ai-driven-insights-c5074c90bee1>

[28] La inteligencia artificial está revolucionando la ciberseguridad  
[Internet]. ZOE 2023 [Consultada en enero 2023] Disponible en:  
<https://zoeweb.es/secgpt-la-inteligencia-artificial-que-esta-revolucionando-la-ciberseguridad/>

[29] Guía completa de chatbots de IA [Internet]. Zendesk 2023  
[Consultada en mayo 2023] Disponible en:  
<https://www.zendesk.com.mx/service/messaging/chatbot/>

[30] Build Applications With Generative AI [Internet]. Nvidia  
Developer 2023 [Consultada en abril 2023] Disponible en:  
<https://developer.nvidia.com/megatron-turing-natural-language-generation>

[31] Los mejores chabots para la web [Internet]. Bloo Mmedia 2023  
[Consultada en enero 2023] Disponible en:  
<https://bloo.media/blog/mejores-herramientas-chatbot/>

[32] Inteligencia Artificial Para la Ciberseguridad [Internet]. David  
Pereira 2023 [Consultada en marzo 2023] Disponible en:  
<https://www.youtube.com/watch?v=sH05p8paBOY> atención

[33] Noam Chomsky sobre ChatGPT [Internet]. Rebelión 2023  
[Consultada en junio 2023]. C J Polychroniou Disponible en:  
<https://rebellion.org/noam-chomsky-sobre-chatgpt/>