UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA DE COMPUTACIÓN

Evaluación de ChatGPT en la Generación Automática de Resúmenes de Artículos Científicos de Conferencias

Trabajo de titulación previo a la obtención del
Título de Ingenieros en Ciencias de la Computación

AUTORES: ERICK BLADIMIR RUEDA PABÓN
BRAYAN STIVEN VANEGAS RODRIGUEZ
TUTORA: PAULINA ADRIANA MORILLO ALCIVAR

Quito - Ecuador
2024

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Erick Bladimir Rueda Pabón con documento de identificación N.º 1501093460 y Brayan Stiven Vanegas Rodriguez con documento de identificación N.º 1757756760; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 29 de febrero del 2024

Atentamente,

Erick Bladimir Rueda Pabón
1501093460

Brayan Stiven Vanegas Rodriguez
1757756760

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotros, Erick Bladimir Rueda Pabón con documento de identificación No. 1501093460 y Brayan Stiven Vanegas Rodriguez con documento de identificación No. 1757756760, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Artículo Académico: "Evaluación de ChatGPT en la generación automática de resúmenes de artículos científicos de conferencias", el cual ha sido desarrollado para optar por el título de: Ingenieros en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 29 de febrero del 2024

Atentamente,

Erick Bladimir Rueda Pabón
1501093460

Brayan Stiven Vanegas Rodriguez
1757756760

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Paulina Adriana Morillo Alcívar con documento de identificación N° 1715646574, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: EVALUACIÓN DE CHATGPT EN LA GENERACIÓN AUTOMÁTICA DE RESÚMENES DE ARTÍCULOS CIENTÍFICOS DE CONFERENCIAS, realizado por Erick Bladimir Rueda Pabón, con documento de identificación N.º 1501093460 y por Brayan Stiven Vanegas Rodriguez con documento de identificación N.º 1757756760, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 29 de febrero del 2024

Atentamente,

Ing. Paulina Adriana Morillo Alcívar, MSc
1715646574

Evaluación de ChatGPT en la Generación Automática de Resúmenes de Artículos Científicos de Conferencias

1st Erick Bladimir Rueda Pabón
eruedap@est.ups.edu.ec

2nd Brayan Stiven Vanegas Rodriguez
bvanegasr@est.ups.edu.ec

3rd Paulina Adriana Morillo Alcívar
pmorillo@ups.edu.ec

***Resumen*—ChatGPT es una herramienta de inteligencia artificial que permite la generación de texto por medio de algoritmos de aprendizaje profundo que a su vez permiten generar contenido multimedia. ChatGPT se ha aplicado a diversas áreas de estudio como la educación, la medicina, entre otras y actualmente su capacidad ha sido aplicada en la redacción de artículos científicos. En este trabajo se realiza una comparación de resúmenes de artículos científicos, que han sido escritos por humanos y resúmenes que han sido generados con chatGPT en base a los títulos de los artículos previamente escritos. Para cumplir este objetivo se utilizan los resúmenes de artículos científicos de las conferencias AAAI 2013, AAAI 2014, ICMLA 2014 y ICMLA 2015 que son conjuntos de datos depositados en *Machine Learning Repository* de la Universidad de California de Irving (UCI) y de la base de datos de Mendeley. En total se compararon 784 resúmenes originales comparados con la misma cantidad de resúmenes generados por chatGPT. El tiempo de generación de cada resumen fue en promedio de 3.107 segundos. Para la comparación de similitud, se usaron 4 métricas Coseno, Jaccard, Sorensen-Dice y Overlap. La media de estos valores fue de 0.795, 0.631, 0.758 y 0.83, para cada métrica, respectivamente. Aunque los resultados no permiten asegurar si existe o no similitud completa entre los resúmenes, se puede observar que en algunos casos la similitud entre los resúmenes generados por chatGPT y los resúmenes originales es alta.**

***Palabras Clave*—ChatGPT, Procesamiento del Lenguaje Natural, pruebas de similitud, Coseno Vectorial, Jaccard, Sorensen-Dice, Overlap.**

***Abstract*—ChatGPT is an artificial intelligence tool that enables text generation through deep learning algorithms, allowing for the creation of multimedia content. It has found applications in various fields, including education, medicine, and, most recently, in the composition of scientific articles. This study involves a comparison between human-written summaries of scientific articles and summaries generated by ChatGPT based on the titles of previously written articles. To achieve this goal, abstracts of scientific articles from the AAAI 2013, AAAI 2014, ICMLA 2014, and ICMLA 2015 conferences were utilized. These datasets are available in the Machine Learning Repository of the University of California, Irvine (UCI), and the Mendeley database. A total of 784 original summaries were compared with an equal number of summaries generated by chatGPT. The average generation time for each summary was 3,107 seconds. For the similarity comparison, four metrics—Cosine, Jaccard, Sørensen-Dice, and Overlap—were utilized. The mean values for these metrics were 0.795, 0.631, 0.758, and 0.83, respectively. While the results do not allow us to ascertain complete similarity between the summaries, it can be observed that, in some cases, the similarity between the chatGPT-generated summaries and the original summaries is high.**

***Keywords*—ChatGPT, Natural Language Processing, similarity tests, Vectorial Cosine, Jaccard, Sorensen-Dice, Overlap.**

I. Introducción

A finales de noviembre de 2022, OpenAI lanzó su herramienta gratuita ChatGPT1, mostrando la capacidad de los modelos de inteligencia artificial (IA) para generar contenido textual de una forma natural. En poco tiempo, se publicaron diversos artículos que exploraban sus posibles aplicaciones y las controversias que podrían surgir a raíz de su uso [1].

ChatGPT es una herramienta tecnológica de Inteligencia Artificial desarrollada por medio de algoritmos de aprendizaje profundo, *Generative Pre-trained Transformer (GPT)*, que son una familia de modelos de redes neuronales que utilizan la arquitectura de transformadores, estos permiten generar texto, imágenes y música de manera muy similar a como lo haría un humano [2].

Los usuarios iniciales de esta herramienta han compartido sus experiencias en plataformas de redes sociales, expresando principalmente sentimientos positivos [3]. El uso de chatGPT en diversas actividades humanas ha tenido un impacto significativo en la sociedad debido a su capacidad para simular las conversaciones humanas y generar respuestas acertadas en función a las peticiones realizadas [4].

chatGPT se ha aplicado en diversos campos como la educación, la medicina, la investigación, etc. Se ha comprobado que es capaz de generar respuestas con altas puntuaciones, es decir, que el desempeño de chatGPT ha sido evaluado positivamente en la generación de texto como la generación de expresiones de pensamiento crítico [5], mantener la coherencia en el contexto de la conversación e incluso ayudar en la generación de código en diferentes lenguajes de programación [6]. Aunque el uso de chatGPT tiene ventajas como el acceso a información relevante, la generación de contenido escrito y la traducción de texto en diferentes idiomas, es crucial supervisar que su uso sea adecuado y que no sobrepase los límites éticos y morales [7].

En el ámbito académico y de investigación científica, el acceso a internet y a herramientas tecnológicas de IA han provocado cambios en la manera de transmitir y generar información científica de forma escrita [8]. A nivel mundial, la producción científica ha tenido un crecimiento significativo en

las últimas décadas, impulsada por avances tecnológicos, colaboraciones internacionales y una creciente conciencia sobre la importancia de la investigación para abordar desafíos globales. En el 2019, la producción científica representó el 2.6 % en investigación y desarrollo a nivel global. En el Ecuador, fueron 17182 artículos en la revista Scopus, en los años 2019, 2020 y 2021, registrando un crecimiento de 30.3 % en su producción científica en los últimos años [9].

Debido a que toda investigación culmina con una publicación, ya sea en revista, conferencia, congreso, etc, es fundamental que este documento sea claro, preciso, y con resultados confiables y replicables. Por esta razón, los artículos científicos están sujetos, en la mayoría de los casos, a una revisión por pares, donde expertos en el campo evalúan la calidad, relevancia y originalidad del artículo. La fase de revisión del artículo se realiza con el fin de mejorar el trabajo y suele llevar varias semanas. Sin embargo, antes y después de la aceptación de un artículo, la edición y escritura, son las tareas que suelen tomar mayor tiempo dentro del proceso de publicación [10]. En promedio la redacción del informe puede extenderse a lo largo de semanas e incluso meses, especialmente cuando se abordan temas complejos [11]. Para reducir este tiempo, algunos autores se apoyan de herramientas de generación automática de texto como chatGPT.

No obstante, la generación automática de texto presenta algunos problemas como la confiabilidad, coherencia y precisión en los contenidos producidos, sumado a la dificultad para evaluar la calidad del texto generado. Por lo tanto, se presenta cierta incertidumbre en los resultados de la herramienta para llevar a cabo investigaciones más serias y para delegar completamente la redacción final del informe a chatGPT [12].

Este trabajo pretende analizar la similitud de resúmenes de artículos científicos publicados en conferencias científicas de inteligencia artificial y reconocimiento de patrones, con resúmenes generados por ChatGPT únicamente a través de los títulos del artículo original. La comparación de similitud entre los resúmenes generados de forma manual y de forma automática servirá para evaluar la capacidad de comprensión y síntesis de información técnica y científica por parte de chatGPT. De esta forma, en la Sección II se muestra lo metodología empleada en generación automática de los resúmenes y la comparación con los resúmenes originales. En la sección III se exponen los resultados experimentales de este proyecto. En la sección III-B se realiza una discusión de los resultados. Finalmente, en la sección IV se presentan las conclusiones y trabajo futuro de este proyecto.

I-A. Trabajos Relacionados

Existen varias investigaciones y artículos relacionados con chatGPT en diferentes áreas como: la medicina, la educación, programación, etc. En el caso de este trabajo interesa analizar aquellos que se enfocan en la evaluación de esta herramienta en diversos ámbitos de generación de texto. Por ejemplo,

Benichou en su investigación [13] usó 10 artículos científico-médicos, los cuales están compuestos de 5 secciones: introducción, materiales y métodos, resultados, discusión y conclusión. La idea del autor fue hacer una comparación entre las secciones escritas naturalmente y las secciones creadas por chatGPT. El enfoque de la investigación era evaluar la claridad, precisión y coherencia de la producción de texto de chatGPT. Los resultados arrojaron que el texto generado en todas las secciones fue claro, preciso y consistente [13], por lo cual, el texto generado por chatGPT tuvo una calidad comparable con el texto producido por un experto humano. En cuanto al tiempo de creación de cada sección de un artículo, a cada autor le tomó 10 minutos en promedio, mientras que a chatGPT tan solo 30 segundos.

Otro trabajo importante es el estudio realizado por Samaan et al [14] que evalúa la precisión de las respuestas de chatGPT frente a varias preguntas sobre la cirugía bariátrica. Se realizaron 151 preguntas con cuatro categorías de respuesta, integral, correcta pero inadecuada, algunas correctas y otras incorrectas y completamente incorrectas. La verificación de la precisión de las respuestas de chatGPT se realizó con un equipo de expertos cirujanos bariátricos quienes realizaron una calificación manual. Los resultados muestran que chatGPT proporcionó el 86.8 % de las preguntas como integrales quedando el 13.2 % de las preguntas entre las categorías: correcta pero inadecuada, algunas correctas y otras incorrectas y completamente incorrectas

Gao et al [15] en su investigación, realizó una comparación de resúmenes científicos generados por chatGPT y resúmenes reales. Se recopilaron 50 resúmenes de artículos publicados en cinco revistas de medicina y se solicitó a chatGPT por medio de un *script* que generara resúmenes en base a sus títulos. Se usó *GPT-2 Output Detector* que encontró una alta probabilidad de que los resultados generados sean falsos, por otro lado, se usaron 2 sitios web para la detección de plagio, donde se evidenció el 62.5 % los resúmenes reales no contenían textos originales, mientras que los resúmenes generados por la herramienta obtuvieron el 0 % de similitud con otros resúmenes previamente publicados. Además, de estos experimentos también se realizó una clasificación manual entre resúmenes reales y generados, por parte de un equipo humano. En este caso, el 68 % de los resúmenes generados fueron clasificados correctamente, mientras que el 86 % de los resúmenes reales fueron clasificados correctamente.

En un contexto académico, la falta de originalidad en las publicaciones es una preocupación constante y la aparición de *chatbots*, como chatGPT ha aumentado el número de textos cuya originalidad no es del autor, ni se desprende del trabajo realizado por los investigadores. Por esta razón en [16] se plantea una evaluación de la autenticidad de las respuestas de chatGPT para generar respuestas novedosas, coherentes y precisas que evadan la detección de coincidencias de texto. Para llevar a cabo esta evaluación se inició realizando una pregunta a chatGPT y regenerando la misma pregunta 2 veces más con el fin de verificar la repetibilidad en la generación de nuevas respuestas. Este proceso se repitió 30 veces con el modelo de *ChatGPT 3.5* y otras 15 veces con *chatGPT4* demostrando así que estos modelos pueden generar respuestas con entre el 10 % y 25 % de coincidencia de texto, observando

que el modelo de *chatGPT 4* tiene una mayor capacidad para generar respuestas más auténticas. Para la comparación de las respuestas se usaron los índices de capacidad *Process Performance Index (PPK)* y *parts per million (PPM)* que brindan la información sobre el desempeño de un proceso al evaluar su capacidad para cumplir con las especificaciones.

De acuerdo con la revisión de la literatura, chatGPT es una herramienta de IA con gran capacidad para generar texto. Esto podría facilitar las tareas de redacción de artículos académicos, rediciendo el tiempo de edición y producción de un *paper*. En este escenario, este artículo propone la comparación de la similitud entre los resúmenes generados por chatGPT y resúmenes de *papers* reales publicados en los *proceedings* de conferencias científicas. El objetivo es realizar la comparación, a través del uso de técnicas de procesamiento natural del lenguaje (NLP) y métricas de similitud de textos.

II. Materiales y Métodos

La comparación de resúmenes académicos generados por ChatGPT 3.5 y resúmenes de artículos académicos reales empieza con la recopilación de los conjuntos de resúmenes de *papers* presentados en conferencias científicas del área de informática y ciencias de la computación. Luego, se utiliza los títulos de los artículos originales para generar los resúmenes artificiales con la herramienta ChatGPT 3.5. En segundo lugar, se realiza el procesamiento de lenguaje natural para depurar los resúmenes. En tercer lugar, se comparan los resúmenes, a través del cálculo de las métricas de similitud. Finalmente, se realizan pruebas estadísticas para comparar los resultados 1.

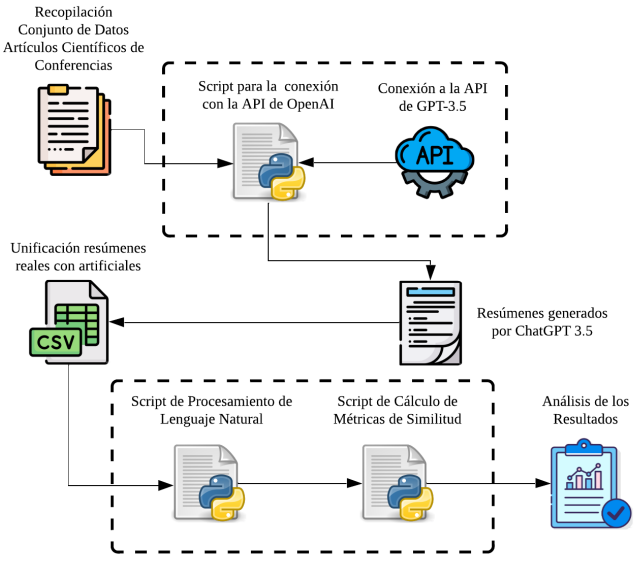

Figura 1: Metodología para realizar la comparación entre los resúmenes reales y artificiales.

II-A. Recopilación y generación de resúmenes de artículos científicos de conferencias

Los resúmenes de los artículos se recopilan del repositorio de conjuntos de datos para aprendizaje automático *Machine Learning Repository* de la Universidad de California de Irving (UCI) y de la base de datos de Mendeley. Los conjuntos de datos seleccionados son: AAAI 2013, AAAI 2014 [17], ICMLA 2014 y ICMLA 2015 publicados por [18].

Los *papers* son artículos académicos publicados en las conferencias: *Association for the Advancement of Artificial Intelligence* (AAAI) e *International Conference on Machine Learning and Applications* (ICMLA). En el caso de ICMLA, la información se recopila de Mendeley, mientras que para AAAI, se obtiene de *Machine Learning Repository* UCI. De cada artículo se considera el título y el resumen. Los resúmenes originales de todas las conferencias se recopilan en un solo archivo de extensión .csv.

Después de la recopilación de los resúmenes de los artículos científicos de conferencias se construyen los *prompts* para generar los resúmenes en ChatGPT 3.5 a través de su API. La estructura del *prompt* utilizado para la generación automática del resumen se muestran en el Pseudocódigo 1. Como se observa, para generar el resumen artificial se realiza la misma pregunta cambiando el título del artículo, el número máximo de palabras del resumen solicitado se establece a través del valor asociado en la variable *Abstract[i]*, la cual corresponde al número de palabras del *Abstract* original.

Algoritmo 1 Generación de resúmenes utilizando ChatGPT 3.5

```
titles ← data[Title]
Abstract ← data[Abstract]
Keywords ← data[Keywords]
res ← []
i ← 0
while i < longitud(titles) do
    pregunta ← "Generar resumen para el artículo científico 'titles[i]'. Debe ser máximo (lenAbstract[i]) palabras. Sin saltos de línea, un solo párrafo."
    completion ← openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[{role": "system", çontent": pregunta}]
    respuesta ← completion.choices[0].message.content
    res.append(respuesta)
    imprimir(titles[i])
    i ← i + 1
end while
df ← pd.DataFrame(res)
df.to_csv(nombre_dataset_gpt.csv)
```

II-B. NLP aplicado a los resúmenes artificiales y reales

Luego de la recopilación de los resúmenes reales y la generación de los resúmenes artificiales, se lleva a cabo el procesamiento clásico de NLP. Este proceso empieza con el reemplazo de los caracteres no alfanuméricos por espacios en blanco, para depurar el texto y eliminar caracteres que no aportan significado al texto. Posteriormente, se cambia todo el texto a minúsculas y se realiza la tokenización, segmentando el documento en una lista de palabras individuales, conocidas

como tokens. En tercer lugar, se define un conjunto de palabras vacías en inglés; luego, se procede a filtrar las palabras del documento, eliminando aquellas que pertenecen a dicho conjunto. Finalmente, se implementa el algoritmo de Porter para llevar a cabo el *stemming*, es decir, reducir cada palabra a su forma base o raíz [19].

Con el conjunto de tokens resultante del proceso de NLP, se construye una bolsa de palabras aplicando la técnica *Term Frequency-Inverse Document Frequency* (TF-IDF) [20]. El uso del TF-IDF se realiza para evaluar la importancia de una palabra en el resumen real y el resumen artificial. Considerando su frecuencia y rareza de forma individual y en conjunto. Esta medida es usada para el cálculo de la métrica de coseno vectorial, ya que permite definir los atributos como coordenadas reales de un vector conformado por los valores de la matriz TF-IDF, teniendo en cuenta la importancia relativa de los términos en cada resumen.

Por otro lado en el caso de Jaccard, Sorensen-Dice y Overlap se usó una matriz binaria para el calculo del nivel de similitud. Lo que realiza esta matriz es una comprobación que verifica si la palabra contenida en el diccionario se encuentra en el resumen real y el artificial, tomando el valor cero si no hay coincidencia y el valor uno si la hay.

II-C. Métricas de Similitud

Para evaluar la similitud entre los resúmenes originales y los resúmenes artificiales se consideran las siguientes métricas Índice de Jaccard (*J*), Coeficiente del Coseno Vectorial (*CV*), Coeficiente de Sørensen-Dice (*QS*) y Overlap (*O*).

El índice de Jaccard [21] varía entre 0 y 1, donde 1 implica similitud exacta entre los resúmenes. Por el contrario, si los resúmenes son diferentes casi en su totalidad el valor esperado sería cercano a cero. Para calcular el índice Jaccard se usa la Ecuación (1), donde R_i hace referencia a cada resumen real con su conjunto de tokens, mientras que A_i representa cada resumen artificial con su respectivo conjunto de tokens. La intersección de R_i y A_i implica el número de tokens que coinciden en ambos resúmenes, este resultado se divide para la unión de R_i y A_i, es decir, se divide para el número total de tokens presentes en R_i y A_i.

$$J(R_i, A_i) = \frac{|R_i \cap A_i|}{|R_i \cup A_i|} \tag{1}$$

El Coeficiente de Coseno Vectorial [22] por su parte mide la similitud entre dos vectores en un espacio multidimensional. El coeficiente coseno vectorial toma valores entre -1 y 1. Sin embargo, por la naturaleza de los vectores formados por las columnas de la matriz TF-IDF, este rango se limita de cero a uno. En este contexto, un valor cercano a 1 indica una similitud alta entre los dos resúmenes, mientras que un valor de cero denota ausencia de similitud.

La Ecuación (2) muestra como se realiza su calculo. En este caso, cada resumen R_i y A_i se denotan vectorialmente como $\overrightarrow{R_i}$ y $\overrightarrow{A_i}$ cuyas coordenadas representan los tokens presentes en los dos resúmenes a comparar y los valores de las coordenadas son tomados de la matriz TF-IDF de los resúmenes reales y artificiales, respectivamente.

$$cos(\overrightarrow{R_i}, \overrightarrow{A_i}) = \frac{\overrightarrow{R_i} \cdot \overrightarrow{A_i}}{|\overrightarrow{R_i}||\overrightarrow{A_i}|} \tag{2}$$

La medida de similitud del coeficiente de Sørensen-Dice [23] es muy similar al índice Jaccard. La fórmula asociada (Ecuación (3)) produce resultados en un rango de 0 a 1, donde 1 representa una similitud perfecta y 0 indica la falta de similitud. En el cálculo de esta métrica el numerador es igual al doble de la intersección de R_i y A_i, es decir, el doble de la cantidad de tokens comunes entre ambos conjuntos, este resultado se divide para la suma de los tamaños de R_i y A_i, ambos valores son extraídos de la matriz binaria asociada a los resúmenes reales y artificiales, respectivamente.

$$QS(R_i, A_i) = \frac{2|R_i \cap A_i|}{|R_i| + |A_i|} \tag{3}$$

La métrica de Overlap [24] también varía en el rango de 0 a 1, y su interpretación es igual a los índices de Jaccard y Sørensen-Dice. Para calcular la métrica de Overlap se usa la Ecuación (4), donde R_i hace referencia a cada resumen real con su conjunto de tokens, mientras que A_i representa cada resumen artificial con su respectivo conjunto de tokens. En el numerador se contabilizan el número de tokens comunes a ambos resúmenes y en el denominador el tamaño del conjunto de tokens más pequeño entre R_i y A_i, estos valores se extraen de la matriz binaria correspondiente a los resúmenes reales y artificiales.

$$O(R_i, A_i) = \frac{|R_i \cap A_i|}{\text{mín}(|R_i|, |A_i|)} \tag{4}$$

III. Experimentos y resultados

Los experimentos se realizaron por medio de la plataforma Google Colab, con Python V3, en una máquina virtual con 12.67 GB de RAM de sistema, 107.72 GB de almacenamiento en disco. El código y el conjunto de datos utilizado en esta investigación se encuentra disponible en el siguiente enlace https://github.com/pure0527/ChatGPT$_{S}imilarity$. Para establecer la conexión con ChatGPT 3.5, se implementó la interfaz API de la herramienta mediante la transmisión de objetos JSON. En este proceso, se incorporaron dos claves fundamentales: la primera, denominada *role* y asignada el valor *system*, cumple la función de identificación; la segunda, bajo el nombre *content*, refleja el *prompt* utilizado para la interacción. Cabe destacar que se optó por el modelo específico gpt-3.5-turbo para llevar a cabo la generación de resúmenes [25].

III-A. Conjuntos de datos

La descripción de los resúmenes originales y los artificiales, se muestra en la Tabla I.

Como se observa se recopilaron 784 resúmenes de las cuatro conferencias seleccionadas: AAAI 2013, AAAI 2014, ICMLA 2014 y ICMLA 2015, el promedio de número de palabras en los resúmenes originales fue de 153.327, mientras que el

Nombre de conferencia	Tipo	N. palabras promedio de resúmenes	Tiempo promedio de generación (segundos)	N. articulo por conferencia	N. palabras mínimo y máximo de resúmenes
AAAI 2013	Real	173	2.864	150	[63 - 343]
	Artificial	142			[63 - 226]
AAAI 2014	Real	167	2.783	398	[50 - 305]
	Artificial	139			[44 - 267]
ICMLA 2014	Real	162	2.977	105	[56 - 312]
	Artificial	128			[45 - 236]
ICMLA 2015	Real	181	3.806	131	[62 - 287]
	Artificial	145			[55 - 281]
Total	Real	170	3.107	784	[58 - 312]
	Artificial	138			[52 - 253]

Tabla I: Descripción de los resúmenes originales y los artificiales.

promedio de palabras generadas en los *abstracts* artificiales fue de 125.346.

El tiempo promedio en segundos de generación de los resúmenes por ChatGPT 3.5 fue de 3.107 segundos. Tanto para los resúmenes artificiales como para los resúmenes originales el número de palabras del título es el mismo.

Luego de la aplicación de NLP a cada resumen se obtuvieron los resultados de la Tabla II.

Se observa una reducción del 39 % del número de palabras en los resúmenes originales versus el 37 % de los resúmenes artificiales. Con esto podemos llegar a la conclusión que en los resúmenes reales se usan más *stopwords* que en los resúmenes generados por *ChatGPT 3.5*.

III-B. Similitud entre resúmenes

En la Figura 2, se representa la distribución de probabilidad de las métricas de similitud. Al observar la métrica de Coseno Vectorial, se identifica un distribución bimodal que revela dos picos bien diferenciados. Esta característica sugiere la presencia de dos grupos distintos de resúmenes, posiblemente reflejando variaciones en términos de estilos de redacción o enfoques temáticos. El valle entre los picos señala una zona donde la similitud disminuye antes de volver a aumentar, indicando una separación clara entre los dos grupos identificados. La distribución de la similitud se centra mayoritariamente alrededor de la media de 0.795, mientras que la dispersión tiene una ligera asimetría hacia la derecha, expresada por la desviación estándar de 0.116, refleja la variabilidad en las similitudes. El rango [0.479 - 1.000] presenta valores excepcionales en la cola superior, indicando similitudes notables entre algunos resúmenes.

En el contexto del Índice de Jaccard, la distribución muestra un configuración bimodal con la presencia de dos picos distintivos, indicando la existencia de dos grupos bien diferenciados de resúmenes con similitudes marcadas. El valle entre los picos señala una clara separación entre estos dos conjuntos, evidenciando una disminución en la similitud antes de volver a aumentar su pico. Esta estructura bimodal puede sugerir disparidades en los estilos de redacción o enfoques temáticos, destacando dos conjuntos claramente definidos. La distribución del Índice de Jaccard presenta una ligera asimetría hacia la derecha, donde la media 0.631 se desplaza sutilmente a la derecha de la mediana 0.579. Con una desviación estándar de 0.193, la dispersión es moderada. El rango [0.185 - 1.000] muestra la presencia de valores extremos en la cola superior, indicando similitudes notables entre algunos resúmenes.

El Coeficiente de Sørensen-Dice presenta una distribución mayormente unimodal, destacando un pico central notable que indica una mayor homogeneidad en las similitudes entre los resúmenes. Aunque carece de la distinción de dos grupos como en las métricas bimodales, la presencia de un único pico sugiere una coherencia más pronunciada en los estilos de redacción. Se observa un valle más sutil, indicando áreas donde la similitud disminuye antes de regresar a niveles más altos, lo que puede sugerir una consistencia en las relaciones entre los resúmenes, aunque menos marcada que en las métricas bimodales. La distribución del Coeficiente de Sørensen-Dice es más simétrica, centrada en la media de 0.758. La dispersión más estrecha, caracterizada por una desviación estándar de 0.138, indica una consistencia más marcada en las similitudes. El rango [0.312 - 1.000] sugiere que la mayoría de las similitudes se encuentran en la parte superior de la distribución, resaltando una coherencia en las relaciones.

Finalmente, en el caso del Índice de Overlap, la distribución presenta una configuración bimodal similar al Coseno Vectorial, caracterizado por dos picos distintos que indican la existencia de dos conjuntos bien definidos de resúmenes. El valle entre los picos resalta una separación clara entre estos conjuntos, donde la similitud disminuye antes de observar un nuevo aumento. Este patrón bimodal sugiere la presencia de diferentes enfoques o temáticas en los resúmenes, destacando dos conjuntos claramente diferenciados. La distribución del Índice de Overlap muestra una concentración robusta alrededor de la media de 0.83, con una baja dispersión representada por la desviación estándar de 0.109. El rango [0.488 - 1.000] revela la presencia de valores extremos en la cola inferior, indicando variabilidad en las similitudes. Este patrón sugiere que la mayoría de los resúmenes comparten similitudes destacadas, pero algunos casos excepcionales presentan diferencias más notables.

III-C. Comparación de las métricas de similitud

Al analizar la similitud entre los resúmenes reales de 784 artículos y aquellos generados automáticamente por ChatGPT 3.5, se realizaron cálculos estadísticos que incluyeron la media, la desviación estándar, la mediana, el primer cuartil, el tercer cuartil y los rangos mínimos y máximos. Estos cálculos se basaron en las métricas de similitud previamente descritas, y los resultados detallados se presentan en la Tabla III. Se observa que el Coseno Vectorial muestra una media de 0.795, con una desviación estándar de 0.116, una mediana de 0.777 y cuartiles que indican una variabilidad moderada. El Índice de Jaccard muestra una media ligeramente superior de 0.631,

Nombre de conferencia	Tipo	N. promedio *tokens* sin caracteres especiales	N. promedio *stopwords*	N. promedio después de steaming
AAAI 2013	Reales	159	97	66
	Artificiales	133	83	58
AAAI 2014	Reales	154	95	65
	Artificiales	130	82	58
ICMLA 2014	Reales	151	93	63
	Artificiales	119	76	54
ICMLA 2015	Reales	167	104	71
	Artificiales	134	86	60
Total	Reales	158	97	66
	Artificiales	129	81	57

Tabla II: Resultados del NLP aplicado a los resúmenes originales y artificiales.

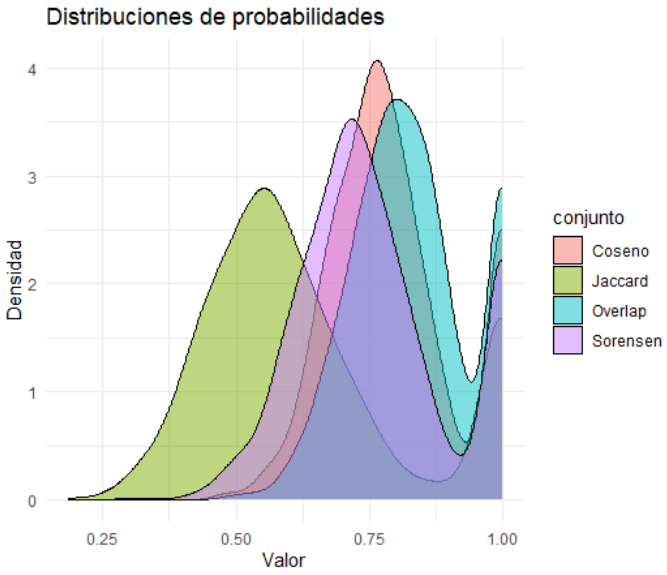

Figura 2: Distribuciones de probabilidad de las métricas de similitud.

con una desviación estándar de 0.193 y una mediana de 0.579, indicando una distribución más centrada. El Coeficiente de Sørensen-Dice presenta una media de 0.758, una desviación estándar de 0.138 y una mediana de 0.734, sugiriendo una similitud más consistente entre los conjuntos de datos. Por último, la métrica Overlap muestra la mayor similitud con una media de 0.83, una desviación estándar de 0.109 y una mediana de 0.823, reflejando una distribución más concentrada hacia similitudes cercanas a 1.

Medida de Similitud	μ	σ	M_e	$1stQ$	$3rdQ$	Min - Max
Coseno Vectorial	0.795	0.116	0.777	0.713	0.858	[0.479 - 1.000]
Índice de Jaccard	0.631	0.193	0.579	0.5	0.711	[0.185 - 1.000]
Coeficiente de Sørensen-Dice	0.758	0.138	0.734	0.667	0.831	[0.312 - 1.000]
Overlap	0.83	0.109	0.823	0.754	0.9	[0.488 - 1.000]

Tabla III: Análisis estadístico de los valores de las métricas.

IV. Conclusiones

En esta investigación, se evaluó la similitud entre resúmenes de artículos científicos publicados en conferencias con los resúmenes generados por chatGPT, con esto se recopilaron cuatro conjuntos de datos que incluían títulos y resúmenes de trabajos académicos. Utilizando *prompts* adecuados, se generaron resúmenes artificiales a partir de los títulos. Posteriormente, se aplicaron métricas de evaluación con el objetivo de determinar el grado de similitud entre los resúmenes producidos por chatGPT y aquellos resúmenes elaborados manualmente por los respectivos autores.

La comparación de similitud automática en la evaluación de resúmenes generados por modelos como chatGPT ofrece eficiencia, objetividad y consistencia en el análisis. Proporciona métricas cuantificables para evaluaciones rápidas y consistentes, eliminando la subjetividad asociada con la evaluación manual. Además, estas métricas ofrecen una base sólida para análisis estadísticos, y la retroalimentación resultante contribuye a mejoras continuas en la generación de texto.

En futuras investigaciones, se sugiere la exploración de la integración de elementos adicionales, tales como palabras clave y tópicos, con el objetivo de enriquecer la generación automática de resúmenes. La evaluación de modelos más recientes, como GPT-4 u otras alternativas de Google, ofrecería la oportunidad de comparar y mejorar la eficacia en la captura de la esencia de la información. Además, la adaptación de modelos a dominios específicos, como la medicina, podría resultar en resúmenes más precisos y contextualmente relevantes.

Se propone realizar pruebas de hipótesis, como análisis de varianza (ANOVA), para comparar los resultados con grupos artificiales generados mediante técnicas de *clustering*. En este contexto, se podría considerar el uso de métodos estadísticos como las tablas Tukey para llevar a cabo una evaluación más detallada de las diferencias entre estos grupos, proporcionando así un análisis más robusto y significativo.

Esta adaptación estratégica podría mejorar la aplicabilidad práctica de la generación automática de texto, contribuyendo a un avance significativo en la calidad y utilidad de los resúmenes generados.

Referencias

[1] S. Shankland, "Chatgpt: Why everyone is obsessed this mind-blowing ai chatbot," *CNET https://www. cnet. com/tech/computing/chatgpt-why-everyone-is-obsessed-this-mind-blowing-ai-chatbot*, 2022.

[2] M. Lahtela and P. P. Kaplan, "Gpt." [Online]. Available: https://aws.amazon.com/es/what-is/gpt/#:~:text=Los%20modelos%20GPT%20son%20modelos,y%20extraer%20datos%20de%20documentos.

[3] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse, and H. Ahmad, "ï think this is the most disruptive technology": Exploring sentiments of chatgpt early adopters using twitter data," *arXiv preprint arXiv:2212.05856*, 2022.

[4] S. Wagholikar, A. Chandani, R. Atiq, M. Pathak, and O. Wagholikar, "Chatgpt-boon or bane: A study from students perspective," in *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. IEEE, 2023, pp. 207–212.

[5] T. Susnjak, "Chatgpt: The end of online exam integrity?" *arXiv preprint arXiv:2212.09292*, 2022.

[6] A. Haleem, M. Javaid, and R. P. Singh, "An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 2, no. 4, p. 100089, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772485923000066

[7] X. Zhai, "Chatgpt user experience: Implications for education," *Available at SSRN 4312418*, 2022.

[8] U. de la frontera, "Orientaciones para el uso de chatgpt en educación superior," 2023. [Online]. Available: https://docencia.ufro.cl/orientaciones-chatgpt-educacion-superior/

[9] C. H. González Parias, J. A. Londoño Arias, and W. A. Giraldo Mejía, "Evolución de la producción científica en américa latina indexada en scopus. 2010-2021," 2022.

[10] S. López Leyva, "El proceso de escritura y publicación de un artículo científico," *Revista Electrónica Educare*, vol. 17, no. 1, pp. 05–27, 2013.

[11] T. Greenhalgh, "Getting your bearings (deciding what the paper is about)," *BMJ*, vol. 315, no. 7102, pp. 243–247, 1997.

[12] F. Juca-Maldonado, "El impacto de la inteligencia artificial en los trabajos académicos y de investigación," *Revista Metropolitana de Ciencias Aplicadas*, vol. 6, no. S1, pp. 289–296, 2023.

[13] L. Benichou, "The role of using chatgpt ai in writing medical scientific articles," *Journal of Stomatology, Oral and Maxillofacial Surgery*, vol. 124, no. 5, p. 101456, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468785523000782

[14] J. S. Samaan, Y. H. Yeo, N. Rajeev, L. Hawley, S. Abel, W. H. Ng, N. Srinivasan, J. Park, M. Burch, R. Watson *et al.*, "Assessing the accuracy of responses by the language model chatgpt to questions regarding bariatric surgery," *Obesity surgery*, pp. 1–7, 2023.

[15] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, and A. T. Pearson, "Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers," *NPJ Digital Medicine*, vol. 6, no. 1, p. 75, 2023.

[16] A. M. Elkhatat, "Evaluating the authenticity of chatgpt responses: a study on text-matching capabilities," *International Journal for Educational Integrity*, vol. 19, no. 1, p. 15, 2023.

[17] C. Brodley, "AAAI 2013 Accepted Papers," UCI Machine Learning Repository, 2014, DOI: https://doi.org/10.24432/C5Q89K.

[18] D. Vallejo, C. Ferri, and P. Morillo, "Icmla 2014/2015/2016/2017 accepted papers data set," Jan 2019. [Online]. Available: https://www.researchgate.net/profile/Diego-Vallejo-6/publication/332010284_A_dataset_of_attributes_from_papers_of_a_machine_learning_conference/links/5cb7a5fda6fdcc1d499c4da6/A-dataset-of-attributes-from-papers-of-a-machine-learning-conference.pdf

[19] S. Sarica and J. Luo, "Stopwords in technical language processing." [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0254937

[20] A. Vuotto, C. Bogetti, and G. Fernández, "Aplicación del factor tf-idf en el análisis semántico de una colección documental," Jan 1970. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=5265905

[21] L. da Fontoura Costa, "Further generalizations of the jaccard index," *ArXiv*, vol. abs/2110.09619, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239024336

[22] X. Saralegi Urizar and I. Alegría Loinaz, "Similitud entre documentos multilingües de carácter científico-técnico en un entorno web," 2007-09.

[23] A. Carass, S. Roy, A. Gherman, J. Reinhold, A. Jesson, T. Arbel, O. Maier, H. Handels, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, D. Pham, C. Crainiceanu, P. Calabresi, J. Prince, W. Roncal, R. Shinohara, and I. Oguz, "Evaluating white matter lesion segmentations with refined sørensen-dice analysis," *Scientific Reports*, vol. 10, p. 8242, 05 2020.

[24] H. Bustince, J. Fernandez, R. Mesiar, J. Montero, and R. Orduna, "Overlap functions," *Nonlinear Analysis: Theory, Methods Applications*, vol. 72, no. 3, pp. 1488–1499, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0362546X09009936

[25] K. Abramski, S. Citraro, L. Lombardi, G. Rossetti, and M. Stella, "Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students," *Big Data and Cognitive Computing*, vol. 7, no. 3, p. 124, 2023.