Análisis sobre la categorización de tesis de grado de las carreras informáticas de la UM, mediante minería de textos

Gabriel Mariuz[1] Marisa Panizzi[1] Iris Sattolo[1]

[1] Universidad de Morón. Escuela Superior de Ingeniería, Informática y Ciencias Agroalimentarias

gmariuz91@gmail.com; marisadanielapanizzi@gmail.com; iris.sattolo@gmail.com

Resumen

Este trabajo presenta una clasificación temática de documentos automática mediante el uso de Inteligencia Artificial. Se utilizó Procesamiento de Lenguaje Natural, el cual busca que las computadoras comprendan los textos no estructurados, y extraigan información relevante de dichos textos.
Se utilizó la metodología KDT propuesta para minería de textos, y la red neuronal GPT-3 para la clasificación. Los resultados de los experimentos permitieron vislumbrar que GPT-3 es una herramienta posible para utilizarse en la clasificación de texto, obteniendo para nuestro caso un 76% de efectividad en la tarea realizada. Si bien presentó un cierto margen de error, en futuras investigaciones y mejoras en la técnica de preprocesamiento de datos, sería posible aumentar su precisión.

Palabras Clave: Minería de texto, Categorización de documentos, Redes neuronales, Aprendizaje Profundo, GPT-3

1. Introducción

La minería de textos ha ganado cada vez más atención en los últimos años debido a las grandes cantidades de datos de texto que se crean en una variedad de redes sociales, web y otras aplicaciones centradas en la información. Cualquier aplicación, en cualquier escenario, genera datos, siendo estos, los más comunes no estructurados. Como resultado, ha habido una tremenda necesidad de diseñar métodos y algoritmos que puedan procesar eficazmente una amplia variedad de aplicaciones de texto. [1].
La categorización de documentos de texto es una aplicación de la minería de texto que asigna a los documentos una o más categorías, etiquetas o clases, basadas en el contenido.
El enfoque tradicional para la categorización de textos, en que los expertos en el dominio definían manualmente las reglas de clasificación, fue reemplazando por otro basado en técnicas de aprendizaje automático, o en combinaciones de éste con otras técnicas [2].
Actualmente en las cátedras de tesis del área de Informática en la Universidad de Morón se dispone con un archivo en formato xls que contiene los datos referidos a las tesis realizadas en las carreras de informática desde el año 2004 hasta la actualidad. Este archivo cuenta, entre otros datos, con el título de la tesis, su resumen, y el área temática a la que corresponde cada tesis según un criterio personal que se usó en su momento. Este criterio, a veces, no coincide con las áreas temáticas propuestas en CACIC (Congreso Argentino de Ciencias de la Computación), originando un problema en el momento de la categorización, y al momento de dar el nombre a su tesis. Con el fin de ayudar a los alumnos al momento de elegir palabras claves y solucionar el problema planteado se propuso aplicar técnicas de minería de textos para una categorización automática y validar si la clasificación obtenida se corresponde con las temáticas abordadas en cada una de las tesis.
Se utilizó el archivo xls, con las clasificaciones otorgadas hasta ahora, como fuente de información y entrada de datos para utilizar en la herramienta GPT-3.

Antes de comenzar con los procesos de minería de textos para la categorización automática, se realizó un mapeo sistemático de la literatura (en inglés, *Systematic Mapping Studies* o SMS) [3] para hallar evidencias de las investigaciones realizadas en el contexto académico, en las cuales se utiliza la minería de texto para la resolución de problemas de categorización de documentos.
El SMS evidenció que:

- En la mayoría de los estudios analizados se presentan principalmente propuestas de evaluación y en menor medida de informar una experiencia como tipo de investigación
- Los algoritmos más utilizados en general son las redes neuronales
- Las herramientas o lenguajes de programación más usados son Weka [4], y Rapid Miner [5], mientras que, en menor medida, para los lenguajes de programación, son R [6], y Phyton [7].
- La metodología más utilizada en la minería de textos es la metodología KDT (*Knowledge Discovery in Text*), una variante de KDD enfocada en el proceso de descubrimiento de conocimiento en texto [8].

Proceso KDT

La metodología KDT [8] es una metodología para la minería de texto que se utiliza para descubrir conocimientos útiles a partir de grandes conjuntos de datos de texto. Esta metodología consta de tres pasos que se detallan a continuación:
1. Procesamiento: donde se engloban las tareas de selección o recopilación, preprocesamiento y transformación de los datos.
2. Minería de texto: que se encarga del descubrimiento de conocimiento, el cual se puede dar a través de detección de patrones, representaciones vectoriales, modelos de aprendizaje supervisado o no supervisado etc.
3. Visualización e Interpretación: en esta etapa se da paso a interpretar y validar el conocimiento obtenido tras realizar el proceso.

GPT-3

Se decidió elegir la herramienta GPT-3 [9] la cual es un tipo de red neuronal que emplea aprendizaje profundo y está enfocada en producir texto que simula la redacción humana.
GPT (*Generative Pre-trained Transformer*) es un modelo de lenguaje basado en la arquitectura Transformer [10], la cual es una arquitectura de red neuronal que permite procesar secuencias de texto muy largas de manera eficiente y que utiliza técnicas de aprendizaje profundo para procesar y generar texto en lenguaje natural.
GPT se basa en el marco teórico del aprendizaje profundo y específicamente en la técnica de preentrenamiento del lenguaje, en la que un modelo se entrena en grandes cantidades de datos de texto sin una tarea específica para aprender patrones en el lenguaje natural. Una vez que el modelo ha sido pre-entrenado, se puede utilizar para una variedad de tareas de procesamiento de lenguaje natural, como la generación de texto, la clasificación de texto y la traducción automática.

2. Desarrollo

Para el descubrimiento de la información se aplican las fases del proceso KDT:

Procesamiento

Para realizar esta fase se utilizó la planilla de cálculo en Excel que la cátedra posee, la cual contiene información sobre todas las tesis realizadas hasta el momento. El formato en que se presentan los datos en el documento es el siguiente: Código Carrera, Código Año, Código, Línea de Investigación, Número, Título, Resumen, Futuras líneas de

investigación, Autor/es, Tutor/Director, Año, Carrera.

De todos ellos, el dato que se utilizó para la clasificación es el de "Título", el cual hace referencia al título de la tesis en cuestión. El total de tesis a clasificar a partir de su título es de 283, aunque 17 fueron utilizadas a modo de ejemplo para GPT-3 a fin de darle un contexto sobre cómo debe realizar la clasificación. Las categorías sobre las cuales se deberá clasificar a las tesis son las áreas temáticas propuestas por el CACIC (Agentes y Sistemas Inteligentes, Procesamiento Distribuido y Paralelo, Tecnología Informática Aplicada en Educación, Computación Gráfica, Imágenes y Visualización, Bases de Datos y Minería de Datos, Ingeniería de Software, Arquitectura, Redes y Sistemas Operativos, Innovación en Sistemas de Software, Procesamiento de Señales y Sistemas de Tiempo Real, Innovación en Educación en Informática, Seguridad Informática). Además de dichas áreas temáticas, el CACIC presenta una subcategorización para cada una de las mismas, las cuales también serán tenidas en cuenta para incluir los títulos en alguna de ellas según corresponda.

Antes de ser usados en la herramienta, los datos pasaron por un proceso de limpieza en el cual se realizó:

1. Corrección de los títulos que tenían errores ortográficos
2. Corrección de los títulos en los que faltaban palabras
3. Ajuste de los títulos que estaban en mayúsculas para que estén todos iguales, con mayúsculas para nombres propios y títulos y el resto en minúscula

Luego, todos los títulos (incluidos los corregidos) se volcaron a una hoja de cálculo de Google, siendo en total 283 registros (Ver Tabla 1).

Palabras claves
Sistema de prevención automático de choques automovilísticos.
Aplicación de la gestión del conocimiento organizacional en la educación.
Orquestador para Aplicaciones Distribuidas.
Algoritmos Inteligentes Genéticos.
Palabras claves
Simulación de Tomografía Axial Computada.
Aprendizaje del lenguaje mediante tecnologías de voz.

Tabla 1. Listado parcial de los títulos de las tesis.

Minería de texto

GPT-3 requiere de ejemplos con un determinado formato para generar una mejor respuesta, es por lo que se probó con diferentes enunciados hasta lograr que la respuesta fuera satisfactoria y se limitara a utilizar solamente las categorías usadas por el CACIC.

Se creó para ello un documento de hoja de cálculo de Google y se le agregó la extensión Apps Script, la cual permite conectarse con la API de GPT-3 ofrecida por OpenAI y así utilizar las funcionalidades que tiene la herramienta para la generación de texto. La estructura de la hoja de cálculo de Google que servirá para darle ejemplos a la herramienta se presenta en la Figura 1.

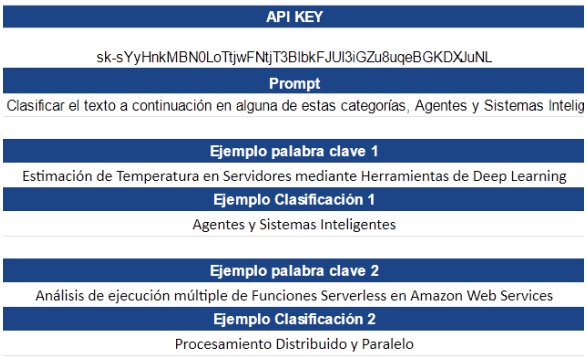

Figura 1. Formato de la hoja de cálculo con los datos que se enviarán a la API de GPT-3.

A continuación, se define qué representa cada apartado en los campos de la hoja de cálculo:

Api Key: es la clave generada por OpenIA para poder consumir el servicio con las funcionalidades que brinda GPT-3.

Prompt: el enunciado que se le da a la herramienta para que tenga contexto sobre lo que se busca que haga.

Ejemplo Palabra clave n: en este apartado van títulos de tesis a modo de ejemplo y se puede incluir una cantidad n de ejemplos que la herramienta considerará para tener contexto.
Ejemplo Clasificación n: es la clasificación a la que corresponde el título de tesis de la palabra clave, también cumple la función de servir como ejemplo, pudiendo incluir una cantidad n mientras se corresponda con la cantidad de palabras clave.

Luego de definido el formato en el que están los datos de ejemplo, se creó el script para que esos campos con información sean tomados por la API que nos brinda GPT-3, los procese y devuelva un resultado. El diagrama conceptual del procesamiento de los títulos de las tesis se presenta en la Figura 2.

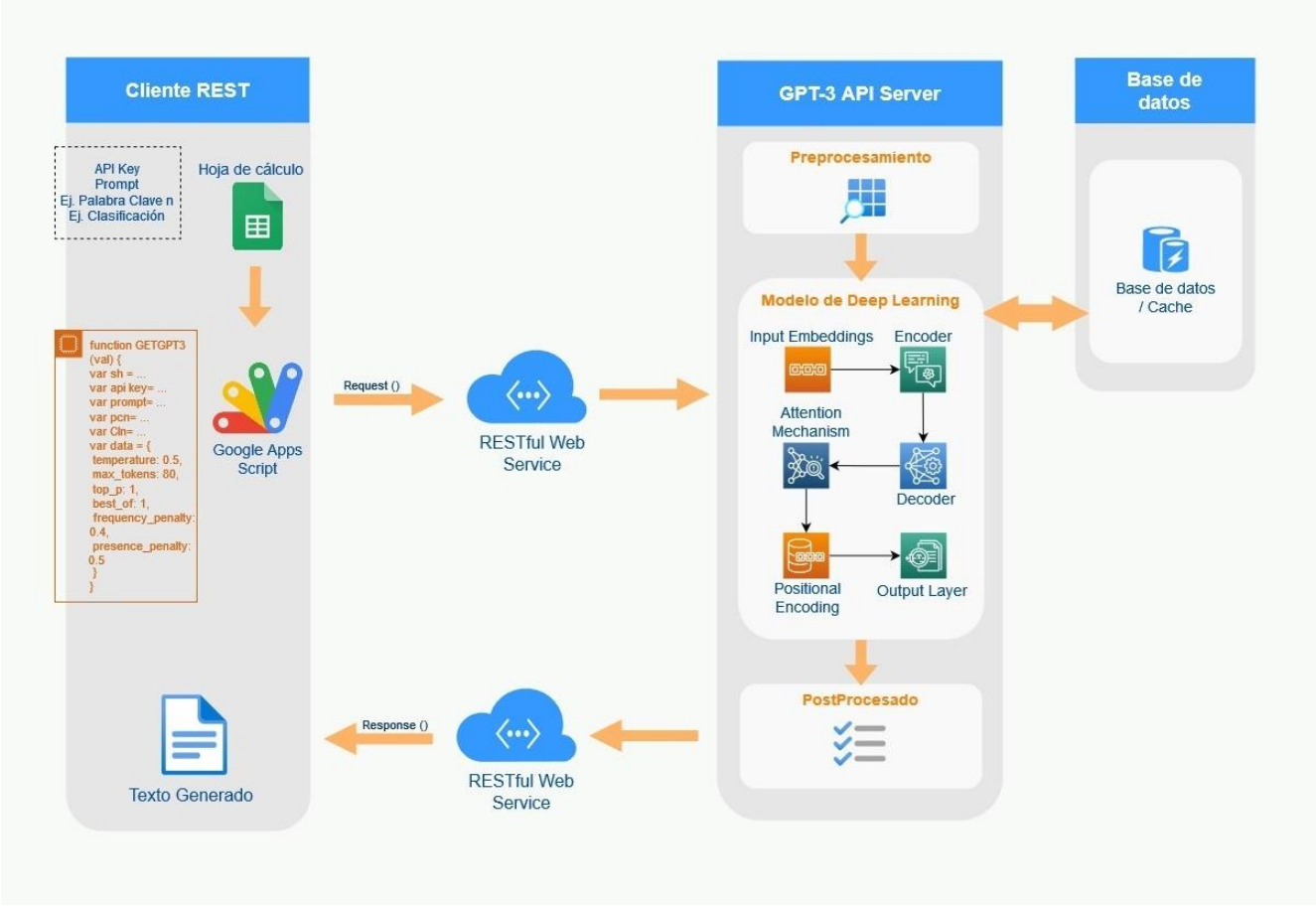

Figura 2. Diagrama conceptual del procesamiento de los títulos de las tesis.

Los datos referidos a "temperature", "max_tokens", "top_p", "best_of", "frequency_penalty" y "presence_penalty" son parámetros de ajuste para controlar la "creatividad" o la aleatoriedad en la generación de texto [9]. En general hacerlo menos aleatorio generará respuestas más predecibles y menos variadas.
Temperature: Cuanto mayor sea la temperatura, más "creativas" y "arriesgadas" serán las respuestas generadas por el modelo. Por otro lado, si la temperatura es baja, el modelo tenderá a generar respuestas más "seguras" y "conservadoras".
Max_tokens: Se utiliza para controlar el número máximo de tokens (palabras o símbolos) que se generarán en la salida del modelo de GPT. En otras palabras, establece la longitud máxima de la secuencia de salida.
Top_p: Establece una probabilidad acumulativa a partir de la cual se seleccionan

los tokens permitidos en cada paso de la generación. Establecer un valor bajo de top_p puede generar respuestas demasiado restrictivas y repetitivas, mientras que un valor alto de top_p puede generar respuestas más creativas y diversas.

Best_of: Este parámetro es particularmente útil en tareas de generación de texto donde se necesitan varias opciones de respuesta para seleccionar la mejor, como en la generación de respuestas en un chatbot.

Frequency_penalty: Se utiliza para controlar la repetición de tokens o palabras en la salida generada por el modelo de GPT.

Presence_penalty: Se utiliza en la generación de texto para alentar al modelo a incluir ciertos tokens o palabras específicas en la salida generada. Este parámetro penaliza los tokens que no aparecen en la lista de tokens o palabras deseadas, y, por lo tanto, alienta al modelo a generar respuestas que incluyan esas palabras o tokens específicos.

A continuación, se comentan los experimentos realizados:

Experimento 1. Se tomaron cuatro títulos del listado de tesis para usar como ejemplo junto a sus categorías correspondientes, ello con el fin de darle contexto a la herramienta sobre el tema del que se está hablando.

Para el campo prompt se usó el siguiente enunciado junto con los ejemplos y sus correspondientes categorías:

"Clasificar los títulos de las siguientes tesis según el tema al que corresponden: "

Ejemplos
Conceptualización de Sistema Experto para la Gestión de Eventos Gastronómicos. Categoría: Agentes y sistemas inteligentes.
Aplicación de la gestión del conocimiento organizacional en la educación. Categoría: Tecnología informática aplicada a la educación.
Aplicación de la gestión del conocimiento organizacional en la educación. Categoría: Tecnología informática aplicada a la educación.
Animación remota en mundos virtuales. Categoría: Computación Gráfica, Imágenes y Visualización.
Prototipo de Sistema para la Gestión de Controles de Tránsito Vehicular. Categoría: Ingeniería de Software.

Tabla 2. Ejemplos de títulos con sus correspondientes categorías.

Los valores usados para los parámetros de configuración para GPT-3 fueron:

- Temperature: 0.7
- max_tokens: 64
- top_p: 1
- best_of: 1
- frequency_penalty: 0
- presence_penalty: 0

Se usaron valores promedio para los campos que controlan el equilibrio del texto para buscar que el resultado presente un cierto grado de "creatividad" en su texto, permitiendo cierta aleatoriedad en las respuestas dadas.

Como último paso, se invocó a la API para que devuelva un resultado usando el comando *=GETGPT3(Coordenada de la celda en la que s e encuentra el título a clasificar en la hoja de cálculo)*, dicho resultado se guardará en la celda en que estemos parado en la hoja de cálculo.

Los resultados obtenidos comparados con la clasificación manual se muestran en la Tabla 3.

Títulos	**Manual**	**GPT-3**
Conceptualización de Sistema Experto para la Gestión de Eventos Gastronómicos.	Agentes y sistemas inteligentes.	**Sistemas expertos**
Aplicación de la gestión del conocimiento organizacional en la educación.	Tecnología Informática Aplicada en Educación.	**Educación**
Animación remota en mundos virtuales.	Computación Gráfica, Imágenes y Visualización.	**Tecnología**
Prototipo de Sistema para la Gestión de Controles de Tránsito Vehicular.	Ingeniería de Software.	**Sistema de gestión**

Tabla 3. Resultados del experimento 1.

Se puede apreciar que, al usar dicha configuración, los resultados obtenidos por parte de la herramienta si bien son coherentes y guardan relación con los títulos de las tesis, no respeta las categorías dadas como ejemplo y crea las suyas propias que no es lo que buscamos, lo cual ocurre debido al grado de "creatividad" que se le asignó, lo que le permite no ser determinista a la hora de generar una respuesta.

Experimento 2. En este caso, se modificó el prompt del Experimento 1 para indicar que se tengan en cuenta solo las 11 categorías dadas por el CACIC a la hora de clasificar los títulos se mantuvieron los mismos 4 ejemplos y sus categorías, además se ajustaron los parámetros de configuración para reducir la aleatoriedad en las respuestas y se pidió que se categorice nuevos títulos que no le fueron dados como ejemplos previamente.

La configuración de los parámetros es la siguiente:

- Temprature: 0.2
- Max_tokens: 64
- Top_p: 0.8
- Best_of: 1
- Frequency_penalty: 0.8

El prompt usado fue:

"Clasificar los títulos de las siguientes tesis en alguna de estas categorías, Agentes y Sistemas Inteligentes, Procesamiento Distribuido y Paralelo, Tecnología Informática Aplicada en Educación, Computación Gráfica Imágenes y Visualización, Bases de Datos y Minería de Datos, Ingeniería de Software, Arquitectura, Redes y Sistemas Operativos, Innovación en Sistemas de Software, Procesamiento de Señales y Sistemas de Tiempo Real, Innovación en Educación en Informática, Seguridad Informática"

Mientas que los nuevos títulos usados son:

- Prototipo de sistema domótico escalable
- Sistema de automatización de rutinas personalizadas para el Método Pilates
- Algoritmos inteligentes genéticos

Los resultados obtenidos con los ajustes realizados se presentan en la Tabla 4.

Títulos	Manual	GPT-3
Conceptualización de Sistema Experto para la Gestión de Eventos Gastronómicos.	Agentes y sistemas inteligentes.	**Sistemas expertos**
Aplicación de la gestión del conocimiento organizacional en la educación.	Tecnología Informática Aplicada en Educación.	**Educación**
Animación remota en mundos virtuales.	Computación Gráfica, Imágenes y Visualización.	**Tecnología**
Prototipo de Sistema para la Gestión de Controles de Tránsito Vehicular.	Ingeniería de Software.	**Sistema de gestión**
Prototipo de sistema domótico escalable.	Agentes y sistemas inteligentes.	**Sistema domótico**
Sistema de automatización de rutinas personalizadas para el Método Pilates.	Agentes y sistemas inteligentes.	**Sistemas de automatización**
Algoritmos inteligentes genéticos.	Agentes y sistemas inteligentes.	**Algoritmos inteligentes**

Tabla 4. Resultados obtenidos del experimento 2.

De los resultados obtenidos podemos deducir que al limitar la aleatoriedad del texto generado para forzarlo a que sea más determinista, se aprecia que la clasificación realizada respeta las categorías dadas en los ejemplos, incluso aunque no los haya catalogado igual, sin embargo, para títulos nuevos de los cuales no se le dio un ejemplo previo, no respeta las categorías dadas y genera

nuevas, lo que implica que se lo debe restringir aún más.

Experimento 3. Considerando lo expuesto en el experimento anterior y si bien los creadores de la herramienta sugieren que con 3 ejemplos es suficiente para darle contexto a GPT-3, se decidió darle un ejemplo nuevo por cada categoría usada por el CACIC, siendo 11 ejemplos en total con sus correspondientes categorías, con el fin de que al clasificar los títulos lo haga limitándose a estas categorías, el prompt inicial se dejó igual que en el caso anterior y los parámetros usados se redujeron con respecto al experimento anterior con el fin de minimizar aún más la aleatoriedad del texto generado:

- Temperature: 0.2
- max_tokens: 64
- top_p: 0.77
- best_of: 1
- frequency_penalty: 0.1
- presence_penalty: 0.1

Los resultados obtenidos se visualizan en la Tabla 5.

Títulos	Manual	GPT-3
Estimación de Temperatura en Servidores mediante Herramientas de Deep Learning.	Agentes y Sistemas Inteligentes.	**Procesamiento de Señales y Sistemas de Tiempo Real**
Análisis de ejecución múltiple de Funciones Serverless en Amazon Web Services.	Procesamiento Distribuido y Paralelo.	**Procesamiento Distribuido y Paralelo**
Un sistema integral modular para la gestión administrativa de la Educación Superior.	Tecnología Informática Aplicada en Educación.	**Tecnología Informática Aplicada en Educación**
Análisis y clasificación de ladrillos de hormigón celular a través de imágenes.	Computación Gráfica, Imágenes y Visualización	**Computación Gráfica, Imágenes y Visualización**
Un Análisis Experimental de Sistemas de Gestión de Bases de Datos para Dispositivos Móviles.	Bases de Datos y Minería de Datos.	**Base de Datos y Minería de Datos**
Ingeniería de Requisitos para Organizaciones Enfocadas en los Procesos.	Ingeniería de Software.	**Ingeniería de Software**
Análisis del comportamiento de variantes de TCP cuando se producen desconexiones.	Arquitectura, Redes y Sistemas Operativos.	**Arquitectura, Redes y Sistemas Operativos**
Detección de Anomalías en Segmento Terreno Satelital Aplicando Modelo de Mezcla Gaussiana y Rolling Means al Subsistema de Potencia.	Innovación en Sistemas de Software.	**Procesamiento de Señales y Sistemas de Tiempo Real**
Control Activo de Ruido Impulsivo Basado en la Entropía del Error con Ancho de Kernel Variable.	Procesamiento de Señales y Sistemas de Tiempo Real.	**Procesamiento de Señales y Sistemas de Tiempo Real**
Propuesta didáctica para el aprendizaje de la especificación de requisitos.	Innovación en Educación en Informática.	**Innovación en Educación en Informática**
Detección de Patrones de Comportamiento en la Red a través del Análisis de Secuencias.	Seguridad Informática.	**Redes y Sistemas Operativos**

Tabla 5. Resultados del experimento 3.

Se observa que, para este caso, al darle todas las categorías posibles a utilizar, sí que se mantuvo dentro de las mismas al categorizar nuevos títulos que no le fueron dados como ejemplos previos, por lo que se considera como óptimo para ser ejecutado sobre todo el universo de títulos disponibles, manteniendo

dicha configuración para los parámetros y el mismo prompt de este experimento.

Experimento 4. Tomando como base el prompt y los parámetros del experimento anterior, se procedió a correr el mismo proceso sobre los 266 registros que contienen los títulos de las tesis. Luego, a partir de las categorías con la mejor clasificación, se usó el mismo proceso para las subcategorías, los resultados obtenidos se analizarán en el siguiente apartado.

Evaluación e interpretación

Al aplicar la modalidad vista en el último experimento que tomaba como base lo obtenido en el experimento 3, se logró una categorización que cumple con las categorías del CACIC sin salirse de la misma para ninguno de los casos, por esto es que fue la utilizada para aplicarse a todo el set de datos completos, obteniendo como resultado el siguiente porcentaje de efectividad de la Tabla 6 para cada una de las categorías en base a la revisión realizada a los resultados obtenidos y comparando con la clasificación manual.

Categoría	Categorizaciones fallidas	Total de categorizaciones	Efectividad por categoría
Agentes y Sistemas Inteligentes.	0	22	100%
Procesamiento Distribuido y Paralelo.	0	2	100%
Tecnología Informática Aplicada en Educación.	13	25	48%
Computación Gráfica, Imágenes y Visualización.	0	7	100%
Bases de Datos y Minería de Datos.	0	10	100%
Ingeniería de Software.	24	79	69%
Arquitectura, Redes y Sistemas Operativos.	0	13	100%
Innovación en Sistemas de Software.	21	78	73%
Procesamiento de Señales y Sistemas de Tiempo Real.	0	4	100%
Innovación en Educación en Informática.	3	8	62%
Seguridad Informática.	3	18	83%
Total:	**64**	**266**	**76%**

Tabla 6. Resultados de la categorización de todos los títulos.

La efectividad general de GPT-3 para clasificar los documentos de tesis es del 76%, funcionando muy bien para categorías cuyos títulos están enteramente relacionados a conceptos del área de sistemas, mientras que para títulos referidos al uso de la tecnología en otras áreas si bien dio una categorización válida, no es la óptima para el tema buscado, por ejemplo, el caso de la tecnología informática aplicada a la educación que tiende a categorizar en este apartado títulos que se refieren a la innovación en educación en informática, pero no es capaz de discernirlo. Tomando tres de las categorías que dieron una efectividad del 100% con muestras más grandes, se aplicó el mismo proceso a las subcategorías dadas por el CACIC, obteniendo resultados satisfactorios, un ejemplo de ello se puede apreciar en la Tabla 7 donde se muestra el resultado obtenido al clasificar según las subcategorías de la categoría Agentes y Sistemas Inteligentes.

Categoría	Categorizaciones fallidas	Total de categorizaciones	Efectividad por categoría
Metaheurística inspirada en la biología.	0	1	100%

Restricciones, Satisfacción y Búsqueda.	1	1	0%
Minería de datos inteligente.	0	1	100%
Robótica inteligente.	0	2	100%
Medición del rendimiento de los sistemas inteligentes.	0	1	100%
Sistemas inteligentes.	0	1	100%
Aprendizaje automático.	0	1	100%
Metaheurística basada en la inteligencia colectiva.	1	1	0%
Sistemas multiagente.	0	1	100%
Razonamiento y lógica.	0	13	100%
Total:	**2**	**23**	**91%**

Tabla 7. Resultado de la subcategorización de la categoría Agentes y Sistemas Inteligentes.

3. Conclusiones y Trabajos Futuros

El uso del modelo de lenguaje GPT para clasificar texto ha demostrado ser relativamente eficaz, logrando una tasa de acierto general del 76% para las categorías principales, mientras que para las tres subcategorías probadas el promedio de efectividad es del 73%. Esto es un resultado muy prometedor en términos de la precisión que se puede lograr con esta herramienta. A pesar de ello, aún existen oportunidades para mejorar aún más la precisión de la clasificación, especialmente en áreas donde el modelo no ha obtenido un rendimiento óptimo ya que la herramienta presenta un cierto margen de error que debe ser tenido en cuenta y revisado, sin embargo, se espera que con la mejora continua del modelo de lenguaje y el aumento del tamaño del conjunto de datos de entrenamiento, se pueda mejorar aún más la tasa de acierto en la clasificación de texto. En resumen, los resultados obtenidos indican que GPT-3 es una herramienta útil y de fácil uso para la clasificación de texto en diferentes ámbitos y aplicaciones como el usado en este caso y ahorra el tener y que entrenar de cero todo el modelo.

Con relación al uso del chat GPT-3 para la categorización de las tesis desde el punto de vista de los alumnos, los orienta a contextualizar el título de esta, de acuerdo con su temática.

Como futuros trabajos se identifican: (a) Continuar con la experimentación usando la nueva versión del modelo, GPT-4, la cual fue entrenada con un mayor volumen de datos e incorpora mejoras en el texto generado (b) Para mejorar la eficacia del modelo en diferentes contextos, se pueden probar diferentes conjuntos de datos, revisar la cantidad de datos de entrenamiento y el ajuste de los hiperparámetros y comparar los resultados.

Bibliografía

[1] C. Aggarwal, C. Zhai. *Mining Text Data*. Springer, 2012.

[2] M. Abelleira, A. Cardoso. *Categorización automática de documentos*. XII Argentine Symposium on Artificial Intelligence (ASAI), 20-31, 2011.

[3] G. Mariuz, M. Panizzi, I. Sattolo. *Hacia el análisis de tesis de grado de carreras informáticas de la UM mediante minería de textos*. En las Actas del XXVIII Congreso Argentino en Ciencias de la Computación (CACIC 2022), La Rioja, Argentina, pp. 870-874, ISBN 978-987-1364-31-2, 2022.

[4] Weka. University of Waikato. Machine Learning Group. Página web: https://waikato.github.io/weka-

wiki/downloading_weka/. Disponible online en abril de 2023.

[5] RapidMiner. Management Team (S/A). RapidMinerStudio. Página Web: https://rapidminer.com/platform/. Disponible online en abril de 2023.

[6] Lenguaje R. R Core Team. Página web: http://mirror.fcaglp.unlp.edu.ar/CRAN/. Disponible online en abril de 2023.

[7] Python. Python Software Foundation. Página web: https://www.python.org/downloads/. Disponible online en abril de 2023.

[8] C. Villalba. *Análisis de sentimiento en Twitter sobre la serie Game of Thrones utilizando técnicas de Aprendizaje Automático Supervisado*. Universidad de Morón, Escuela Superior de Ingeniería, Informática y Ciencias Agroalimentarias, 2020.

[9] M. Carmona, R. Aranda, Á. Diaz-Pacheco, J. de Jesús Ceballos-Mejía. *Generador automático de resúmenes científicos en investigación turística*. Unidad de Transferencia Tecnológica Tepic, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE-UT3), 2022.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei. *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems 33, 2020.