

Caracterización de personajes y generación de libretos para cortometrajes usando LLMs

William Salamanca, *Ingeniería de Sistemas y Computación, UPTC*

Yasser Cristancho, *Ingeniería de Sistemas y Computación, UPTC*

Resumen

El siguiente proyecto presenta el análisis de modelos LLMs para realizar las dos tareas fundamentales en para la la elaboración de cortometrajes, la caracterización de personajes de manera detallada y la generación de un libreto, lo anterior mediante la aplicación de dos modelos LLM que permitan solventar las actividades respectivamente. Para llevar acabo la ejecución de este trabajo, se realiza a exploración de documentación acerca de los modelos LLM generadores de texto ya existentes que pueden ser aplicados o modificados de la manera mas adecuada para el desarrollo de los objetivos de la dos tareas, por medio de uso de herramientas como librerías que sean de utilidad en el proceso diseño, implementación y entrenamiento de los modelos LLMs.

Index Terms

LLM, AI, PLN, Fine Tuning, Transformer

I. INTRODUCCIÓN

EN la industria cinematográfica también se ha evidenciado el crecimiento de la revolución digital, aportando avances tecnológicos en la generación de productos visuales [de Souza et al., 2023]. Sin embargo, uno de los aspectos más fundamentales como la creación de personajes y la escritura de libretos sigue siendo muy manual.

Dentro del entorno cinematográfico, los cortometrajes permiten brindar una visualización de entretenimiento en corto tiempo a los espectadores para dejar un mensaje o transmitir emociones a partir de una historia. Para la creación de un cortometraje, la caracterización y la creación de un libreto son fundamentales para construir historias memorables y cautivadoras. Los personajes bien definidos con sus características únicas son la base de cualquier narración. Por otro lado, la creación de libretos requiere tener un amplio dominio en escritura y el arte de generar historias de manera atractiva, lo que puede llevar mucho tiempo.

Los Modelos de Lenguaje Largos o Extensos, de su traducción en sus siglas en inglés LLM (Language Large Models), han permitido evidenciar la capacidad potencial de generar texto o contenido para diferentes tareas [Hadi et al., 2023]. Contemplando lo anterior, en este proyecto se propone desarrollar un sistema que pueda analizar la narrativa del fragmento de un libro con el fin de explorar la aplicación de LLM [Du et al., 2022] para la identificación de personajes con sus características detalladas y la generación de libretos dependiendo del contexto que se ingrese en términos de ubicación, época, entre otros.

II. PLANTEAMIENTO DEL PROBLEMA

En la etapa de preproducción de una obra, guion o cortometraje, la creación de personajes detallados y bien definidos es fundamental para construir una historia cautivadora y convincente. Sin embargo, este proceso suele ser laborioso y requerir una amplia investigación y esfuerzo creativo por parte de los guionistas y directores. Mediante la implementación de un modelo LLM especializado en caracterización de personajes, se busca automatizar y agilizar esta tarea, permitiendo la generación de descripciones detalladas y coherentes de personajes a partir de un fragmento de una narrativa corta.

Por otra parte, la escritura del libreto es una etapa crucial en la producción cinematográfica de cortometrajes. Requiere habilidades narrativas, conocimiento del formato de guiones y una gran inversión de tiempo. Al utilizar un modelo LLM entrenado específicamente para la generación de libretos, se pretende agilizar este proceso, brindando a los guionistas y directores una herramienta poderosa que les permita obtener borradores iniciales de libretos a partir de una idea.

El objetivo principal de este proyecto es desarrollar una solución innovadora que aproveche la potencia de los modelos de lenguaje grandes (LLM) para facilitar y agilizar dos tareas clave en la producción de cortometrajes: la caracterización de personajes y la generación de libretos. Estos modelos serán entrenados con conjuntos de datos específicos y relevantes para cada tarea, utilizando técnicas de aprendizaje supervisado, buscando ahorrar tiempo y recursos.

III. MARCO CONCEPTUAL

III-A. Procesamiento de Leguaje Natural

El PLN es un campo de la informática que se ocupa de la interacción entre las computadoras y el lenguaje humano. Se centra en el desarrollo de métodos para que las computadoras puedan comprender, procesar y generar lenguaje natural.

III-B. Inteligencia Artificial

Es una rama de la informática que se ocupa del desarrollo de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, la percepción, el razonamiento y la resolución de problemas.

III-C. LLM

Los Modelos de Lenguaje Grandes (LLM, por sus siglas en inglés) son modelos de lenguaje entrenados en grandes cantidades de datos de texto, con millones o miles de millones de parámetros, lo que les permite capturar y representar relaciones complejas y sutilezas en el lenguaje natural

III-D. Transformer

El Transformer es una arquitectura de redes neuronales que se ha convertido en el estándar de facto para tareas de procesamiento de lenguaje natural, como la traducción automática, el resumen de texto y la generación de texto

III-E. Fine Tuning

El Fine-Tuning, o ajuste fino, es una técnica utilizada en el aprendizaje de transferencia, donde un modelo preentrenado se ajusta aún más en un conjunto de datos específico de la tarea objetivo, lo que permite adaptar el modelo a un dominio o tarea particular

IV. METODOLOGÍA

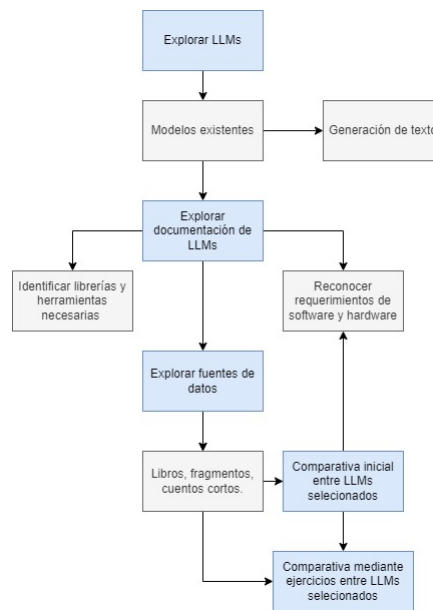


Figura 1. Metodología planteada

IV-A. Explorar LLMs

Se necesita tener una vista general de la variedad de LLMs que se encuentran actualmente implementados y desplegados.

IV-B. Explorar documentación de los LLMs

Identificar aspectos relacionados a los requerimientos básicos y/o necesarios de los LLMs para poder hacer finetune de estos. Requerimientos de software y hardware, así como posibles comparativas que indiquen su desempeño en tareas de NPL, específicamente en las de identificar, caracterizar y clasificar personajes de libros. Además de la generación de libretos para cortometrajes a partir de personajes identificados, caracterizados y clasificados, y de un contexto histórico en el que se desarrolla el libreto.

IV-C. Explorar fuentes de datos

Para el proceso de finetune de un LLM seleccionado se necesita tener varias fuentes de datos que en un principio nos provean de una descripción de las características de los personajes que se encuentran en este. Adicionalmente a lo anterior, se necesitan datos sobre contextos históricos con el fin de tener un contexto en el que se desarrolla el libreto para el cortometraje.

IV-D. Comparativa inicial entre LLMs seleccionados

- Capacidad de los modelos para realizar las tareas de identificación, caracterización y clasificación de personajes, además de la generación de libretos con base a lo anterior y un contexto historio dado.
- Tamaños y complejidades asociados a los modelos.
- Disponibilidad y acceso.
- Requerimientos de hardware.

IV-E. Comparativa mediante ejercicios entre LLMs seleccionados

Partiendo de las mismas entradas y con las mismas condiciones de hardware, identificar el que tenga mejor rendimiento en términos de tiempo pero, teniendo en cuenta la calidad de la salida.

IV-F. Herramientas

- Lenguaje de programación: Python.
- Bibliotecas: Hugging Face Transformers.
- Modelos Preentrenados: T5, GPT-Neo.
- Entorno: GoogleColab.
- Control de Versiones: Git y GitHub.
- Pypdf: Librería para extraer texto de PDFs

V. EXPLORACIÓN DE HERRAMIENTAS SELECCIONADAS

```
from pypdf import PdfReader
'''
# creating a pdf reader object
reader = PdfReader('example.pdf')
'''
# printing number of pages in pdf file
print("Numero de paginas",len(reader.pages))
'''
# getting a specific page from the pdf file
page = reader.pages[11]
'''
# extracting text from page
text = page.extract_text()
print(text)
```

Figura 2. Uso de la librería PyPDF

```
python > python3 read_pdf.py
Numero de paginas 96
12Hace 6 años, estaba volando
sobre el desierto del Sahara
y se estropeó el motor de mi avión.
Estaba completamente solo.
No tenía mecánico ni pasajeros.
Tenía que reparar el avión yo mismo
y solo tenía agua para 8 días.
Era una cuestión de vida o muerte.
La primera noche me quedé dormido
sobre la arena a miles de kilómetros
de cualquier lugar habitado.
Me sentía como un naufrago en medio del océano.
Os podéis imaginar mi sorpresa
cuando me despertó una extraña voccecita
que me dijo:
voccecita – Por favor, ¡dibújame un cordero!
Di un respingo y me froté los ojos.
Vi una personita extraordinaria
que me miraba fijamente.
```

Figura 3. Uso de la librería PyPDF

Modelo	Capacidad de Tokenización (Límite de Contexto)	Tipo	Número de Parámetros
GPT-1	512 tokens	Pago	117M
GPT-2	1024 tokens	Código libre	1.5B
GPT-3	2048 tokens	Pago	175B
GPT-4	Hasta 8192 tokens o más, dependiendo de la configuración	Pago	175B+
Megatron-LM	2048 tokens	Código libre	8.3B
PaLM	2048 tokens	Pago	540B
LaMDA	Hasta 2048 tokens	Pago	137B
GPT-Neo	2048 tokens	Código libre	1.3B y 2.7B
GPT-J	2048 tokens	Código libre	6B
BERT	512 tokens	Código libre	110M
RoBERTa	512 tokens	Código libre	355M
T5	512 tokens	Código libre	11B
DistilBERT	512 tokens	Código libre	66M

Tabla I

CAPACIDAD DE TOKENIZACIÓN, TIPO Y NÚMERO DE PARÁMETROS DE LLMs IDENTIFICADOS

Fuente: Autor

VI. LLMs SELECCIONADOS

VI-A. GPT-Neo

The screenshot shows a Jupyter Notebook titled "Copy of GPT-Neo for Beginners". The code cell contains a prompt and parameters for GPT-Neo:

```
#@title Your first completion

prompt = "Hace 6 años, estaba volando sobre el desierto del Sahara y se estropeó el motor de mi avión. Estaba completamente solo. Al día siguiente hubo mucha lluvia, pero desde el máximo grado de deshielo aprendí que el océano había recuperado su estatura y no las tempestades desencadenantes sino una actividad agradable. Un día, como pude, aprendí a seguir, y aunque el viento nos trajo nada que pensar a nuestro propio volcánico, creo que no podíamos dormir.\n\nMe apetecía pensar en esa noche, en este desierto de siete nódles, en ese día librado, ese día, en esta terraza de araquares fríos y fríos. ¿Cómo se "
```

The parameters are set to:

- min_length: 50
- max_length: 500
- temperature: 1

The output shows the generated text, which is a creative continuation of the prompt, describing a journey in the Sahara desert and the challenges faced.

Figura 4. Ejemplo prompt capacidad creativa de GPT-Neo

VI-B. T5

Figura 6. Ejemplo caracterización de personajes T5.

Figura 7. Ejemplo generación libreto T5.

VI-C. GPT-4

Identificación, clasificación y caracterización de los personajes de *El código Da Vinci*

Personajes Principales

1. Robert Langdon

- **Identificación:** Profesor de simbología religiosa en la Universidad de Harvard.
- **Clasificación:** Protagonista principal.
- **Características:** Intelectual, erudito en simbología y arte, racional y lógico, pero también abierto a lo misterioso y lo no convencional. Langdon se caracteriza por su habilidad para resolver acertijos y su vasto conocimiento en símbolos religiosos y arte.

Figura 8. Identificación, caracterización y clasificación de personajes: El código Da Vinci

2. Sophie Neveu

- **Identificación:** Criptóloga de la policía francesa.
- **Clasificación:** Protagonista femenina.
- **Características:** Inteligente, ingeniosa y valiente. Tiene una conexión personal con el misterio central de la historia, ya que es la nieta de Jacques Saunière. Su habilidad en criptología y su conocimiento de la obra de su abuelo son cruciales para el desarrollo de la trama.

3. Jacques Saunière

- **Identificación:** Curador del Museo del Louvre.
- **Clasificación:** Personaje clave cuyo asesinato desencadena la trama.
- **Características:** Misterioso y enigmático, dejó una serie de pistas criptográficas antes de morir. Saunière es un erudito en arte y símbolos, y su relación con el Priorato de Sion es central en la historia.

Figura 9. Identificación, caracterización y clasificación de personajes: El código Da Vinci

4. Silas

- **Identificación:** Monje albino y asesino al servicio de "El Maestro".
- **Clasificación:** Antagonista.
- **Características:** Devoto y fanático, Silas está profundamente influenciado por su fe y su necesidad de redención. Su carácter está marcado por su pasado traumático y su ferviente creencia en su misión divina.

5. Leigh Teabing

- **Identificación:** Historiador y ex académico británico.
- **Clasificación:** Antagonista oculto.
- **Características:** Erudito en temas del Santo Grial, Teabing es carismático y persuasivo. Sin embargo, su ambición y deseo de revelar la verdad sobre el Santo Grial lo llevan a traicionar a Langdon y Neveu. Es astuto y manipula a otros para alcanzar sus objetivos.

Figura 10. Identificación, caracterización y clasificación de personajes: El código Da Vinci

Personajes Secundarios	
1. Bezu Fache	
• Identificación:	Capitán de la Policía Judicial francesa.
• Clasificación:	Antagonista secundario.
• Características:	Determinado y autoritario, Fache sospecha inicialmente de Langdon en relación con el asesinato de Saunière. Es persistente y está decidido a resolver el caso, a menudo utilizando métodos directos y agresivos.
2. Manuel Aringarosa	
• Identificación:	Obispo y líder del Opus Dei.
• Clasificación:	Antagonista secundario.
• Características:	Devoto y ambicioso, Aringarosa está profundamente involucrado en las maquinaciones detrás del conflicto central. Su lealtad al Opus Dei y su deseo de proteger la Iglesia lo llevan a tomar decisiones cuestionables.

Figura 11. Identificación, caracterización y clasificación de personajes: El código Da Vinci

VII. FUENTES DE DATOS SELECCIONADOS

- Libros de cuentos y relatos cortos: Los libros de cuentos y relatos cortos son una fuente viable de fragmentos narrativos concisos, que podrían servir como base para desarrollar cortometrajes. Algunos ejemplos son los cuentos de Edgar Allan Poe, Anton Chéjov, Jorge Luis Borges, entre otros.
- Antologías de historias cortas: Las antologías recopilan historias cortas de diversos autores y géneros, lo que te brindaría una amplia variedad de estilos y temáticas para trabajar.
- Novelas cortas: Las novelas cortas, también conocidas como "novellas", son obras narrativas más extensas que un cuento, pero más breves que una novela completa. Podrías extraer fragmentos o capítulos completos de estas obras.
- Obras de teatro de un acto: Las obras de teatro de un acto, al ser más cortas y condensadas, son una excelente fuente de diálogos y tramas compactas que podrían adaptarse fácilmente a cortometrajes.
- Repositorios de guiones y libretos: Existen repositorios en línea que compilan guiones y libretos de películas, series de televisión y obras de teatro. Aunque no son fragmentos literarios, podrían ser útiles para aprender sobre la estructura y el formato de los libretos.
- Sitios web y plataformas de escritura creativa: Algunas plataformas en línea, como Wattpad, Inkitt o WritingPrompts, albergan historias cortas, relatos y escritos creativos de autores emergentes. Estos podrían ser una fuente interesante de material.

VIII. HALLAZGOS

Existen diferentes opciones de LLMs, particularmente de código y uso libre. Se evidenció que los modelos que mejor rendimiento y calidad de salida muestran son los que tienen costo por su uso, específicamente GPT-4 realiza exitosamente la tarea de identificar, caracterizar y clasificar personajes. Dentro de la capa gratuita de GPT-4 podemos obtener datos de la identificación, caracterización y clasificación de personajes a partir de un libro seleccionado. Lo anterior puede ser usado como fuente de información para el proceso de entrenamiento del LLM seleccionado.

IX. CONCLUSIÓN

- Partiendo de los aspectos generales de los LLMs identificados, descartamos los que tuvieran asociado un costo para su uso.
- Se identificó y seleccionó el LLM que provee una mayor capacidad de tokenización, siendo en este caso GPT-Neo.
- Respecto a los recursos de hardware, se optará por priorizar el uso de Google Colab, que en su capa más básica ofrece 12GB de RAM y 130GB de almacenamiento en disco duro.
- Podemos aprovechar la capa gratuita de modelos ya desplegados como fuente de datos para el entrenamiento del modelo.

X. BIBLIOGRAFÍA

REFERENCIAS

- [de Souza et al., 2023] de Souza, F., Ricieri, D., Barreto, R., and Farias, A. (2023). Anais do iv cobicet -trabalho completo congresso brasileiro interdisciplinar em ciência e tecnologiaa simulaÇÃO de diÁlogos e personagens no chat-gpt4: AnÁlise comparativa do desempenho em idiomas inglês e português.
- [Du et al., 2022] Du, M., He, F., Zou, N., Tao, D., and Hu, X. (2022). Shortcut learning of large language models in natural language understanding: A survey.
- [Hadi et al., 2023] Hadi, M. U., Al-Tashi, Q., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M., Akhtar, N., Wu, J., and Mirjalili, S. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects.