**UNIVERSIDAD MAYOR REAL Y PONTIFICIA DE SAN FRANCICO XAVIER DE CHUQUISACA**



GENERADOR DE POEMAS UTILIZANDO TECNICA FINE TUNING EN GPT2

**UNIVERITARIO**: Juan Sergio Pita Colque

**CARRERA**: ING. EN CIENCIAS DE LA COMPUTACIÓN

**MATERIA**: DESARROLLO DE APLICACIONES INTELIGENTES (SIS-330)

SUCRE - BOLIVIA

Contenido

[2 Introducción 3](#_Toc71896075)

[2.1 Antecedentes. 3](#_Toc71896076)

[2.2 Problemática 3](#_Toc71896077)

[2.3 Propuesta. 3](#_Toc71896078)

[2.4 Objetivo General 3](#_Toc71896079)

[2.5 Objetivos específicos 3](#_Toc71896080)

[2.6 Marco Teórico 3](#_Toc71896081)

[2.6.1 El Procesamiento del Lenguaje Natural (PLN) 3](#_Toc71896082)

[2.6.2 Transformers en Procesamiento del Lenguaje Natural 4](#_Toc71896083)

[2.6.3 ¿Cómo se trabaja con los Transformers? 4](#_Toc71896084)

[Pre-training. 4](#_Toc71896085)

[2.6.4 GPT (Generative Pre-trained Transformer) 4](#_Toc71896086)

[2.6.5 FINETUNING 4](#_Toc71896087)

[2.6.6 Poema 5](#_Toc71896088)

[Características de un poema 5](#_Toc71896089)

[2.6.7 Métricas Poema 5](#_Toc71896090)

[3 Metodología 6](#_Toc71896091)

[3.1 Método de investigación o desarrollo 6](#_Toc71896092)

[3.1.1 Método cuantitativo 6](#_Toc71896093)

[3.2 Técnicas de recolección de datos 6](#_Toc71896094)

[3.2.1 La recopilación documental y bibliográfica 6](#_Toc71896095)

[3.3 Materiales y herramientas utilizadas 7](#_Toc71896096)

[3.4 Etapas de trabajo 7](#_Toc71896097)

# Introducción

## Antecedentes.

Los poemas fueron una de las primeras maneras en las cuales los grandes literatos de la historia expresaron sus sentimientos y emociones mediante la palabra escrita. Un poema es considerado tal, en la actualidad, cuando está escrito en verso, a diferencia de aquellos textos escritos en prosa.

Como podría ser por ejemplo un cuento o una novela. El poema tiene a su vez otras características que lo vuelven un género propio y que merece, en quienes los escriben, una habilidad especial.

La importancia de la poesía radica principalmente en que nos permite ampliar nuestros horizontes. Leer poesía puede servirnos para entender mejor el mundo o para buscar nuevas formas de decir las cosas.

Unos de los mayores retos de la Inteligencia Artificial están ligados con la creación de modelos capaces de realizar esas tareas que son propias de los humanos, y en uno de los últimos avances de la misma, se han desarrollado las redes neuronales tipo transformer que 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.

## Problemática

La falta de la poesía en la vida cotidiana de las personas, hace que esta ya no radica en su vida cotidiana y en su bienestar mental.

## Propuesta.

La propuesta es un sistema que generar poemas Mediante el uso de Redes Neuronales tipo Transforme capaz de generar poesía partir del entrenamiento con una gama de Poemas que tengan características similares entre sí, tales como el género, estructura, autor, etc.

## Objetivo General

Crear un sistema inteligente, capaz de generar poemas, con sus características propias: rima, escrita en verso, formato, estrofa.

## Objetivos específicos

* Recopilar un banco de datos
* Hacer una preparación a los datos para su implementación.
* Hacer FINE TUNING en gpt2 para la generación de poemas
* Hacer una evaluación del modelo con 90% de exactitud.
* Implementar el modelo.

## Marco Teórico

### El Procesamiento del Lenguaje Natural (PLN)

El Procesamiento del Lenguaje Natural (PLN) es una de las ramas del Machine Learning que más atención ha recibido en los últimos años, tanto por parte de la comunidad científica como por parte de las empresas, que ven en sus técnicas un arma para diferenciarse de la competencia, automatizar procesos, mejorar sus productos y conocer mejor a sus clientes, entre otros.

Algunas de las aplicaciones del PLN más conocidas son la generación de textos, detectar entidades (Named Entity Recognition o NER) o clasificar textos, aunque hay muchas más. Esta versatilidad del NLP, y su aplicabilidad a múltiples problemas del mundo real en numerosas disciplinas, resalta la importancia de estas técnicas para diferentes sectores. Además, la investigación en nuevos modelos, especialmente los Transformers, ofrece nuevas posibilidades y está mejorando sus resultados al tener en cuenta el contexto.

### Transformers en Procesamiento del Lenguaje Natural

En un paper de Google de finales del año 2017 -“Attention is All You Need”- se presentó la arquitectura del Transformer, un modelo que tenía como principal innovación la sustitución de las capas recurrentes, como las LSTMs que se venían usando hasta ese momento en PLN, por las denominadas capas de atención.

Estas capas de atención codifican cada palabra de una frase en función del resto de la secuencia, permitiendo así introducir el contexto en la representación matemática del texto, motivo por el cual a los modelos basados en Transformer se les denomina también Embeddings Contextuales.

La arquitectura de Transformer incluye otras innovaciones, como los embeddings posicionales, que permiten al algoritmo conocer la posición relativa de cada palabra del texto. Además, en el paper original se examina la aplicación de esta arquitectura a tareas de traducción, demostrando ser mucho más efectiva que los métodos anteriores en este aspecto, como se puede ver en el gráfico inferior.

### ¿Cómo se trabaja con los Transformers?

Con Transformers, se suele trabajar en dos fases:

Pre-training.

En esta fase, el modelo aprende cómo se estructura el lenguaje de forma general, además de conseguir un conocimiento genérico del significado de las palabras. Esto se hace de manera parecida a los exámenes de idiomas, resolviendo “ejercicios” en los que el modelo tiene que predecir qué palabra o palabras faltan en una frase.

### GPT (Generative Pre-trained Transformer)

El modelo de lenguaje GPT aparece en el año 2018. El avance que se buscaba con su creación era que esta tecnología pudiera aprender de textos ya escritos. De esta manera, podría dar alternativas para terminar una frase ahorrando mucho tiempo de trabajo y logrando variabilidad y riqueza lingüística en los textos. Todo esto con una gramática perfecta.

¿Cómo se consigue? La tecnología GPT no es una solución de lenguaje natural clásica, sino que es una red neuronal artificial con millones de parámetros. Son necesarios grandes cantidades de datos para así hacerlo funcional, es decir, es necesario un ‘aprendizaje previo’ para que pueda tener un contexto y crear el texto deseado.

El lanzamiento de su segunda versión (GTP-2) en 2019 no estuvo exenta de polémica. La tecnología que utiliza este modelo podría utilizarse para crear noticias falsas o para cualquier otro uso fraudulento, por lo que los creadores decidieron no dar a conocer sus últimos avances.

Al margen de los problemas éticos que pueda suponer, tecnológicamente el modelo GPT-2 superaba a su primera versión con 10 veces más de parámetros (1.5 mil millones vs. 117 millones).

### FINETUNING

FINE TUNING significa tomar pesos de una red neuronal entrenada y usarla como inicialización para un nuevo modelo que se entrena con datos del mismo dominio.

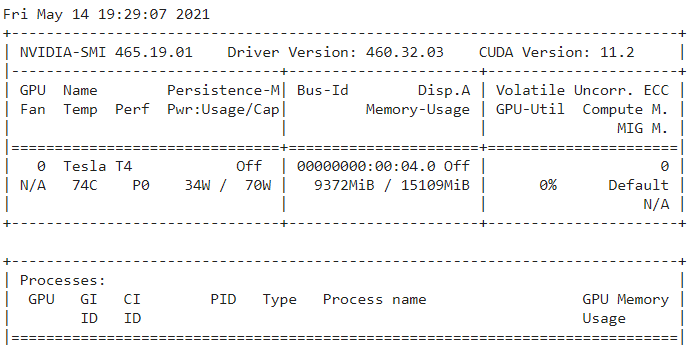
 Es usado para:

1. acelerar el entrenamiento
2. superar el tamaño pequeño del conjunto de datos

### HiperParámetros

#### BachSize

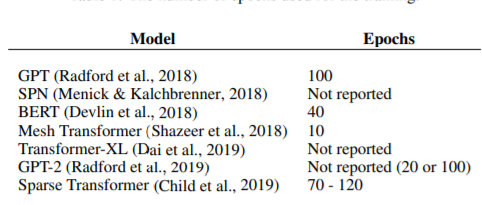
Los diferentes modelos de lenguaje requieren diferentes cantidades de memoria para contener todos los pesos y sesgos en la memoria. Según la memoria que tenga disponible su máquina, esto determinará cómo establecerá el tamaño de su lote.



Por este concepto se definió un batchSize = 24

#### Epoch

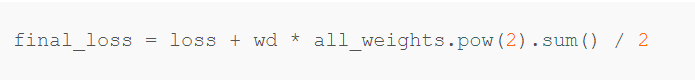
Se consideró el número de epoch en documentos de generación de texto



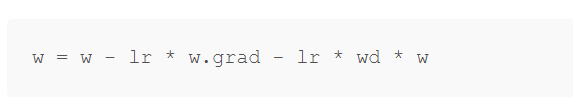
#### Optimizador Adamw

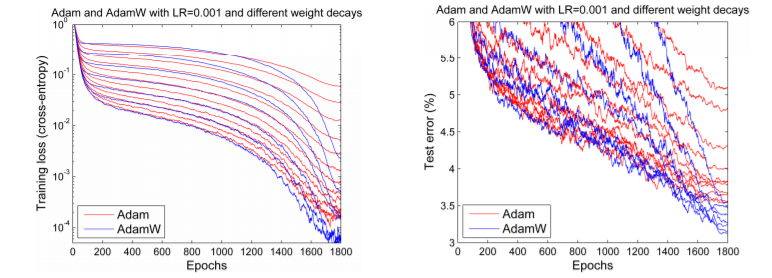
Pero a fines de 2017, Adam parecía tener una nueva vida. Ilya Loshchilov y Frank Hutter señalaron en su artículo que la forma en que se implementa la disminución del peso en Adam en todas las bibliotecas parece ser incorrecta y propusieron una forma simple (a la que llaman AdamW) de solucionarlo. Aunque sus resultados fueron ligeramente mixtos.

La regularización L2 es un método clásico para reducir el sobreajuste, y consiste en agregar a la función de pérdida la suma de los cuadrados de todos los pesos del modelo, multiplicada por un hiperparámetro dado.



Donde w es el hiperparámetro a establecer. Esto también se llama pérdida de peso, porque cuando se aplica vainilla SGD es equivalente a actualizar el peso de esta manera:





#### **Learning Rate**

Los pesos de una red neuronal no se pueden calcular mediante un método analítico. En cambio, los pesos deben descubrirse mediante un procedimiento de optimización empírico llamado descenso de gradiente estocástico.

Cada actualización de los pesos realizada en la propagación hacia atrás se pondera con un parámetro llamado tasa de aprendizaje. Si es más pequeño, está avanzando con un tamaño de paso más pequeño hacia el mínimo.

A veces, si está sobreajustado, se recomienda disminuir la tasa de aprendizaje (y aumentar el número de épocas), pero también hay otras formas de combatir el sobreajuste. Una buena tasa de aprendizaje debería permitirle alcanzar un buen mínimo en un tiempo adecuado. Esto es algo con lo que necesitas experimentar.

Se llego probrando a un Lr d=0.0001

### Poema

Se denomina poema a la composición literaria escrita en verso, que pertenece al género de la poesía y cuya estructura métrica se encarga de la formación rítmica.

Características de un poema**:**

Rima:

La rima es una figura retórica que se enfoca en los elementos morfológicos de las palabras.

Consiste en semejanzas o en el fenómeno de homofonía, que se puede clasificar como un tipo de aliteración (eco en inglés). Se basa en la repetición de fonemas al usar la métrica o ritmo.

Verso:

El verso consiste en la unión de una serie de palabras que son colocadas conforme a unas regolas que se acomodan al ritmo y al metro, y el verso no siempre coincide con una sola unidad métrica, sino que puede variar dependiendo del género literario poético.

El verso obedece a diferentes principios lo que produce que existan diferentes versos como el verso libre, el verso de arte mayor y el verso de arte menor entre otros.

Estrofa:

La estrofa es en sí el metro que es la medida silábica que se utiliza para los versos poéticos, así el verso se forma de estrofas que son versos agrupados los cuales se ajustan al metro que son el conjunto de sílabas.

Esta definición corresponde sólo al español, pues en otras lenguas como el latín se forman con el “pie”, que se forma por la duración y cantidad de las vocales.

Ritmo:

Este se produce cuando se repite en diversos intervalos en forma regular que permiten armonizar en el oído los versos y sus partes.

El ritmo se cuantifica por lo que existen diversos versos, que pasan del arte menor al arte mayor, formándose al repetirse en forma periódica los metros (pies) dentro dela poesía.

### Métricas Poema

La métrica es la medida de los versos, su estructura y cómo el autor los combina para conseguir un determinado poema y, con ello, una determinada forma de expresar su mensaje. Dicho de otra forma, la métrica es la construcción de un poema, los elementos de un verso y las combinaciones de éstas.

#### TIPOS DE MÉTRICA

Dentro de la poesía podemos encontrar una gran variedad de construcciones de **versos y estrofas**, lo que provocará que entren dentro de una u otra categoría.

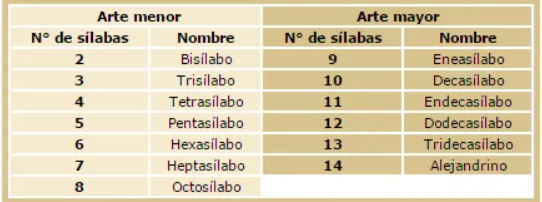
Lo primero que tenemos que distinguir son los versos de arte mayor y los de arte menor.

#### VERSOS DE ARTE MAYOR

Estos versos están compuestos de nueve sílabas métricas (veremos más adelante qué es una sílaba métrica) o más. Es decir, los versos de nueve sílabas son eneasílabos y, sumando una a la anterior son decasílabo, endecasílabos, dodecasílabo, tridecasílabos, alejandrinos o tetradecasílabos, etc.

#### VERSOS DE ARTE MENOR

Aquellos versos que tienen ocho sílabas métricas o menos. Esto se corresponde con: versos de dos sílabas se denominan bisílabos y, sumando una sílaba más a la anterior nos encontramos con los trisílabos, tetrasílabos, pentasílabo, hexasílabos, heptasílabo, octosílabos.



A su vez, también tienen que tenerse en cuenta las**licencias métricas** a las que pueda acogerse el autor, algunos ejemplos serían la sinalefa, la diéresis, hiato y la sinéresis.

#### SINALEFA

Cuando una palabra termina en vocal y la siguiente comienza en vocal, las dos sílabas se unen y suman sólo una. Ejemplo: «termina en» –> ter-mi-na+en, son tres sílabas métricas.

#### DIÉRESIS

La diéresis es la separación de dos vocales que deberían ir juntas porque son un diptongo pero el autor quiere que se cuenten separadas para mantener la métrica. A veces, esto suele ir representado por la diéresis (¨). Ejemplo: sed insaciable –> sed-in-sa-cï-a-ble, lo que son 6 sílabas en lugar de 5.

#### SINÉRESIS

Es lo contrario a la diéresis, y es que dos vocales que van juntas pero que no son un diptongo, por tanto deberían contarse separadas, se cuentan unidas. Ejemplo: Poesía –> poe-sí-a son tres sílabas lugar de cuatro.

#### HIATO

El hiato es lo contrario a la sinalefa, es decir, que la última sílaba de una palabra que termina en vocal y la primera que comienza por vocal no se unen. «Termina en» –> ter-mi-na-en, son cuatro sílabas métricas.

# Metodología

## Método de investigación o desarrollo

### Método cuantitativo

La investigación cuantitativa es aquella donde se recogen y analizan datos cuantitativos, por su parte la cualitativa evita la cuantificación; sin embargo, los registros se realizan mediante la narración, la observación participante y las entrevistas no estructuradas.

## Técnicas de recolección de datos

### La recopilación documental y bibliográfica

La técnica de recolección de información, consiste en “detectar, obtener y consultar bibliografía y otros materiales que parten de otros conocimientos y/o informaciones recogidas moderadamente de cualquier realidad, de modo que puedan ser útiles para los propósitos del estudio”.

Esta modalidad de recolección de información parte de las fuentes secundarias de datos; es decir, aquella obtenida indirectamente a través de documentos que son testimonios de hechos pasados o históricos.

Existe una gran variedad de fuentes documentales, entre las que se destacan: hemerográfica, bibliográfica, escrita, audio gráfica, videográfica, iconográfica, cartográfica y de objetos (vestidos, instrumentos de trabajo, obras de arte o artesanía, construcciones, entre otras.

## Materiales y herramientas utilizadas

* Se utilizará Python como lenguaje de programación.
* Se utilizará como plataforma hasta la fase de entrenamiento y evaluación del modelo, la herramienta Colaboraty Pro de Google.
* Se utilizará la librería Pytorch para la implementación de la red neuronal.
* Se usará como banco de datos poemas de ruben Dario
* GTP-2 para la generación de textos y para el Fine Tuning.
* Se usará Redes tipo transforme para el modelo

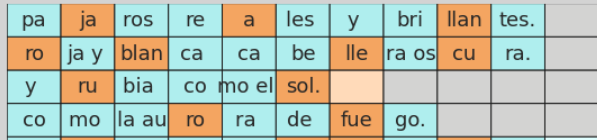
## Etapas de trabajo

|  |  |  |  |
| --- | --- | --- | --- |
| Nº | ACTIVIDAD | INICIO | FIN |
| 1 | RECOLECCIÓN DE DATOS | 8-4-2021 | 16-4-2021 |
| 2 | PREPROCESAMIENTO DE DATOS | 17-4-2021 | 23-4-2021 |
| 3 | DISEÑO DE ARQUITECTURA DEL MODELO | 24-4-2021 | 7-5-2021 |
| 4 | PROTOTIPO DEL MODELO | 8-5-2021 | 14-5-2021 |

# Resultados y conclusiones

## Resultados coinciden con los objetivos

No se pudo generar poemas que respondan a las métricas estándares de poesía.



## Resultados descritos de lo general a lo particular

* Hubo éxito en la recopilación de datos
* Se pudo hacer un procesamiento a los datos
* Se logró reentrenar el modelo gpt2 de un corpus español para la generación de poemas

## Conclusiones relacionadas al trabajo

# Bibliografia

Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more Edición Kindle

arXiv:1907.00151v5 [cs.CL] 5 Sep 2019 GPT-based Generation for Classical Chinese Poetry

https://openai.com/blog/gpt-2-1-5b-release/

<https://github.com/huggingface/transformers>

https://www.fast.ai/2018/07/02/adam-weight-decay/#understanding-adamw-weight-decay-or-l2-regularization