Análisis Temático Automatizado de Presentaciones Científicas del Sitio de CONACYT

Autores

Grupo 3

- Monica Fleitas
- Zulema Silguero
- Carlos Bustamantei

Curso: Tópicos Especiales en Ciencias de Datos

Resumen

El crecimiento exponencial de la literatura científica plantea desafíos en cuanto a la organización y recuperación de información relevante. Este trabajo presenta un sistema de clasificación automática de artículos científicos paraguayos utilizando modelos de lenguaje preentrenados. El objetivo es identificar de manera precisa si un trabajo está relacionado con redes neuronales. Se construyó un corpus a partir del repositorio de la CONACYT, procesado y etiquetado manualmente. Se aplicó un modelo basado en robertuito-base con tokenización y ajuste fino (fine-tuning) para clasificación binaria. Los resultados muestran una exactitud del 73.8%, con análisis de métricas F1 y matriz de confusión. Finalmente, se implementó un sistema RAG (Retrieval-Augmented Generation) para consultas semánticas. El sistema propuesto demuestra efectividad, aunque con limitaciones que se analizan en detalle.

Planteamiento del problema

El volumen de artículos científicos relacionados con redes neuronales crece rápidamente. Sin embargo, estos trabajos se encuentran dispersos, mal clasificados o sin etiquetas temáticas adecuadas. Esto obstaculiza su localización por parte de investigadores que necesitan información precisa para análisis comparativos, revisiones sistemáticas o inspiración para nuevos desarrollos. El presente proyecto busca solucionar este problema utilizando modelos de lenguaje preentrenados para identificar automáticamente si un artículo trata sobre redes neuronales, integrando herramientas de NLP y aprendizaje profundo.

Descripción del corpus

Se utilizó como fuente el repositorio institucional de la **CONACYT Paraguay**, accediendo a trabajos finales de investigación en formato JSON.
Cada entrada incluía los siguientes campos:

- titulo, resumen, autores, fecha_publicacion, idioma, materia, link_documento
- Si el idioma era inglés, se utilizaba titulo_traducido y resumen_traducido.

Se utilizó un modelo FastText preentrenado en español para determinar la similitud de cada artículo con la expresión “redes neuronales”. Esto permitió filtrar automáticamente documentos que, aunque no mencionaran explícitamente el término, presentaban alta similitud semántica.

Se usó como referencia una lista de términos relacionados como: "deep learning", "perceptron", "machine learning", "modelo computacional".

Ejemplo en formato JSON


```
{

        "titulo": "Deep Learning for Traffic Prediction with an Application to Traffic Lights Optimization.",

        "imagen": "",

        "autores": "Gamarra, Walter, Bogado, Maira Santacruz, Cikel, Kevin, Martínez, Elvia",

        "fecha_publicacion": "2021",

        "tipo_publicacion": "research article",

        "materia": "TRAFFIC SIMULATION DEEP LEARNING GENETIC ALGORITHMS",

        "resumen": "Resumen This work proposes the use of deep neural networks for the prediction of traffic variables for measuring traffic congestion. Deep neural networks are used in this work in order to determine how much time each vehicle spends in traffic, considering a certain amount of vehicles in the traffic network and traffic light configurations. A genetic algorithm is also implemented that finds an optimal traffic light configuration. With the implementation of a deep neural network for the simulation of traffic instead of using a simulation software, the computation time of the fitness function in the genetic algorithm improved considerably, with a decrease of precision of less than 10%. Genetic algorithms are used in order to show how useful deep neural networks models can be when dealing with vehicular flow slowdown.",

        "link_documento": "https://repositorio.conacyt.gov.py/bitstream/handle/20.500.14066/3588/PINV15-66art.pdf?sequence=1&isAllowed=y",
```

```
        "link_tema_investigacion":
"https://repositorio.conacyt.gov.py/handle/20.500.14066/3588",

        "idioma": "ingles",

        "titulo_traducido": "Aprendizaje profundo para la predicción del tráfico con una
aplicación para la optimización de los semáforos.",

        "resumen_traducido": "Resumen Este trabajo propone el uso de redes neuronales
profundas para la predicción de las variables de tráfico para medir la congestión del
tráfico. Las redes neuronales profundas se utilizan en este trabajo para determinar cuánto
tiempo pasa cada vehículo en el tráfico, considerando una cierta cantidad de vehículos en
la red de tráfico y las configuraciones de semáforo. También se implementa un algoritmo
genético que encuentra una configuración óptima de semáforo. Con la implementación de una
red neuronal profunda para la simulación del tráfico en lugar de usar un software de
simulación, el tiempo de cálculo de la función de aptitud en el algoritmo genético mejoró
considerablemente, con una disminución de la precisión de menos del 10%. Los algoritmos
genéticos se utilizan para mostrar cuán útiles pueden ser los modelos de redes neuronales
profundas al tratar con la desaceleración del flujo vehicular.",

        "label": 1

    },,
```

Metodología

El objetivo fue construir un sistema que clasifique automáticamente artículos científicos del repositorio de CONACYT Paraguay según su relación con redes neuronales.

Herramientas utilizadas

- Lenguaje de programación: Python
- Frameworks NLP y ML:
    - transformers (HuggingFace)
    - datasets
    - scikit-learn
    - fastText (semántica inicial)
    - langchain + FAISS (búsqueda semántica)
- Modelos utilizados:
    - FastText preentrenado en español (clasificación semántica inicial)

- dccuchile/bert-base-spanish-wwm-cased y pysentimiento/robertuito-base-uncased (fine-tuning)

3.3 Hiperparámetros y configuración:

python

CopyEdit

- num_train_epochs = 3
- learning_rate = 2e-5
- batch_size = 8
- weight_decay = 0.01
- save_strategy = "no"
- evaluation_strategy = "epoch"

3.4 Métricas utilizadas:

- Accuracy: precisión general
- F1-score (macro/micro): balance de precisión y recall
- Confusion Matrix: errores por clase

3.5 Aceleración

- Se utilizó `gradient_checkpointing` para economizar memoria en sistemas Apple Silicon (MPS backend).

---

4. Evaluación de Resultados

4.1 Métricas obtenidas:

lua

CopyEdit

- ☐ Accuracy: 73.8%
  La exactitud (accuracy) mide qué proporción de predicciones del modelo coinciden con las verdaderas etiquetas.

- ☐ F1-macro: 42.4%
  Promedia el F1-score de cada clase (0 y 1), tratando cada clase por igual. Aquí, su valor es bajo (~0.42), lo que indica que al menos una clase no fue predicha correctamente (en este caso, la clase 1).

- ☐ F1-micro: 73.8%
  El F1-micro pondera el F1-score por cantidad de muestras. En este caso, coincide con el accuracy porque el modelo solo predice la clase mayoritaria.

- ☐ Matriz de Confusión: [[886, 0], [314, 0]]
  Esto confirma que el modelo nunca predijo la clase 1. Clasifica absolutamente todos los ejemplos como 0.
  ¿Qué nos dicen estos resultados?

- El modelo aprendió a identificar solo los artículos irrelevantes (negativos).
- Hay una ausencia total de detección de artículos relevantes (positivos).
- El modelo sufre de sesgo hacia la clase mayoritaria.
- Las métricas de evaluación engañan si no se interpreta la matriz de confusión.
- El problema no está en el modelo, sino en los datos: desequilibrio de clases, poca representación de label=1.

4.2 Análisis cualitativo:

- El modelo logra capturar una buena proporción de artículos etiquetados como 0, pero **falla sistemáticamente en detectar la clase 1**, lo que indica un problema de desbalance de clases.
- Esto se evidencia en la matriz de confusión: **todos los artículos fueron clasificados como clase 0**.
- El `eval_loss` fue `NaN`, lo que sugiere que hubo inestabilidad numérica (posiblemente activaciones muy grandes, o etiquetas desequilibradas).

---

5. Integración con sistema RAG

Para mejorar la recuperación de información, se implementó un sistema **Retrieval-Augmented Generation (RAG)** utilizando FAISS + OpenAI.
Este sistema:

- Indexa fragmentos del resumen y título.
- Permite responder preguntas como "¿existe un trabajo que use aprendizaje profundo?".
- En pruebas controladas, mostró la capacidad de identificar documentos relevantes, pero también evidenció limitaciones semánticas.

Ejemplo de las preguntas

```
Pregunta: ¿Hay trabajos sobre aprendizaje profundo?

📄 Documento 1 (fuente: https://repositorio.conacyt.gov.py/handle/20.500.14066/3783):

Título no disponible

Resumen no disponible

Hibridación de aprendizaje profundo y neuroevolución: aplicación al pronóstico de consumo de energía eléctrica a corto plazo español. Renovar la producción de energía eléctrica sería mucho más eficiente si hubiera estimaciones precisas de la demanda futura, ya que estos permitirían asignar solo los recursos necesarios para la producción de la cantidad correcta de energía requerida. Con esta motivación en mente, proponemos una estrategia, basada en la neuroevolución, que puede usarse para este ob...

📄 Documento 2 (fuente: https://repositorio.conacyt.gov.py/handle/20.500.14066/3588):

Título no disponible

Resumen no disponible

Aprendizaje profundo para la predicción del tráfico con una aplicación para la optimización de los semáforos.. Resumen Este trabajo propone el uso de redes neuronales profundas para la predicción de las variables de tráfico para medir la congestión del tráfico. Las redes neuronales profundas se utilizan en este trabajo para determinar cuánto tiempo pasa cada vehículo en el tráfico, considerando una cierta cantidad de vehículos en la red de tráfico y las configuraciones de semáforo. También se im...

📄 Documento 3 (fuente: https://repositorio.conacyt.gov.py/handle/20.500.14066/2667):
```

```
Título no disponible

Resumen no disponible

Encuentros que abren paso al descubrimiento, al deseo, a la palabra : Proyecto
Prácticas didáctico-pedagógicas innovadoras en escuelas públicas. Resumen El
Proyecto se orientó a generar procesos de reflexión y aprendizaje sobre las
prácticas pedagógicas, en diálogo con otras propuestas filosófico-pedagógicas.
Buscó también dar vida a procesos de co-construcción e implementación de propuestas
didáctico-pedagógicas innovadoras en escuelas públicas....


📘 Respuesta: Sí, hay trabajos sobre aprendizaje profundo, como el que se menciona
en los documentos sobre la aplicación del aprendizaje profundo para la predicción
del tráfico y la optimización de semáforos, así como para el pronóstico de consumo
de energía eléctrica a corto plazo en España.
```

---

6. Conclusiones y Recomendaciones

- **Conclusión principal**: Es posible construir un clasificador temático funcional con modelos preentrenados y datos locales, pero el desequilibrio en las clases requiere atención inmediata.
- **Problemas detectados**:
  - `NaN` en pérdida por falta de ejemplos positivos.
  - MPS (MacOS) con límites de memoria bajos, ocasionando errores.
- **Recomendaciones**:
  - Balancear las clases (por ejemplo, submuestreo de clase 0 o aumento de datos positivos).
  - Evaluar modelos multiclase o jerárquicos si se incorporan más etiquetas temáticas.
  - Hacer fine-tuning con batches más pequeños y checkpoints intermedios.