UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO INFORMÁTICA

Deep learning

2023-1

Guillermo Cabrera
Julio Godoy
Manuel Perez

Informe:

"Tarea 2"

Nicolas Esteban Parra Garcia
2019422588

Concepción, - 7 de julio de 2023

Introducción

Se aborda el desafío de determinar la contribución personal que cada individuo puede realizar para construir el Chile deseado. Se considera que esta contribución puede abarcar una amplia gama de temas estructurales y emergentes. El objetivo de esta es desarrollar un modelo capaz de clasificar los tipos de contribuciones que las personas desean realizar en la sociedad, utilizando una base de datos que contiene contribuciones escritas por diferentes grupos o individuos.
El modelo utilizado se basa en RoBERTa, un modelo pre entrenado para procesamiento de lenguaje natural. Utiliza una capa de tokenización, para generar incrustaciones contextuales y una capa de clasificación para asignar etiquetas de salida.

Materiales y métodos

Adaptacion del modelo de clasificacion de texto en base a RoBERTa en español, propuesto en https://somosnlp.org/recursos/tutoriales/02_clasificacion_texto_amazon
A continuación se presentan todos los detalles respecto a la implementación del modelo. El modelo se llevó a cabo en google collab con utilización de GPU.

Preprocesamiento de datos:

- Se cargó el conjunto de datos de contribuciones utilizando la librería pandas. Los datos fueron leídos desde un archivo CSV.

Tokenización:

- Se utilizó el modelo preentrenado RoBERTa ("PlanTL-GOB-ES/roberta-base-bne") y su tokenizador correspondiente de la librería Transformers.
- Se implementó una función tokenize_category para tokenizar los textos de las contribuciones utilizando el tokenizador de RoBERTa. Se realizaron operaciones de truncamiento y relleno para asegurar una longitud uniforme de las secuencias.

Modelo de clasificación:

- Se utilizó el modelo AutoModelForSequenceClassification de la librería transformers para construir un modelo de clasificación de secuencias basado en RoBERTa. El número de etiquetas se estableció según la cantidad de categorías únicas en el conjunto de datos.
- Se definieron métricas de evaluación, como la precisión y el puntaje F1 macro, utilizando la función compute_metrics de la librería datasets.

Entrenamiento del modelo:

- Se configuraron los hiperparámetros de entrenamiento, como el número de épocas, el tamaño de lote y la tasa de aprendizaje, utilizando la clase TrainingArguments de la librería transformers.
- Se creó un objeto Trainer para entrenar el modelo. Se utilizaron los conjuntos de datos de entrenamiento y validación, así como el tokenizador de RoBERTa.

Detalles de la arquitectura e hiperparametros

- **tamaño de batch:** 16
- **Épocas:** 20
- **Optimizador:** Adam
- **Metricas:** accuracy, macro f1
- **weight_decay =** 0.01
- **Learning rate** = 2e-5
- **Loss function =** Cross-Entropy
- **Set de entrenamiento:** 16449 textos con categoría
- **Set de validación:** 8758 textos con categoría
- **Set de entrenamiento:** 2735 textos con categoría

Overview de PlanTL-GOB-ES/roberta-base-bne

- **Arquitectura:** Roberta-base
- **Lenguaje:** español
- **Tarea:** fill-masl
- **Data:** BNE

Resultados y discusión

Para los parámetros anteriores se establece el gráfico de pérdida obtenido del modelo pre entrenado al obtener el state log history.

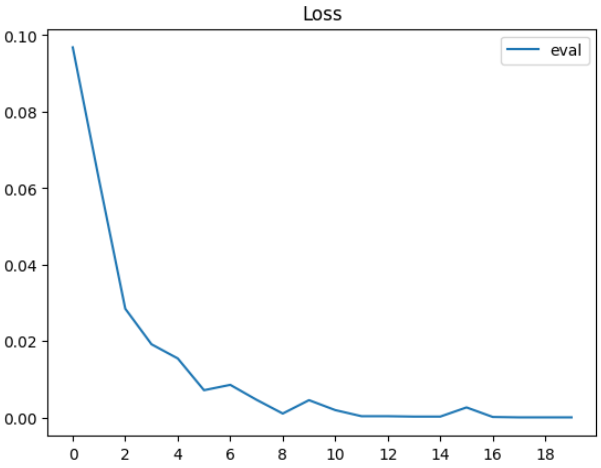

gráfico 1: curva de aprendizaje

Se puede puede observar un crecimiento acelerado para las primeras iteraciones, estabilizando la pérdida en la época 5.

Se hace uso de las métricas accuracy y Macro F1 score que representan la proporcionalidad de muestras correctamente clasificadas y una medida que combina la precisión y la exhaustividad para evaluar el rendimiento respectivamente.

Estas métricas son obtenidas del state log history que se guarda al final de cada época para el set de validación.

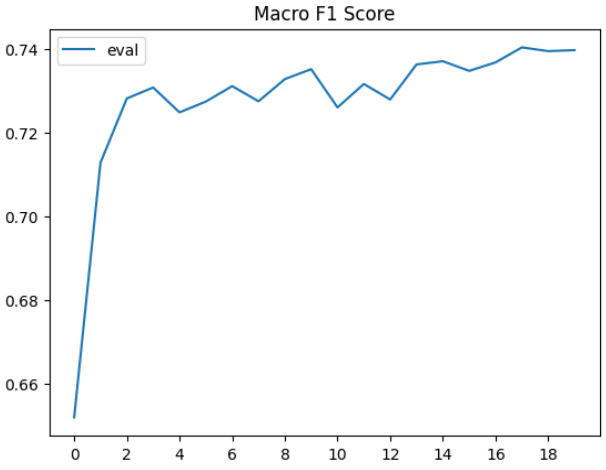

gráfico 2: F1 score vs epoca

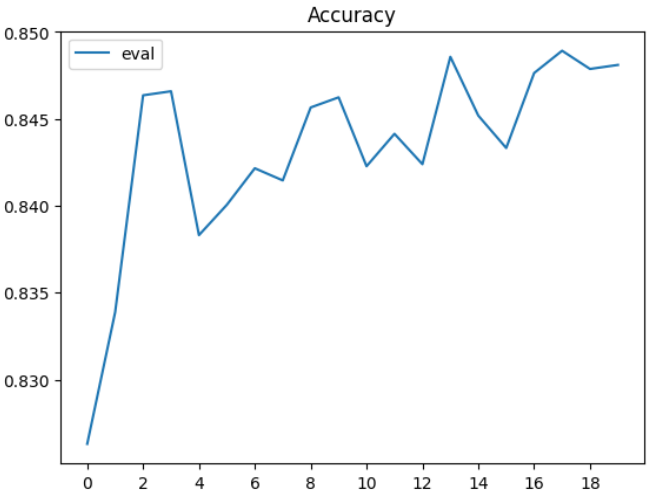

gráfico 3: Accuracy vs epoca

A Partir de estos gráficos se puede establecer que tanto la puntuación F1 como la accuracy indican que el modelo mejora hasta alcanzar un nivel de rendimiento sólido en clasificación alrededor del 80% de aciertos. También se puede concluir que el modelo está obteniendo una mejor exactitud respecto al equilibrio de precisión y exhaustividad. Considerando esta última como dividiendo el número de verdaderos positivos entre la suma de los verdaderos positivos y los falsos negativos.

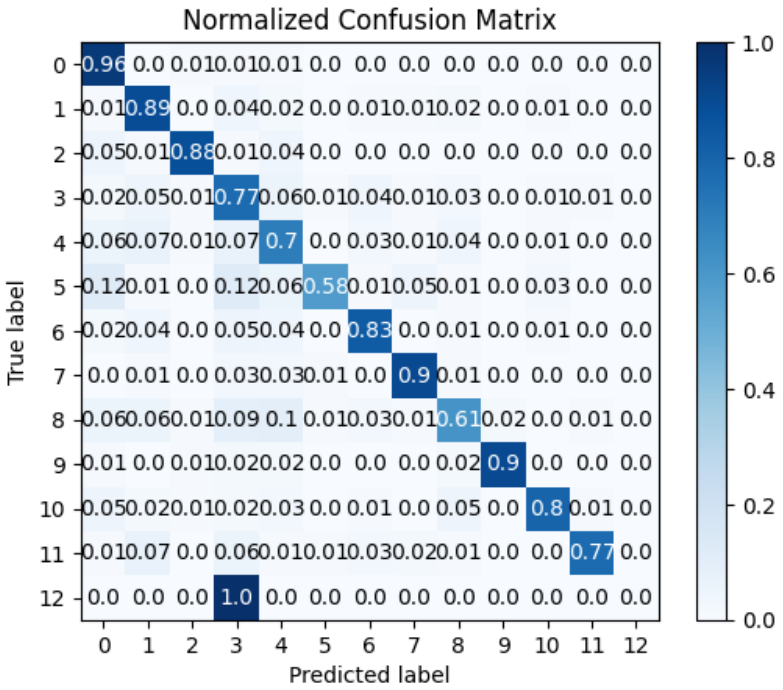

gráfico 4: matriz de confusión normalizada

Matriz de confusión normalizada obtenida de la predicción del set tokenizado de validación , se puede observar una buena predicción para la mayoría de categorías.

Observaciones

- Observando los datos se puede establecer que algunas categorías poseen cantidades desproporcionadas por ejemplo en el caso de la última categoría se presentan solo 8 textos, esto causa problemas al modelo para que aprenda a categorizar.
- Las categorías enumeradas son las siguientes en orden de 0 a 12: Participación, Reciprocidad-Redes, Protesta Social, Educación y autoeducación, Voluntariado, Cultural, Trabajo, Sustentabilidad Ambiental, Confianza en las instituciones, Combatir Delincuencia, Inclusión y Diversidad, Autocuidado y Salud, Erradicar violencia contra la Mujer.

Conclusiones

El modelo basado en RoBERTa muestra un rendimiento sólido en la clasificación de las contribuciones al construir el Chile deseado, con una exactitud y puntuación F1 macro alrededor del 80%.

Existe una diferencia entre la exactitud y la puntuación F1 macro, lo que sugiere un desequilibrio en términos de precisión y exhaustividad del modelo.

Propuestas de mejora incluyen equilibrar el conjunto de datos, ajustar hiperparámetros, utilizar técnicas de mejora de rendimiento, realizar análisis de errores y evaluar el modelo en datos nuevos.

Se sugiere explorar modelos alternativos de procesamiento del lenguaje natural para mejorar el rendimiento y la precisión del modelo.

En general, el modelo actual es sólido pero hay áreas para la mejora y el refinamiento con el fin de lograr una clasificación más precisa y equilibrada de las contribuciones al construir el Chile deseado.

Bibliografía

- Deep Learning. (s. f.). https://www.deeplearningbook.org/ Team, K. (s. f.).
- Guia del modelo base: Somos NLP - Democratizando el NLP en español. (s. f.). https://somosnlp.org/recursos/tutoriales/02_clasificacion_texto_amazon%20Roberta-base-bne
- Roberta-base-bne documentation: PlanTL-GOB-ES/roberta-base-bne · Hugging Face. (s. f.). https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne#model-description
- Transformers documentation: 🤗 Transformers. (s. f.). https://huggingface.co/docs/transformers/index
- pytorch documentation: PyTorch documentation — PyTorch 2.0 documentation. (s. f.). https://pytorch.org/docs/stable/index.html