Práctica IV

Detección de clickbait

Segunda etapa

Objetivo

Entrenar un LLM para detectar si un texto es un clickbait o no

Especificación

En equipos de 3-4 miembros hagan lo siguiente:

1) Cargue el corpus TA1C_dataset_detection_train.csv

- La columna de texto teaser se utilizará como función

- La columna Valor de la etiqueta se utilizará como objetivo (clase)

2) Divida el corpus en conjuntos de entrenamiento y desarrollo utilizando el 75 % para entrenamiento y el 25 % para desarrollo.

3) Establecer métricas para la evaluación

4) Tokenizar la entrada

5) Entrene un LLM utilizando el conjunto de entrenamiento y el conjunto de desarrollo como conjunto de datos de evaluación

6) Realizar los ajustes necesarios al modelo para mejorar el rendimiento sobre el conjunto de desarrollo.

7) Utilice el modelo mejor ajustado para predecir las instancias del conjunto de pruebas disponibles en el
Archivo TA1C_dataset_detection_dev.csv

División del corpus

Utilice la función train_test_split de scikit-learn

La instancia del corpus debe barajarse (barajarse = Verdadero)

Establezca la semilla aleatoria fija en 0 (random_state=0)

Activar la función estratificar (estratificar=y)

Métricas de evaluación

Se debe aplicar un informe de clasificación (classification_report) a las predicciones del conjunto de desarrollo para cada experimento.

Se debe generar la matriz de confusión para verificar Cómo se clasifican las instancias

- La métrica principal es f1_macro, esta se utilizará para seleccionar el mejor modelo

Tokenización y formación LLM

Se recomienda un modelo basado en Bert ajustado al español (dccuchile/bert-base-spanish-wwm-cased) como línea base.

Se debe utilizar el mismo modelo para la tokenización.

capacitación

La evaluación del modelo se realiza utilizando el conjunto de desarrollo (25% del conjunto de entrenamiento original)

Ajuste de LLM

Los parámetros del LLM deben ajustarse para mejorar la actuación

Algunos parámetros relevantes son:

- pasos_de_evaluación
- num_train_epochs
- tasa de aprendizaje

También puedes probar diferentes modelos de LLM

Predicciones sobre el conjunto de pruebas

Se debe utilizar el mejor modelo con los parámetros seleccionados para predecir el valor de etiqueta de las instancias en el archivo TA1C_dataset_detection_dev.csv

Predicciones sobre el conjunto de pruebas

El archivo CSV de salida debe tener lo siguiente características:

- Nombre: detección.csv
- Columnas: ID del tweet y valor de la etiqueta
- Carácter separador: Coma ","

Evidencia

Código fuente

Un informe en formato PDF que describa lo siguiente:

- Tarea a resolver
- Métodos de aprendizaje automático seleccionados
- Hiperparámetros ajustados
- Informe de clasificación de cada experimento

Evidencia

Una tabla que describe los experimentos realizados mostrando la mejor configuración de cada LLM en el conjunto de desarrollo

Maestría en Derecho	Hiperparámetros LLM	Puntuación f promedio macro
dccuchile/bert-base-español-wwm-cased	eval_strategy = ="pasos",eval_steps=100	0,85
dccuchile/bert-base-español-wwm-sin mayúsculas y minúsculas	por defecto	0.88
...	...	...
FacebookAI/xlm-roberta-large	num_train_epochs=3, tasa_de_aprendizaje=0,001	0.9

Evidencia

El archivo detection.csv