Machine learning techniques
Cross XQuAD

Integrantes:
Sara Sánchez
Leonardo Ángel
Juan Álvarez
Edinson Cáceres

Objetivos

- Crear un modelo de Q&A capaz de responder las preguntas formuladas por un usuario sobre un tema específico.
- Integrar al modelo la capacidad de inferir información clave de un párrafo y almacenarla para ser usada posteriormente.
- Incluir el procesamiento de 12 idiomas y 9 alfabetos distintos al modelo.
- Crear una interfaz web para el despliegue del modelo y que los usuarios puedan interactuar con la herramienta.
- Aplicar Fine-Tuning al modelo y evaluar los resultados obtenidos.

Estado del arte

- **Transformer**
  Arquitectura base, cambia el paradigma y permite trabajar en paralelo reduciendo tiempos de entrenamiento. Se basa en focalizar la atención.

- **BERT**
  Basado en transformer, trabaja bidireccionalmente, captando la representación según el contexto.

- **XQuAD**
  Trabaja con respuestas en 12 idiomas, utiliza formato JSON.

- **Fine-Tuning**
  Aprender de instancias típicas de varios dominios para adquirir conocimiento altamente transferible.

Experimentación

BERT con un Fine-Tuning sobre XQuAD.

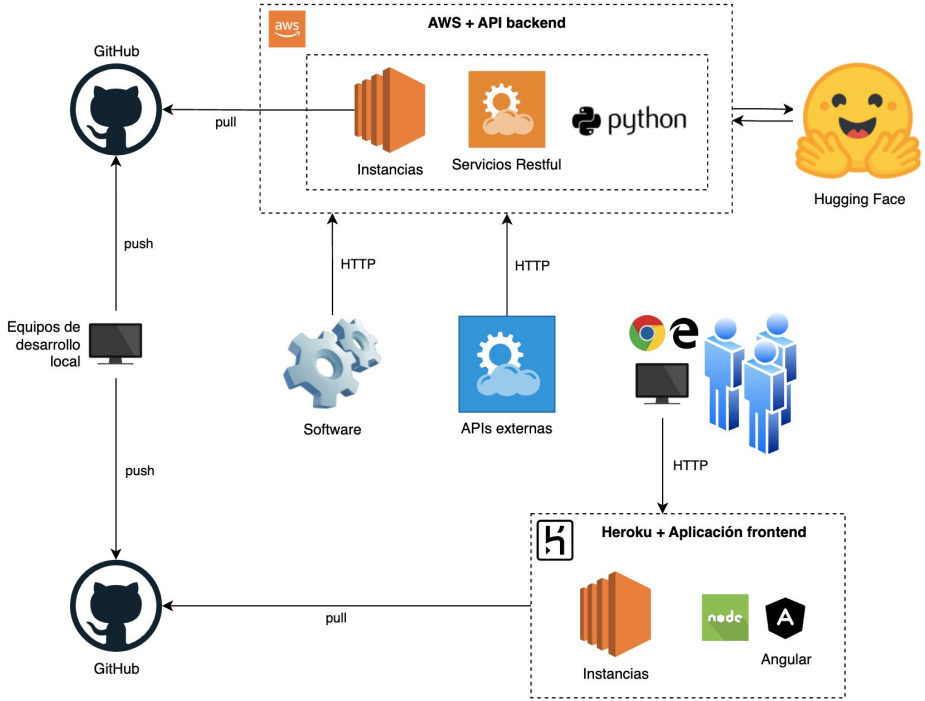
En el experimento expuesto se evalúa la capacidad del XBERT para realizar una tarea para la que no fue precisamente entrenado: responder preguntas en español dado un contexto en cualquiera de los 12 idiomas presentes en el dataset.

Resultados

- El modelo *XBERT* superó la capacidad de respuesta con lenguajes cruzados (usando F1 score y exact match como métricas).
- No fue posible mejorar la capacidad de respuesta con lenguajes cruzados del modelo base usando fine tuning con pocos pocos registros por idioma.

	Number of samples used for Fine-Tuning						
	0	**1**	**2**	**5**	**10**	**20**	**25**
Spanish	99.00	87.06	75.36	80.69	65.42	79.32	68.95
English	95.80	81.37	42.12	60.95	49.94	73.23	73.72
German	90.69	59.16	44.11	44.14	40.51	60.83	49.19
Arabic	82.95	48.86	31.77	36.51	22.00	44.24	38.72
Russian	81.31	66.38	34.63	39.12	24.89	49.18	48.53
Romanian	77.32	49.77	35.62	43.15	28.95	46.42	40.11
Vietnamese	75.47	53.92	46.47	53.46	36.99	58.66	54.03
Chinese	71.24	44.44	27.45	34.64	35.29	49.02	49.67
Greek	68.50	48.19	35.42	26.67	27.55	44.17	38.05
Hindi	67.85	51.84	37.07	33.63	39.30	46.49	44.14
Turkish	66.04	45.64	34.10	28.81	27.56	38.65	47.65
Thai	54.30	33.71	24.23	29.46	24.36	22.55	22.92

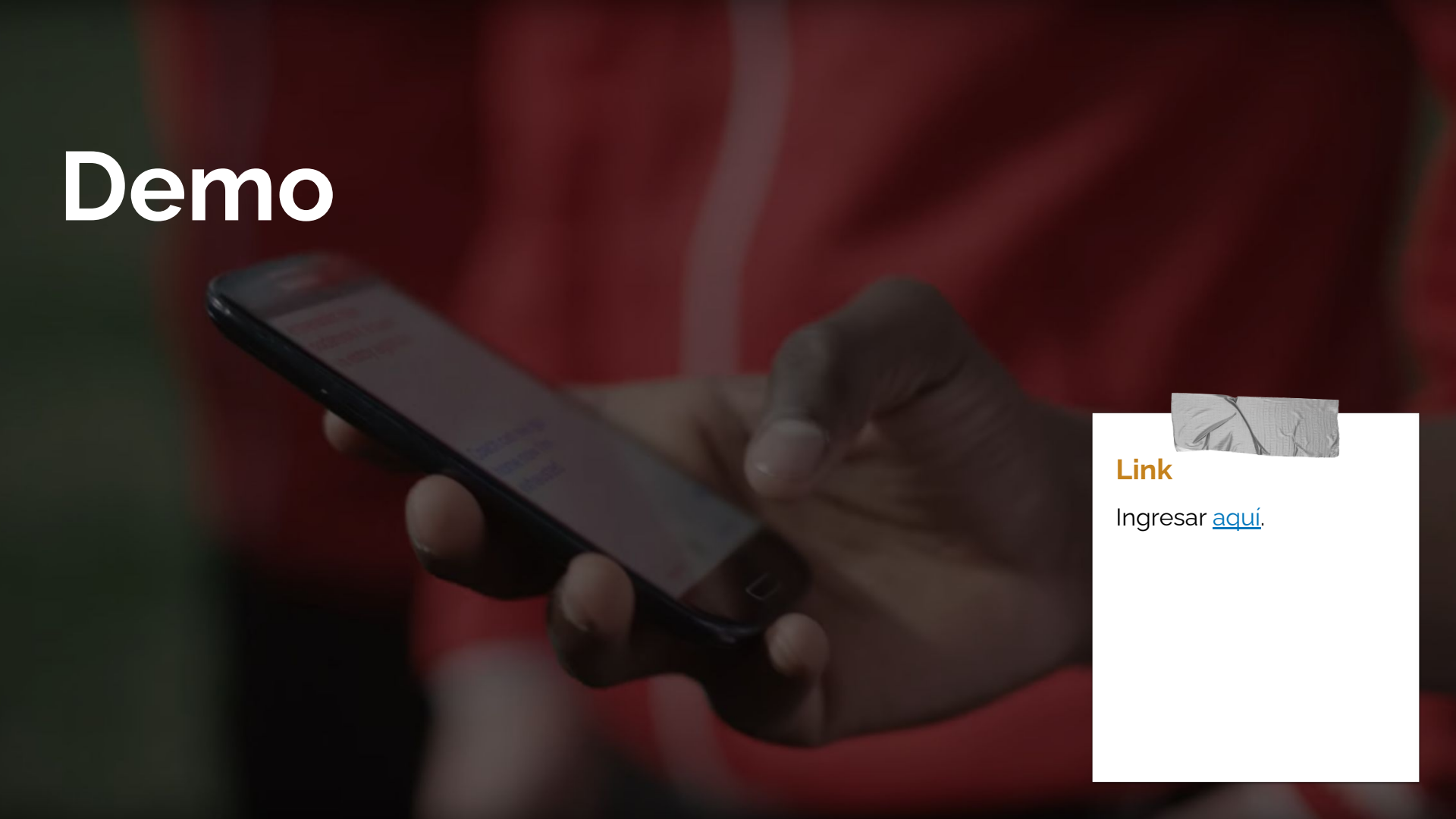
F1 score over Cross-XquAD



Despliegue de la solución

Usamos el poder de AWS junto con la facilidad de despliegue de soluciones en Hugging Face y Heroku

Demo



Link

Ingresar aquí.

Discusión

- Resultados buenos, pero insuficientes.
- Limitaciones en el ambiente de procesamiento y tiempos de cómputo.
- Se logra diseñar la herramienta que responde en un idioma dado, independiente del idioma de entrada.

Conclusiones

Costos tiempo-máquina

Tiempo de entrenamiento tardado y prueba en diferentes ambientes (NVIDIA, Colab y Colab Pro).

Necesario escalar la solución a ambientes más robustos para mejorar la predicción y automatizar mejor el despliegue web.

Rendimiento del modelo

El modelo desarrollado es capaz de recibir párrafos de información, preguntas sobre esos párrafos y generar respuestas en 12 idiomas diferentes y más de 9 alfabetos distintos.

Interfaz

La interfaz web desarrollada permite al usuario interactuar con la herramienta, se le pide ingresar un texto y una o varias pregunta, el modelo responde con base a esa información.

Se recomienda automatizar mejor el despliegue para hacerlo más ameno al desarrollador.

A futuro

Escalar la herramienta a un ambiente de desarrollo con capacidad para análisis de big data.

Despliegue más formal, en una herramienta especializada para facilidad del usuario.

Aumentar la data para la respuesta más precisa del modelo.