UNIVERSIDAD TECNOLÓGICA (UTEC)

Planilla de desafío

AiUTEChallenge

Propuesta desafio: Ética en Inteligencia Artificial - Detección de Sesgos en Sistemas de Procesamiento de Lenguaje Natural

Ficha del desafio: “EDIA / Estereotipos y Discriminación en Inteligencia Artificial”

Pauta de trabajo y estructura del informe.	Pauta de trabajo: Se pide que, luego de ver los videos tutoriales de cada herramienta del Proyecto EDIA: ● Seleccionen dos o más hipótesis de sesgo (por ejemplo: gordofobia) que no figuren en los videos. ● Testear la hipótesis y sus interseccionalidades con las diferentes herramientas del DEMO del proyecto EDIA. ● Documente el proceso en un informe. A continuación se presenta una estructura básica para el informe. Contenido básico del Informe **1**. **Introducción**: - Breve reseña sobre el desafío "EDIA / Estereotipos y Discriminación en Inteligencia Artificial". - Explicar la hipótesis de sesgo que se pretende abordar y fundamentar su relevancia aportando datos del impacto social de ese tipo de discriminación. **2**. **Fundamentos teóricos y contexto** - Definir los siguientes conceptos clave sesgo, estereotipo, discriminación algorítmica, técnicas de procesamiento del lenguaje natural (PLN) y modelos de lenguaje. - Explicar cuáles son las causas más conocidas de sesgo en los modelos de lenguaje natural. - Identificar posibles usos de modelos de lenguaje natural en sitios o aplicaciones uruguayas. En caso de que haya información disponible sobre usos nacionales, detectar su usos en plataformas globales. Analizar si estos sitios/plataformas abordan el problema de los sesgos y si siguen algún marco o protocolo en su desarrollo. **3. Metodología**: - Explicar cómo realizó el testeo de su/sus hipótesis de sesgo: Planteo de hipótesis Herramientas seleccionadas Interseccionalidades **4. Resultados y soluciones**: - Presentar casos de uso específicos donde se detectaron sesgos en los sistemas de procesamiento de lenguaje natural.

- Proponer posibles soluciones y mejoras para mitigar los sesgos.

Estructura mínima del informe:

1. Portada:
- Título del informe.
- Nombre de los integrantes del equipo y carrera que cursan
- Fecha de presentación.

2. Resumen:
- Resumen de los objetivos, metodología y principales resultados del desafío. ¿Qué sesgos se exploran? ¿En qué aplicaciones esos sesgos pueden producir daños (discriminación, invisibilización, condicionamiento de elecciones,...)?

3. Introducción:
- Presentación del desafío y su contexto.
- Objetivos del informe, incluyendo los sesgos que se han explorado y los modelos de lenguaje sobre los cuales se han explorado.

4. Fundamentos teóricos y éticos:
- Definiciones clave y conceptos relacionados con sesgos, estereotipos y discriminación algorítmica utilizados por el equipo para el desarrollo del trabajo.
- Importancia de abordar los aspectos éticos en sistemas de IA: daños y riesgos.

5. Metodología:
- Descripción de la metodología utilizada para el análisis de sesgos.
- Herramientas y descripción de los conjuntos objetivos evaluados en el desafío: sesgos y modelos explorados.

6. Caracterización de sesgos:
- Descripción de los conjuntos de palabras y expresiones multipalabra (oraciones, expresiones) con los que queda caracterizado cada uno de los sesgos explorados
- Descripción de la capacidad diagnóstica de cada uno de los conjuntos y de algunas de las palabras y expresiones en particular:
  - qué palabras resultan más caracterizadoras de cada uno de los extremos del sesgo? Cómo fue el proceso de descubrimiento de esas palabras?
  - qué palabras son más útiles para el diagnóstico, cuáles quedan más fuertemente asociadas a los extremos del sesgo? Cómo fue el proceso de descubrimiento de esas palabras?

	7. Conclusiones y recomendaciones: - Recapitulación de los hallazgos más relevantes. - Reflexiones sobre la importancia de una IA ética y la reducción de la discriminación algorítmica. - Recomendaciones para futuras investigaciones y desarrollos en el campo de la IA y la ética. **8. Bibliografía**: - Listado de las fuentes utilizadas para respaldar el análisis y las recomendaciones. **9. Anexos (opcional)**: - Detalles adicionales sobre los sitios analizados o la metodología empleada.
Referente	Referentes por parte de UTEC y contraparte. Laboratorio de Datos y Sociedad de DATA Uruguay, Fundación Vía Libre y Universidad Nacional de Córdoba.
Bibliografía	• DCC-UCHILE. (2022, Junio). Beto: Spanish BERT model by DCC-UCHILE. GitHub. Disponible en: https://github.com/dccuchile/beto • Deixi Labs. (s.f.). ELIZA. Disponible en: http://deixilabs.com/eliza.html • Google AI. (2018, Noviembre). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. Disponible en: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html • Hugging Face. (s.f.). dccuchile/bert-base-spanish-wwm-cased. Hugging Face Model Hub. Disponible en: https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased • Hugging Face. (s.f.). vialibre/edia. Hugging Face Model Hub. Disponible en: https://huggingface.co/spaces/vialibre/edia • Wikipedia. (2023, julio 19). ELIZA. Disponible en: https://es.wikipedia.org/wiki/ELIZA