Resumen Ejecutivo del Proyecto de Obtención de Grado

DETECCIÓN DE ELEMENTOS SINTÉTICOS EN MODELOS DE LENGUAJE DE GRAN ESCALA (LLMS)

Dr. Luis Miguel Escobar Vega, luis.escobar@iteso.mx

1) Descripción general del proyecto propuesto

Los modelos de lenguaje actuales tienen la capacidad de generar textos que simulan ser escritos por seres humanos. Los Modelos de Lenguaje a Gran Escala (LLMs) e.g. chatGPT[1], Palm2[2] o Llama[3] se destacan por su sofisticación y habilidad para replicar la manera en que escribimos. Sin embargo, esto representa un desafío cuando buscamos verificar la autenticidad de una noticia o estudio. Es esencial contar con herramientas que distingan si un texto ha sido producido por una máquina o por una persona. Aunque existen métodos, como la textometría, su eficacia puede verse limitada debido a la constante evolución de los LLMs. Dado el potencial de estos modelos para propagar desinformación, es crucial emplear estas herramientas de manera ética e informar al público sobre las capacidades de estos sistemas.

Este proyecto se centra en potenciar la capacidad de detectar textos generados por máquinas. Investigamos el desempeño de los modelos utilizando métricas de perplejidad, análisis semántico, y métodos de prueba de Turing Inversa (RTT), buscando identificar la naturaleza artificial de los textos. Estamos convencidos de que existen patrones en el lenguaje humano que las máquinas no han replicado completamente. Por ello, no solo examinamos modelos específicos, sino que también abordamos modelos más generales basados en estructuras lingüísticas, buscando aumentar la precisión en la detección de Elementos Sintéticos.

1) Objetivo General

Objetivo específico: Desarrollar herramienta para identificar textos sintéticos.

Algunas preguntas que podría considerar al desarrollar el plugin:

- ¿Qué técnicas de textometría, análisis semántico y pruebas de Turing Inversa (RTT) utilizará la herramienta?
- ¿Cómo será entrenado el conjunto de datos de texto sintético y texto humano?
- ¿Cómo se proporcionará al usuario un informe detallado sobre los resultados del análisis?

2) Entregabe esperado

Aplicación o especificación del planteamiento, deseable desarrollar plugin con el que se mida los resultaos obtenidos y se pueda comparar con otros modelos del estado del arte.

3) Vinculación o colaboración

Este proyecto se vincula con el programa de investigación de doctorado con el número de 399053 del Conacyt. Además, refuerza la base teórica del alumno, porque pone en práctica su aplicación pues permitirá a los interesados relacionarse con problemáticas reales en la implementación de tecnologías de inteligencia conversacional.

4) Asignaturas de la MSC relacionadas con el desarrollo del proyecto

Análisis y diseño de algoritmos [MSC2229A],
Aprendizaje automático (Machine learning) [MSC1007A] y
Aprendizaje profundo (Deep learning) [MSC2498A].

5) Participación en el proyecto

Este TOG pretende contar con un alumno de la Maestría en Sistemas Computacionales. Se requiere conocimientos generales de frameworks para procesamiento natural del lenguaje, y herramientas de desarrollo de software como Python y conocimiento básico de al menos alguna plataforma de ML como tensorflow, keras o torch.

Bibliografía relacionada

En este apartado se debe incluir una lista de bibliografía (formato IEEE) que se relacione con el desarrollo del TOG para que el alumno pueda revisarla y ahondar más en el proyecto propuesto. A manera de ejemplo:

[1] *OpenAI.* (s.f.). OpenAI. Source https://openai.com.

[2] *Bard. Google.* 2023. [En línea]. Source https://bard.google.com.

[3] *Meta AI.* (2022). Code Llama: Large Language Model Coding. Retrieved from https://ai.meta.com/blog/code-llama-large-language-model-coding

[4] *L. Varshney,* "Limits of Detecting Text Generated by Large-Scale Language Models," arXiv preprint, arXiv:2002.03438, pp. 1-10, 2020.

[5] OpenAI. (2023, February 14). Introducing ChatGPT Enterprise. Retrieved from https://openai.com/blog/introducing-chatgpt-enterprise>

[6] *I. Goodfellow, Y. Bengio, A. Courville,* (2016). Deep Learning. MIT Press.

[7] S. Raschka, V. Mirjalili, (2017). Python Machine Learning. *Packt Publishing; Edición 2nd.*

[8] K. Wu, " LLMDet: A Large Language Models Detection Tool," arXiv preprint, arXiv:2305.15004, 2023.

CV del proponente

El Dr. Luis Miguel Escobar Vega es profesor en el ITESO y especialista en semántica interpretativa, QASs y modelos de lenguaje. Ha contribuido significativamente a la semántica computacional a través de sus publicaciones, enfocadas en optimizar sistemas de recuperación de datos. Además, ha representado activamente su campo en congresos internacionales, discutiendo los beneficios de combinar sistemas estadísticos con análisis semántico.