Anexo I. Registro del Título del Trabajo Fin de Grado (TFG-BA)

NOMBRE DEL ALUMNO: Dolores Zamácola Sánchez de Lamadrid

PROGRAMA: E3-Analytics GRUPO: BA FECHA: 20/10/2024

Director Asignado: Garrido Merchán (Apellidos), Eduardo Cesar (Nombre)

Título provisional del TFG-BA:

"Detección automática de violencia o contenido sexual explícito en canciones con modelos largos de lenguaje"

ADJUNTAR PROPUESTA (máximo 2 páginas: objetivo, bibliografía, metodología e índice preliminares)

Firma del estudiante:

Fecha 20/10/2024

***Detección automática de violencia o contenido sexual explícito en canciones con modelos largos de lenguaje*"**

OBJETIVO:

Utilizando modelos de lenguaje avanzados, el objetivo de este Trabajo de Fin de Grado es crear un sistema automatizado que detecte contenido explícito, como alusiones a violencia o sexo, en las letras de canciones. Esta tecnología tiene como objetivo proteger a públicos vulnerables, como niños y adolescentes, identificando y etiquetando canciones con contenido potencialmente dañino. Esto permitiría a plataformas como Spotify o YouTube dar a los padres y educadores mayor control y transparencia. Este sistema proporciona a todos los usuarios información clara para que tomen decisiones informadas sobre el contenido musical que consumen, reduciendo el impacto negativo de los mensajes explícitos en el desarrollo emocional y cognitivo de los oyentes, además de proteger a los más jóvenes.

METODOLOGÍA

La metodología se centrará en la creación y personalización de modelos de lenguaje automatizados basados en técnicas de aprendizaje automático, particularmente modelos tipo GPT. El proceso comenzará con la selección y el preprocesamiento de un conjunto de datos compuesto por canciones del género reguetón y trap. Se utilizarán herramientas de análisis para encontrar patrones y palabras clave que indiquen la presencia de contenido inadecuado.

Utilizaremos una técnica supervisada para crear un modelo personalizado de GPT. Para lograr este objetivo, se recopilarán 100 canciones, de las cuales 50 contendrán contenido explícito, que ya he etiquetado como experto, y las otras 50 no. Para que el modelo pueda aprender a identificar qué contenido debe etiquetarse como violento o sexual y cuál no, se utilizará una base de datos con todas las letras. Se utilizarán técnicas de validación para evitar sesgos durante el entrenamiento, utilizando canciones de las listas más escuchadas, como los tops de Spotify o las radios comerciales, para que el sistema generalice correctamente sobre las canciones más populares. Por último, se evaluará su habilidad para identificar automáticamente y precisamente el contenido explícito de las canciones nuevas.

ÍNDICE

BIBLIOGRAFÍA

Fell, M., Cabrio, E., Corazza, M., & Gandon, F. (2019, September). Comparing Automated Methods to Detect Explicit Content in Song Lyrics. In *RANLP 2019 - Recent Advances in Natural Language Processing*. Varna, Bulgaria.

Markov, T., Zhang, C., Agarwal, S., Eloundou Nekoul, F., Lee, T., Adler, S., Jiang, A., & Weng, L. (2023). A Holistic Approach to Undesired Content Detection in the Real World. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12), 15009-15018.

Gutfeter, W., Gajewska, J., & Pacut, A. (2024, June). Detecting sexually explicit content in the context of the child sexual abuse materials (CSAM): End-to-end classifiers and region-based networks. *arXiv*.

Addanki, S., & Murthy, N. (2022). Text content moderation model to detect sexually explicit content. *CS230: Deep Learning*, Stanford University.

Gangwar, A., Fidalgo, E., Alegre, E., & González-Castro, V. (2017, December). Pornography and child sexual abuse detection in image and video: A comparative evaluation. *8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017)*.