Anexo 1. Formato de protocolo de investigación

1. Datos generales del proyecto

- **Título de proyecto:** Identificación automática de frases nocivas utilizando Procesamiento de Lenguaje Natural
- **Etapas/Etapa única:** Etapa única
- **Fecha inicio:** 1/12/2022
- **Fecha fin:** 30/11/2024
- **LGAC de INFOTEC en la que incide el proyecto:**

Núm.	**LGAC** (Línea General y Aplicación del Conocimiento)	Marcar con **X**
1	La SIC y la apropiación social de las TIC	
2	Las TIC y la gestión de la información y el conocimiento	
3	Ciberseguridad y delitos informáticos	
4	Protección de datos digitales	
5	Regulación de las TIC	
6	Diseño y desarrollo de sistemas embebidos inteligentes para aplicaciones industriales, biomédicas e internet de las cosas	
7	Regulación y política pública de las telecomunicaciones	
8	Analítica de datos e información	
9	Combinatoria, modelado y análisis de algoritmos	
10	Inteligencia computacional en la Ciencia de Datos	
11	Analítica de grandes cúmulos de información	X

- **Palabras clave:** NPL, identificación, frases, nocivas, clasificación.

2. Descripción del proyecto:

- **Resumen** (ejecutivo):

  Objetivo general:

  Desarrollar y analizar algoritmos de clasificación de texto para la identificación de frases nocivas.

  Objetivos particulares:

  -Analizar la relación entre diferentes categorías de frases nocivas como podría ser la misoginia, el lenguaje ofensivo, el machismo, el sexismo, entre otros.
  -Analizar la aplicabilidad de los algoritmos desarrollados para la generación de estadísticas en mensajes de Twitter.

Metodología:
-Revisión de literatura
-Etiquetado de la base de datos
-Desarrollo de algoritmos
-Pruebas de rendimiento
-Comparación con el estado del arte

- Resultados esperados:
  El envío de un artículo científico a una revista indexada en el JCR, una plática de divulgación y el desarrollo de un software libre para detección de frases nocivas y ver su aplicabilidad.

- Antecedentes:
  El proyecto propuesto se encuentra en el área de procesamiento de lenguaje natural, en particular de clasificación de texto visto como un problema de aprendizaje supervisado, esta área ha sido parte de la investigación que ha desarrollado el Dr. Graff en los últimos ocho años. En estos años la investigación del Dr. Graff está enfocada al estudio de algoritmos de aprendizaje computacional y cómputo evolutivo en tareas de procesamiento de lenguaje natural como la clasificación de texto y el estudio de algoritmos evolutivos en el campo de aprendizaje computacional. En estas áreas el Dr. Graff cuenta con más de 20 artículos en revistas indexadas en el JCR y una obra total de 87 documentos de acuerdo con Scopus.

- Justificación:
  Las contribuciones esperadas de este proyecto serán en el área de procesamiento de lenguaje natural a través de crear un algoritmo que identifique frases nocivas a través de las representaciones de texto propuestas. Además, esta contribución será enfocada en el lenguaje español ya que actualmente la mayor parte de las contribuciones en el área son realizadas para inglés. Cabe mencionar, que este proyecto toma de punto de partida el trabajo anterior de grupo de trabajo (análisis de sentimientos y clasificación de texto) contribuyendo a esta línea de investigación con nuevas técnicas, así como el problema específico de la identificación de frases nocivas en texto.

- Objetivo general:
  Desarrollar y analizar algoritmos de clasificación de texto para la identificación de frases nocivas.

- Objetivos específicos:
  -Analizar la relación entre diferentes categorías de frases nocivas como podría ser la misoginia, el lenguaje ofensivo, el machismo, el sexismo, entre otros.
  -Analizar la aplicabilidad de los algoritmos desarrollados para la generación de estadísticas en mensajes de Twitter.

- Meta (especificación de la finalidad del proyecto de investigación):
  -Desarrollar algoritmos de categorización lineales en los conjuntos de datos de frases nocivas.
  -Analizar las similitudes y diferencias de los modelos de los diferentes tipos de frases nocivas.
  -Desarrollar y analizar representaciones específicas para la identificación de frases nocivas.
  -Comparar la aptitud de los modelos que utilizan las representaciones desarrolladas y representaciones realizadas mediante aprendizaje semi supervisado, por ejemplo, mediante emoticones o palabras claves.
  -Medir la aptitud de los modelos generados en datos extraídos de Twitter seleccionados mediante consultas geográficas.
  -Generar estadísticas de datos de Twitter tanto en datos con referencia geográfica y sin ella.

- Metodología:
  El primer paso de cualquier proyecto de investigación es la revisión de literatura, este proyecto no es la excepción; en una búsqueda preliminar se encontraron investigaciones relacionadas donde se han generado datos para el estudio de algún tipo de lenguaje nocivo, en particular en [4, 6] se presentan dos conjuntos de datos que sirven como base para la detección de misoginia, complementando este trabajo en [5] proponen un conjunto para el estudio del sexismo. Estos trabajos muestran la actualidad del proyecto propuesta. Sin embargo, es necesario realizar una revisión minuciosa del estado del arte. Por otro lado, esta revisión incluirá aquellas contribuciones que contengan bases de datos que puedan y estén relacionadas a frases nocivas. Después de haber realizado la revisión de la literatura y de las bases de datos disponibles se analizará si estas

son suficientes para realizar la investigación o se requiere complementar estas bases de datos con el etiquetado de frases que presenten algún tipo de lenguaje nocivo. Una vez que se han seleccionado o construido las bases de datos con las cuales se trabajará, el proyecto continuará siguiendo una metodología tradicional de aprendizaje supervisado. Éstos diferentes conjuntos servirán como los conjuntos de prueba donde los algoritmos que se desarrollarán serán probados. La aptitud de los algoritmos será medida con métricas tradicionales de clasificación como son el f1, recall, precisión, entre otras. Se desarrollarán diferentes algoritmos de clasificación de texto probando inicialmente una bolsa de palabras con un pesado conocido como TFIDF y un clasificador lineal para tener un modelo base siguiendo la metodología presentada en [7]. Después, se analizará el comportamiento de incluir modelos pre-entrenados como son los obtenidos por emoticones o incluso incluir modelos generados mediante conjuntos de datos similares, esto siguiente inicialmente la metodología presentada en [2]. Una característica que diverge de la metodología tradicional de aprendizaje supervisado es que se utilizarán algoritmos que permitan identificar las similitudes y diferencias de los modelos obtenidos de tal manera que se puedan comparar los diferentes tipos de mensajes nocivos a través de la comparación de los modelos que se generaron. La primera idea que se probará es una bolsa de palabras fija para todos los modelos y entrenar un modelo lineal en los conjuntos de datos, esto hace que el modelo sea un vector y de esta manera se puede medir la similitud (e.g, similitud coseno) y así distinguir cuáles de los tipos de frases nocivas son más similares. Una vez que los modelos han sido desarrollados y probados en los conjuntos de datos seleccionados, se probarán en datos provenientes de Twitter.

- Beneficios esperados (según sea aplicable: los avances de la ciencia y/o tecnología que se prevén alcanzar con el logro de los objetivos; beneficios potenciales en términos de mercado, económicos, sociales, ambientales y recursos tecnológicos, entre otros; y modalidades de protección en materia de propiedad intelectual):

- Resultados esperados (productos entregables):
  - Análisis comparativos de los algoritmos que actualmente se utilizan para la detección de frase nocivas.
  - El conjunto de datos etiquetado
  - Librería de uso libre de algoritmos de identificación de frases nocivas.
  - Escritura de artículo para ser enviado a una revista indexada en el JCR.
  - Escritura de un artículo de divulgación.

3. Cronograma de actividades

Etapa 1/Etapa única:				
#	Actividad	Resultado esperado (producto entregable)	Fecha inicio (dd/mm/aaaa)	Fecha fin (dd/mm/aaaa)
1	Revisión de literatura	Análisis comparativos de los algoritmos que actualmente se utilizan para la detección de frase nocivas.	01/12/2022	01/02/2023
2	Selección de Base de Datos	El conjunto de datos etiquetado	01/02/2023	01/03/2023
3	Desarrollo de algoritmos	Librería de uso libre de algoritmos de identificación de frases nocivas.	01/04/2023	01/07/2023
4	Comparación de las representaciones generadas	Comparación de las representaciones generadas	31/07/2023	01/12/2023
5	Escritura de artículo	Escritura de artículo para ser enviado a una revista indexada en el JCR	01/12/2023	01/03/2024
6	Preparación de un artículo de divulgación	Escritura de un artículo de divulgación	01/03/2024	01/06/2024
7	Desarrollar aplicación	Código cargado en un repositorio de Git asociado a la publicación.	01/07/2024	31/11/2024

4. Referencias

1.- Mario Graff, Daniela Moctezuma, Sabino Miranda-Jiménez, Eric S. Tellez, “A Python library for exploratory data analysis on twitter data based on tokens and aggregated origin–destination information” in Computers & Geosciences, Volume 159, 2022,105012,ISSN 0098-3004, https://doi.org/10.1016/j.cageo.2021.105012.

2.- M. Graff, S. Miranda-Jimenez, E. S. Tellez and D. Moctezuma, "EvoMSA: A Multilingual Evolutionary Approach for Sentiment Analysis [Application Notes]," in IEEE Computational Intelligence Magazine, vol. 15, no. 1, pp. 76-88, Feb. 2020, doi: 10.1109/MCI.2019.2954668.

3.- Iberian Languages Evaluation Forum (Iberlef 2022), https://sites.google.com/view/iberlef2022.

ATENTAMENTE

DRA. MIREYA PAREDES LÓPEZ
POSDOCTORANTE ACADÉMICO

SIN ANEXOS

C.c.p. **Mtro. Carlos Josué Lavandeira Portillo**, Director Adjunto de Innovación y Conocimiento. Presente.