La investigación reportada en esta tesis es parte de los programas de investigación del CICESE (Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California).
La investigación fue financiada por el CONAHCYT (Consejo Nacional de Humanidades, Ciencias y Tecnologías).

Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Maestría en Ciencias
en Ciencias de la Computación

Detección de eventos violentos en publicaciones de redes sociales

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Maestro en Ciencias

Presenta:

Esteban Ponce León

Ensenada, Baja California, México

2023

Tesis defendida por

Esteban Ponce León

y aprobada por el siguiente Comité

Dr. Irvin Hussein López Nava

Codirector de tesis

Dr. Manuel Montes y Gómez

Codirector de tesis

Dr.Hugo Homero Hidalgo Silva

Dr.Humberto Pérez Espinosa

Dr. Gilberto López Mariscal

Coordinador del Posgrado en Ciencias de la Computación

Dra. Ana Denise Re Araujo

Directora de Estudios de Posgrado

Resumen de la tesis que presenta Esteban Ponce León como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la Computación.

Detección de eventos violentos en publicaciones de redes sociales

Resumen aprobado por:

Dr. Irvin Hussein López Nava	Dr. Manuel Montes y Gómez
Codirector de tesis	Codirector de tesis

En los últimos años, ha habido un interés creciente en el monitoreo de redes sociales para recopilar información y, en algunos casos, para examinar la ocurrencia de delitos. Sin embargo, gran parte de las investigaciones hasta ahora solo se han centrado en ciudades de EE. UU. o extranjeras, y por ende, en publicaciones y conjuntos de datos en inglés. El objetivo principal de esta tesis es diseñar un método que permita la identificación de publicaciones de eventos violentos en español y en Twitter, utilizando información multimodal y técnicas de aumento de datos que mejoren el rendimiento de los modelos. Para esto, el trabajo de investigación se dividió en dos fases experimentales. La primera orientada a identificar publicaciones a partir de solo texto, explorando diferentes técnicas de aumento de datos para texto y modelos de aprendizaje máquina y profundo. En la segunda fase, se extendió el método propuesto para abordar la identificación en un contexto multimodal, es decir, considerando tanto los textos de los tweets como las imágenes compartidas que los acompañan. En este caso el método propuesto consideró utilizar descripciones textuales de las imágenes y abordar la problemática desde el dominio textual, además se hicieron 2 tipos de aumento de datos para cada tipo de información. La evaluación de los métodos se hizo utilizando las colecciones de la tarea de evaluación DA-VINCIS 2022 y 2023. Los resultados demostraron una mejora en el rendimiento de los modelos al considerar el uso de información multimodal y el uso de aumento de datos.

Palabras clave: Detección de Violencia, Redes Sociales, Aumento de Datos, Procesamiento del Lenguaje Natural, BERT, BETO, Descripción de Imágenes

Abstract of the thesis presented by Esteban Ponce León as a partial requirement to obtain the Master of Science degree in Computer Science.

Detection of violent events in social media publications

Abstract approved by:

PhD Irvin Hussein López Nava	PhD Manuel Montes y Gómez
Thesis Co-Director	Thesis Co-Director

In recent years, there has been a growing interest in monitoring social networks to gather information and, in some cases, to examine the occurrence of crime. However, much of the research so far has only focused on US or foreign cities, and thus on English-language publications and data sets. The main objective of this thesis is to design a method that allows the identification of publications of violent events in Spanish and on Twitter, using multimodal information and data augmentation techniques that improve the performance of the models. For this, the research work was divided into two experimental phases. The first aimed at identifying publications from only text, exploring different data augmentation techniques for text and machine and deep learning models. In the second phase, the proposed method was extended to address identification in a multimodal context, that is, considering both the texts of the tweets and the shared images that accompany them. In this case, the proposed method considered using textual descriptions of the images and addressing the problem from the textual domain, in addition, 2 types of data augmentation were made for each type of information. The evaluation of the methods was done using the collections of the DA-VINCIS 2022 and 2023 evaluation task. The results demonstrated an improvement in the performance of the models when considering the use of multimodal information and the use of data augmentation.

Keywords: Violence Detection, Social Networks, Data Augmentation, Natural Language Processing, BERT, BETO, Image Captioning

Dedicatoria

A mis padres por todo su amor y apoyo incondicional a lo largo de toda mi vida, especialmente estos años y por motivarme a superarme. A mis amigos y compañeros por su apoyo a lo largo de mis estudios.

Agradecimientos

Al Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California, por brindarme de las herramientas necesarias para completar este trabajo.

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) por brindarme el apoyo económico para realizar mis estudios de maestría.

A mis directores de tesis **Dr. Irvin Hussein López Nava** y **Dr. Manuel Montes y Gómez (INAOE)** por toda su ayuda, tiempo, atención, comentarios, y paciencia durante el desarrollo de este trabajo y sobre todo, su forma de ser como personas e investigadores.

A mi comité de tesis, Dr. Hugo Homero Hidalgo Silva y al Dr. Humberto Pérez Espinosa, por sus comentarios, observaciones y sugerencias a lo largo de este trabajo.

Tabla de contenido

Página

Lista de figuras

Figura Página

Figura Página

Figura Página

Lista de tablas

Tabla Página

Capítulo 1. Introducción

De acuerdo a la Organización de las Naciones Unidas, los países de América tienen las peores tasas de homicidio por un amplio margen, representando el 37 % del total mundial en una región que representa el 13 % de la población mundial[1].

Además, es frecuente que las personas que viven o presencian algún tipo de delito, ya sea perpetrado por un individuo o por un grupo, experimenten un impacto negativo. Esto puede llevar a un aumento en los niveles de depresión, ansiedad o incluso estrés postraumático[2].

A pesar de la relevancia de este tipo de hechos, en algunos países es complicado medir la tasa de incidencia delictiva con precisión debido a la falta de confianza que la población tiene hacia las autoridades, causando una insuficiencia de datos y en el peor de los casos, llevando a conclusiones alejadas de la realidad. Por ejemplo, en la Ciudad de México 9 de cada 10 crímenes no son reportados, donde se acentúa la falta de confianza hacia las autoridades, y solo 1 de cada 100 casos reportados llega a una sentencia (Piña-García & Ramírez-Ramírez (2019)).

Esta situación abre la oportunidad a explorar nuevas fuentes de información que puedan complementar la información que las autoridades disponen. Actualmente, el uso de las redes sociales ha cambiado la manera en que la información es compartida y ahora es una parte relevante de las agencias de comunicación gubernamentales y de compañías de ámbito privado (Kaplan & Haenlein (2010)). La información recolectada de las redes sociales es una valiosa entrada para analizar el flujo de información, opiniones y sentimientos (Prieto Curiel et al. (2020)). Por ejemplo, en los últimos años, ha habido un interés creciente en el monitoreo de redes sociales en línea y el análisis de Big Data. Las plataformas de redes sociales se utilizan para recopilar información y, en algunos casos, para examinar y predecir la ocurrencia de delitos. Sin embargo, gran parte de las investigaciones hasta ahora solo se han centrado en ciudades de EE. UU., y por ende, en publicaciones escritas en inglés (Piña-García & Ramírez-Ramírez (2019)).

Twitter, al ser una de las grandes redes sociales predominantes en la actualidad y que cuenta con alrededor de 465 millones de usuarios activos mensuales, ha sido usada principalmente para la transmisión de noticias (Kwak et al. (2010)). Dado su modelo de datos simple, con una API de acceso directo a estos, hace que sea la red social ideal para estudios que involucran estas áreas. Esta plataforma es una potente herramienta que ha sido explotada para investigaciones en el área de Procesamiento del

[1] https://www.un.org/en/un75/new-era-conflict-and-violence, Accedido 15/ago/2022
[2] https://sites.google.com/view/davincis-iberlef/home, Accedido 15/ago/2022

Lenguaje Natural (PLN o NLP, por sus siglas en inglés) para tareas como la detección de discursos de odio y detección de noticias falsas (Antonakaki et al. (2021)), permitiendo varios enfoques de estudio que puedan llegar a tratar con la ocurrencia de eventos violentos. Por ejemplo, durante una recolección de tweets masiva durante 70 días realizada en 18 países de habla hispana de América Latina, 15 de 1000 publicaciones se encontraban relacionadas, de manera textual, con crímenes o alguna opinión sobre estos (Prieto Curiel et al. (2020)). A su vez, se menciona que se presenta un sesgo hacia crímenes de origen violento y sexual, como llega a ocurrir en los medios de comunicación.

De acuerdo a lo que se ha investigado, existe mucha variabilidad en el número de instancias para cada tipo de evento violento y medio de divulgación, ya sea por incidencias reales de baja frecuencia o falta de reportes oficiales. Lo anterior puede generar un desbalance de clases al momento de realizar una recolección de datos propia y, si no se toman las medidas adecuadas, pueden suponer un problema de sobre ajuste al momento de implementar técnicas y herramientas computacionales para resolver tareas específicas.

Alternativas como el aumento de datos, que incluye desde el uso de datos sintéticos, la recolección de más datos o hasta el uso de grandes modelos de lenguaje (LLM, por sus siglas en inglés), pueden ayudar a reducir el efecto negativo que puede traer un conjunto de datos desbalanceado. Además de resolver la escasez de datos, puede permitir que un modelo sea más robusto mediante la implementación de determinadas técnicas que pueden inyectar ruido en diferentes niveles, generando nuevas instancias.

A pesar de los inconvenientes que se pueden presentar, como los mencionados anteriormente, se considera la detección de reportes de eventos violentos en redes sociales como un paso para incentivar a los ciudadanos a denunciar cualquier acto o evento violento de los cuales sean testigos o víctimas, tomando ventaja de la accesibilidad y facilidad de transmisión de información que las redes sociales aportan a sus usuarios y poder ayudar las autoridades, como a los ciudadanos, a tomar medidas adecuadas con base en la información que se recopile.

1.1. Definición del problema

Con base en lo mencionado hasta el momento, se observa que hay publicaciones con ciertas temáticas relacionadas con eventos violentos que se pueden encontrar con mayor o menor frecuencia, lo cual puede aumentar la dificultad de proyectos que se quieran llevar a cabo con este tipo de información o temática.

Si bien, noticieros locales y nacionales han estado fomentando la denuncia ciudadana a través de sus redes sociales, estas suelen ser, por una parte, encuestas de opinión sobre las condiciones de la infraestructura o sobre eventos periódicos, y denuncias libres.

Sin embargo, lo que respecta a eventos violentos, suelen presentar una variabilidad en cuanto a lo que se suele reportar en el dominio público. Si bien, lo ideal es recolectar la mayor cantidad de datos posibles para mantener la información lo más natural, esto no siempre se puede llevar a cabo debido a que la recolección y el procesamiento de los datos puede ser costoso en términos de tiempo y recursos. En estos casos, la generación de datos sintéticos como alternativa puede ser la respuesta, que aunque pudiera llegar a parecer un trabajo de carácter trivial, en realidad una técnica de aumento de datos que haya funcionado para una tarea, no suele funcionar para la gran variedad de problemas que se presentan en el área de PLN (Bayer et al. (2022)). Es por eso que es necesario realizar un análisis exploratorio de las técnicas (SMOTE, *back translation*, reemplazo por sinónimo y el uso de GPT-3) para poder encontrar aquella que tenga mayor afinidad a la tarea deseada. Además, muchos de los reportes en redes sociales, gracias a las funcionalidades de los actuales teléfonos inteligentes, son de naturaleza multimodal, viniendo acompañados en gran parte de imágenes. Con respecto a esto último, pocos trabajos se han realizado en esta dirección y no es claro la relevancia y complementariedad de las imágenes respecto al texto.

Por lo tanto, en este trabajo se busca encontrar una solución a la detección de eventos violentos en redes sociales por medio de un enfoque basado en técnicas de aumentos de datos, como alternativa a la recolección de nuevos datos, y buscar la más adecuada que permita adaptarse y mejorar el rendimiento de los modelos presentados.

1.2. Preguntas de Investigación

Con base a la problemática expuesta anteriormente, se definen las siguientes preguntas de investigación para guiar el trabajo:

- ¿De qué manera se puede utilizar/combinar la información multimodal para realizar una mejor identificación de los reportes de eventos violentos?
- ¿Con qué técnica de aumento de datos y en qué medida se puede mejorar el desempeño de los modelos de detección de eventos violentos en Twitter?

1.3. Objetivos

1.3.1. Objetivo general

Diseñar un método que permita la clasificación de publicaciones de eventos violentos en Twitter a partir del uso de información multimodal, utilizando técnicas de aumento de datos que mejoren el desempeño.

1.3.2. Objetivos específicos

Para cumplir con el objetivo general se plantean los siguientes objetivos específicos:

1. Estudiar las técnicas de aumento de datos a partir de información multimodal, texto e imágenes.
2. Diseñar un método de aprendizaje profundo de clasificación binaria (violento y no violento) y multiclase (no violento, robo, secuestro, accidente, homicidio) que utilice solo información textual e incorpore las técnicas estudiadas previamente.
3. Diseñar un método de aprendizaje profundo de clasificación binaria y multiclase que utilice la información multimodal incorporando las técnicas de aumento de datos.
4. Evaluar y comparar los modelos propuestos de forma individual y respecto a los resultados reportados en la literatura.

1.4. Metodología de la Investigación

Para cumplir con los objetivos propuestos, el trabajo de investigación se organizó en 7 pasos principales:

1. **Adquisición de una colección de tweets con temática sobre violencia (unimodal):** Este paso consistió en obtener y analizar una base de datos orientada a la temática de eventos violentos publicados en el idioma en español con información textual. Se consideró la base de datos de DA-VINCI y, a su vez, se realizó un preprocesamiento y eliminación del ruido presente en esta base de datos.

2. **Diseñar un método de clasificación binario y multiclase con información textual:** El segundo paso consistió en hacer uso de la base de datos adquirida en un modelo de clasificación binaria (violento y no violento) y multiclase (no violento, robo, secuestro, accidente, homicidio). Basándonos en lo observado, se realizaron los ajustes necesarios para mejorar su rendimiento.

3. **Implementación de técnicas de aumento de datos y para la clasificación binaria y multiclase con información textual:** En este paso se buscó explorar con mayor detalle las diferentes técnicas disponibles para el aumento de datos y eventualmente aplicarlas a la base de datos original. Después se procedió a entrenar y evaluar los métodos de clasificación previos y comparar su rendimiento con los obtenidos en el paso 2.

4. **Adquisición de una colección de tweets con temática sobre violencia (multimodal):** Este paso consistió en obtener y analizar una base de datos extendida en las modalidades de texto e imágenes, en el idioma en español. Se consideró la base de datos de DA-VINCI extendida. A su vez, se realizó un preprocesamiento y una eliminación del ruido.

5. **Diseñar un método de clasificación binario y multiclase con información multimodal:** En el quinto paso, se buscó replicar el paso 2 pero ahora incluyendo información multimodal (texto e imágenes) proveniente del paso anterior, trabajando en un sólo dominio con ayuda de los modelos visión-lenguaje para extraer las descripciones de las imágenes. Iniciando con la clasificación binaria y concluyendo esta etapa con la clasificación multiclase, se evaluó cada parte para realizar los ajustes correspondientes tras evaluar sus resultados.

6. **Implementación de técnicas de aumento de datos para la clasificación binaria y multiclase con información multimodal:** El sexto paso involucró explorar e implementar diferentes técnicas de aumento de datos y aplicarlas a la base de datos original. Similar al paso 3, se entrenarán los modelos previamente diseñados para el paso 5 y se realizarán los ajustes adecuados.

7. **Evaluación de los distintos modelos:** Al final se analizó y evaluó el efecto que tienen las técnicas de aumento de datos en los modelos de clasificación diseñados en etapas anteriores y comparar su desempeño respecto a los resultados del estado del arte.

1.5. Estructura de la tesis

El resto de la tesis se encuentra organizada de la siguiente manera: En el Capítulo 2 se detallan los fundamentos teóricos importantes que se abordan en el resto del trabajo. En el Capítulo 3 se revisan los trabajos relacionados con el tema de investigación. El Capítulo 4 detalla la metodología seguida para realizar el presente trabajo de investigación. A continuación, en el Capítulo 5 se presentan los resultados obtenidos junto a la discusión de los mismos. Finalmente, en el capítulo 6 se presentan las conclusiones derivadas del trabajo de tesis, las contribuciones, limitaciones y el trabajo a futuro.

Capítulo 2. Fundamentos

En este capítulo se introducen los conceptos básicos y técnicas que se utilizaron en el presente trabajo de tesis. En la primera parte, se detallan algunos conceptos generales relacionados con el problema de clasificación. En la segunda parte, se describen las técnicas y herramientas computacionales que se utilizaron durante el trabajo.

2.1. Tarea de clasificación

Para poder resolver el problema que involucra la identificación de tweets que reporten eventos violentos, se decidió modelarlo como un problema de clasificación. Dicho enfoque en el área de inteligencia artificial consiste en que un algoritmo de inferencia, previamente entrenado en un conjunto de datos, realice una predicción categórica con base a un dato o información de entrada. Esta predicción puede ser considerando dos posibles resultados (clasificación binaria), varios resultados independientes entre sí (clasificación multiclase) o que considere varios resultados a la vez (clasificación multietiqueta).

Para poder llevar a cabo la clasificación existen dos tipos de procesos de aprendizaje, *lazy learners* los cuales no crean ningún modelo con base en los datos destinados al entrenamiento, al memorizar los datos de entrenamiento, y cuando es necesario hacer una predicción, buscan al vecino más cercano que comparta características similares entre todos los datos de entrenamiento, ejemplos son el algoritmo kNN. El otro tipo de proceso de aprendizaje es denominado *eager learners*, estos algoritmos primero construyen un modelo a partir del conjunto de datos de entrenamiento antes de hacer cualquier predicción, ejemplos de lo anterior son Support Vector Machines o redes neuronales (Sabharwal & Selman (2011)).

Para que los algoritmos de inferencia deseados puedan ser capaces de realizar las predicciones deseadas es necesario extraer de los datos información o características que puedan asociar a las clases de interés y, donde estas nuevas características suelen ser una representación numérica de los datos originales (fácilmente interpretables por humanos).

2.2. Técnicas

En esta sección, se describen las técnicas computacionales empleadas para realizar la extracción de características, la selección automática de características, aumento de datos y los algoritmos de inferencia utilizados en el presente trabajo.

2.2.1. Extracción de características

2.2.1.1. Indexado

La forma más básica de representar un texto para que un algoritmo de inferencia pueda procesarlo es a través de un espacio vectorial o también conocido como un enfoque de N-gramas (Aas & Eikvil (1999)). En esta representación, la colección de documentos es representada en una matriz de palabra por documento, y suele ser una matriz dispersa porque no todas las palabras aparecen en cada documento.

Cada palabra en la matriz cuenta con un peso asignado, el cual puede ser tratado de diferentes maneras. En este trabajo se exploraron principalmente las siguientes:

- **Peso booleano:** El más simple, consiste en asignarle un peso de 1 si la palabra ocurre en un documento y 0 si no.

$$a_{ik} = \begin{cases} 1, & \text{si } f_{ik} > 0 \\ 0, & \text{de lo contrario} \end{cases} \tag{1}$$

- **Term frequency – Inverse document frequency (Tf-idf):** Este enfoque toma en cuenta la frecuencia de las palabras en todos los documentos. Asigna el peso a la palabra "i" en el documento "k" en proporción al número de ocurrencias de la palabra en el documento, y en proporción inversa al número de documentos en la colección para los cuales la palabra aparece al menos una vez.

$$a_{ik} = f_{ik} * log(\frac{N}{n_i}) \tag{2}$$

2.2.1.2. Representación por vectores contextualizados

La representación por vectores, o *embeddings*, contextuales consiste en la generación de estos por medio de redes neuronales, donde previamente se utilizaban métodos estadísticos. Estos tipos de vectores poseen ventajas como ser de baja dimensionalidad, en relación con el número de palabras, e intentan preservar la información sintáctica y semántica.

Representación generada por BERT. El modelo pre-entrenado de transformers BERT(Bidirectional Encoder Representations from Transformers) genera los vectores mediante la combinación de la tokenización del texto de entrada, la codificación posicional para capturar la posición de los tokens, y el proceso de atención basado en Transformers para capturar las relaciones contextuales entre los tokens en múltiples capas de atención. Esto resulta en embeddings contextualizados que capturan información semántica y sintáctica del texto (Devlin et al. (2019)), en la Figura 1 se puede observar de manera general el proceso para obtener los vectores contextualizados. BERT fue pre-entrenado en grandes cantidades de texto utilizando *BookCorpus* y Wikipedia en el idioma inglés, con un total de 3,300 millones de palabras y con un aproximado de 110 millones de parámetros para su versión base. Durante este proceso, BERT se pre-entrenó en dos tareas no supervisadas donde le permite aprender a predecir palabras enmascaradas (*Masked LM*) dentro de oraciones y a determinar si dos oraciones son continuaciones lógicas una de la otra (*Next Sentence Predicion*). Permitiendo que el modelo BERT adquiera un conocimiento general del lenguaje (Devlin et al. (2019)). La aplicación de este modelo se enfoca en *transfer learning* a través de un ajuste fino (*Fine-tuning*), donde en lugar de entrenar un modelo desde cero para cada tarea, el *transfer learning* aprovecha el conocimiento previamente aprendido de una tarea fuente y lo aplica en una tarea objetivo. Por otro lado, **BETO** es un modelo BERT pre-entrenado en el idioma Español de tamaño similar al modelo BERT original. Para su entrenamiento se utilizó información de Wikipedia, Naciones Unidas, TEDxTalks, noticias y documentos gubernamentales, entre otros. Además, se consideraron detalles como el enmascaramiento dinámico, el cual implica utilizar diferentes máscaras para la misma oración en el corpus original (Cañete et al. (2020)).

2.2.1.3. Visualización de representaciones: t-SNE

La incrustación de vecinos estocásticos distribuidos en t (Van Der Maaten & Hinton (2008)) es una técnica no determinista que permite visualizar características de alta dimensionalidad al calcular las

similitudes, como la distancia euclidiana o la similitud del coseno, entre todas las parejas de los puntos de datos en el espacio de alta dimensionalidad para determinar qué tan similares son los puntos entre sí. Con base a las similitudes calculadas, calcula las probabilidades de similitud entre puntos y cuanto más similares sean dos puntos, mayor será la probabilidad de que sean vecinos cercanos en el espacio de baja dimensión. Crea un nuevo espacio de baja dimensión (generalmente 2D o 3D) donde los puntos se visualizarán.

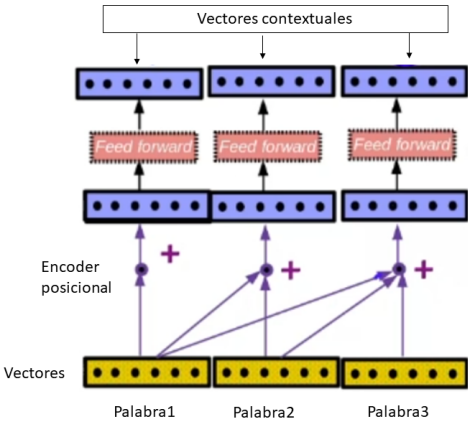


Figura 1. Vectores de palabras contextualizados. A través del mecanismo de atención de Transformers, se obtiene su representación vectorial con respecto a su posición y al de las otras palabras antes y después.

Inicialmente, los puntos se colocan aleatoriamente en este espacio y ajusta las ubicaciones de los puntos en el espacio de baja dimensión para que las probabilidades de similitud en este espacio se aproximen a las probabilidades de similitud en el espacio de alta dimensión donde, a diferencia de la técnica SNE, utiliza una distribución t de Student en lugar de una Gaussiana para calcular la similitud entre dos puntos en el espacio de baja dimensión. La optimización se realiza minimizando una función de costo que mide la discrepancia entre las probabilidades de similitud en los dos espacios. t-SNE utiliza el gradiente descendente estocástico para ajustar las posiciones de los puntos de manera iterativa hasta que se alcanza una configuración en la que las similitudes sean aproximadamente consistentes entre ambos espacios.

2.2.2. Reducción de dimensionalidad

Un problema con los métodos de extracción de características de tipo estadísticos en el área de procesamiento del lenguaje natural, por ejemplo los mencionados en 2.2.1.1, son su alta dimensionalidad en

el espacio de características, generando un alto costo computacional.

La selección de características busca remover palabras que proporcionan poca o ningún tipo de información para mejorar la clasificación y reducir el costo computacional.

2.2.2.1. Estadístico χ^2

El estadístico chi-cuadrada (χ^2), al ser catalogado como una de las 3 técnicas más efectivas para la selección de características, mide la independencia entre la palabra y la clase (Aas & Eikvil (1999)) y se define de la siguiente manera:

$$\chi^2(w, c_j) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (C + D)} \tag{3}$$

Donde:

- A: Numero de documentos de la clase c_j que contiene la palabra w
- B: Numero de documentos que contienen la palabra w pero no pertenece a la clase c_j
- C: Número de documentos de la clase c_j que no contiene la palabra w
- D: Número de documentos que no pertenecen a la clase c_j ni contiene w
- N: Número total de documentos

2.2.3. Aumento de datos

El aumento de datos se comprende como el incremento de instancias, generalmente, de una clase o categoría que se encuentra poco representada en un conjunto de datos. Esto se realiza de dos maneras, de manera natural, recolectando nueva información del medio, y de manera sintética, generando instancias nuevas por medio de alguna transformación o recurso externo. En la Figura 2 se observa que las técnicas de aumento de datos para texto se pueden dividir en dos ramas principales: en el espacio de características y en el espacio de instancias (Bayer et al. (2022)).

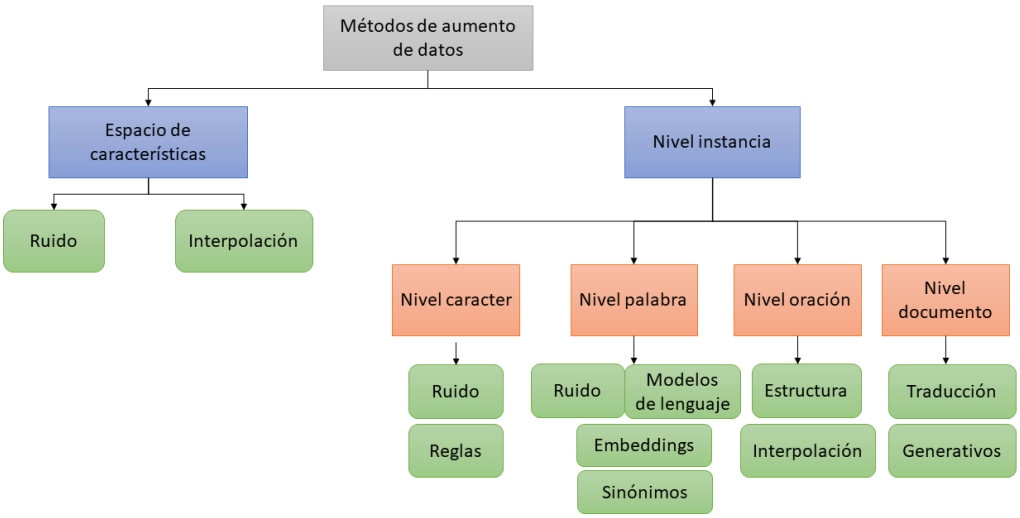


Figura 2. Taxonomía de aumento de datos propuesta por Bayer et al. (2022).

SMOTE.

El aumento de datos en el espacio de características implica generar transformaciones en la representación numérica del conjunto de datos de entrada. La técnica de sobre muestreo de minorías sintéticas (SMOTE por sus siglas en inglés) es una técnica de interpolación donde se realiza una búsqueda de múltiples vecinos próximos a una instancia particular dentro del espacio de características (Chawla et al. (2002)) con el propósito de ser interpolados utilizando la siguiente fórmula:

$$\tilde{x} = x_i + \lambda * dis(x_i, x_j) \tag{4}$$

Llevándose a cabo la interpolación entre instancias que pertenecen a la misma clase. La justificación para calcular vecinos con las mismas etiquetas de clase radica en que las interpolaciones tienden a mantener la clase original, lo que incrementa la confiabilidad de la técnica. Sin embargo, esta restricción limita la novedad y diversidad de las instancias generadas. En la figura 3 se observa de manera ilustrativa el proceso de aumento de datos con esta técnica.

Reemplazo por sinónimo.

Las transformaciones que esta técnica realiza se consideran que están a nivel de palabra. Reemplazo por sinónimo consiste en reemplazar una o más palabras escogidas de manera aleatoria, que no se consideren preposiciones, por uno de sus posibles sinónimos escogido aleatoriamente y se encuentra dentro de las técnicas de aumento fácil (*Easy Data Augmentation*). Este conjunto de técnicas demuestran una mejora, pero deben ser usadas según la tarea, ya que ciertas técnicas pueden hacer que pierdan el significado de la etiqueta original. (Wei & Zou (2019)), en la Figura 4 se puede observar un ejemplo.

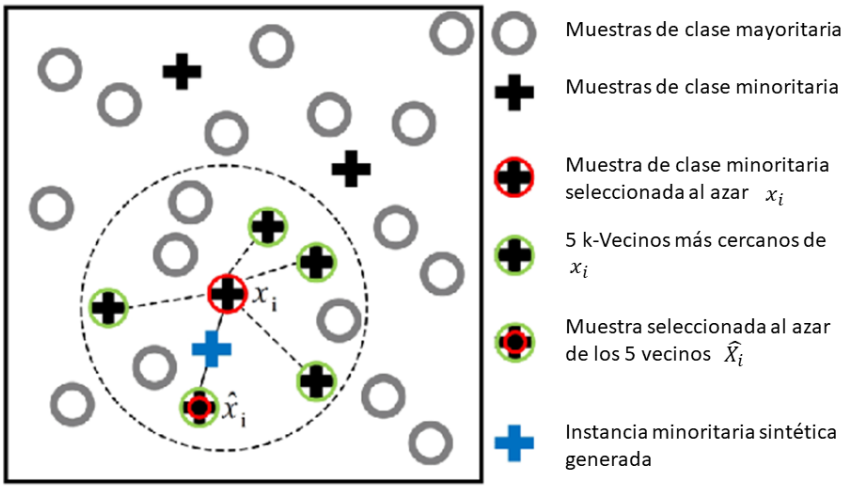


Figura 3. Ejemplo ilustrativo del proceso de la técnica SMOTE, (Ma et al. (2019))

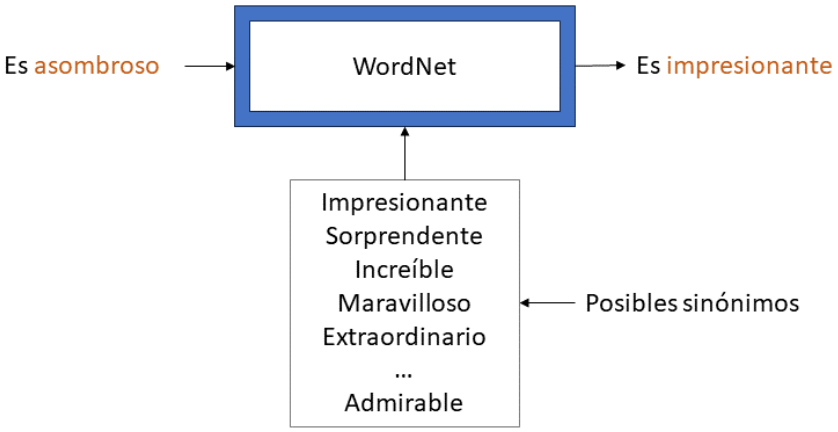


Figura 4. Ejemplo ilustrativo del proceso de la técnica Reemplazo por sinónimos.

Back translation.

Es un método en el que las transformaciones ocurren a nivel de documento, generando transformaciones que pueden ser en palabras, frases, oraciones o todo un documento. Esta técnica se basa en la traducción del texto, en un idioma origen, a un idioma objetivo y después traducir dicho texto al idioma origen. Con esto, se busca aprovechar la complejidad lingüística para generar textos con el mismo contexto que el original, pero aprovechando sinónimos y el parafraseo del texto original que pueda generarse (Aiken & Park (2010)), tal como se puede observar en la Figura 5.

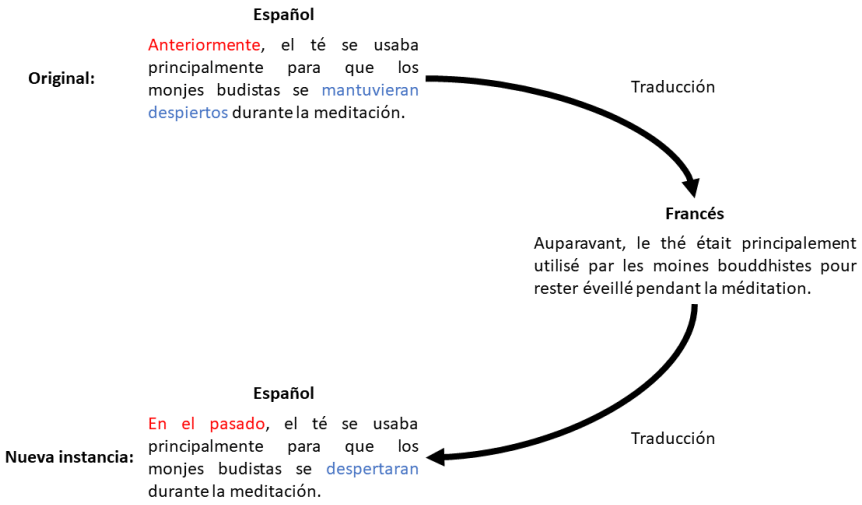


Figura 5. Ejemplo del proceso de *Back Translation*

El aumento de datos para imágenes comparte técnicas similares a las aplicadas para texto. Basándonos en la taxonomía propuesta por Mumuni & Mumuni (2022) (Figura 6), observamos que existen varias técnicas que se pueden englobar en dos principales grupos, transformaciones y síntesis.

Espacio de entrada.

En este espacio, las modificaciones se hacen directamente a las imágenes de entrada (similar al nivel instancia en el aumento de datos para texto). Las técnicas van desde tradicionales, como modificaciones geométricas (rotaciones, traslaciones y cortes) y fotométricas (ajuste de brillo, distorsión del lente e inversión de colores), hasta avanzadas, que realizan modificaciones en una región (como eliminar, reemplazar o cambiar), y píxeles.

Espacio de características.

En este conjunto de técnicas, se busca obtener una mayor representación de características diversas,

manipulando los vectores de características en las capas intermedias de las redes neuronales. Por ejemplo, técnicas de interpolación que permiten generar características adicionales de manera artificial basándose en la relación con características vecinos existentes.

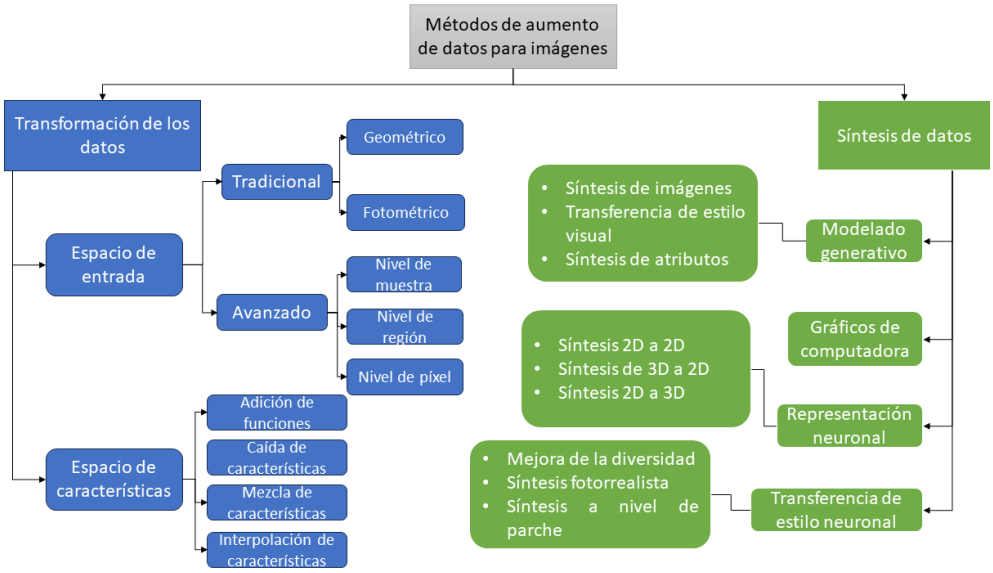


Figura 6. Taxonomía propuesta por Mumuni & Mumuni (2022) para el aumento de datos en el área de visión por computadora.

Modelado de gráficos por computadora.

Con la ayuda de herramientas de diseño asistido por computadora (CAD) se crean modelos 2D y 3D de objetos que se pueden usar como datos de entrenamiento. Los enfoques de modelado basados en motores de juegos 3D modernos pueden sintetizar escenas dinámicas muy grandes. El fundamento esencial del modelado de gráficos por computadora radica en la representación de la información de una escena mediante elementos geométricos fundamentales, como vértices y conjuntos de bordes que los conectan. Además de estas entidades geométricas básicas, también se definen las características funcionales de los parámetros de la escena, como materiales, iluminación, texturas y posición de la cámara.

2.2.4. Algoritmos y modelos de inferencia

- **Naive Bayes**

  Este es un modelo que utiliza el teorema de Bayes y que, como característica principal, asume la independencia entre las palabras y, a partir de los datos de entrenamiento, estima la probabilidad

de cada clase considerando los valores de las características presentes en un nuevo documento (Kowsari et al. (2019)).

- **k-Nearest Neighbour (kNN)**

  Este algoritmo busca las distancias entre las características del documento con sus vecinos más cercanos dentro del conjunto de vectores de documentos de entrenamiento, utilizando las etiquetas de clase de los *k vecinos* más similares para predecir la clase del documento de entrada y se otorga un peso a las clases de estos vecinos, basado en la similitud de cada vecino con el vector del documento (Aas & Eikvil (1999)).

- **Support Vector Machines (SVM)**

  Este algoritmo busca encontrar un hiperplano óptimo que pueda separar los datos de diferentes clases, maximizando el margen entre ellos. Este método solo se puede utilizar en tareas de clasificación binaria, lo cual implica que al utilizar este enfoque, la clasificación de texto debe abordarse como una serie de problemas de clasificación con dos categorías (Joachims (1998)).

- **Random Forest**

  Es una implementación de la estrategia bagging que utiliza un ensamble de árboles de clasificación. Este algoritmo combina múltiples árboles de decisión, donde cada árbol genera una predicción, y la predicción final es el consenso de las predicciones individuales (Breiman (2001)).

- **Perceptrón multicapa**

  Un perceptrón multicapa, también conocido como MLP (Multilayer Perceptron), es un tipo de red neuronal artificial que consta de múltiples capas de neuronas interconectadas. El perceptrón multicapa, siendo uno de los componentes fundamentales en el campo del aprendizaje profundo, es utilizado para resolver problemas de clasificación y regresión, entre otras tareas de procesamiento de datos. La estructura básica de un perceptrón multicapa consta de tres tipos de capas: la de entrada, donde los datos son introducidos en la red y cada neurona en esta capa representa una característica o atributo del conjunto de datos. Las capas ocultas: Estas son capas intermedias entre la capa de entrada y la capa de salida y puede haber múltiples de ellas, donde cada una consta de un número variable de neuronas. Estas capas ocultas son responsables de capturar patrones y relaciones más complejas en los datos. Finalmente, la capa de salida: Esta es la capa final de la red, donde se generan las predicciones o resultados. El número de neuronas en esta capa depende del tipo de problema que se esté abordando (Marius-Constantin et al. (2009)). Si se conecta arriba de la representación generada por modelo BERT en un enfoque *end-to-end* es posible realizar un ajuste fino.

- **Ajuste fino (Fine Tuning)**. Proceso en el cual, a partir de un modelo pre-entrenado, se ajustan sus pesos y parámetros utilizando un conjunto de datos específico a la tarea que se desea resolver. El ajuste de pesos se efectúa congelando las capas iniciales, en las que se suelen capturar características de bajo nivel y generales. Después, se modifican las capas finales (que están más cerca de la salida) o reemplazan para adaptarse a la tarea específica (estas capas son entrenadas utilizando el conjunto de datos de la tarea destino). Finalmente, durante el proceso de ajuste fino, los pesos y parámetros del modelo se ajustan utilizando el algoritmo de optimización. En la Figura 7 se puede observar de manera general el proceso de ajuste fino para la tarea de " Pregunta-Respuesta".

2.2.5. Clasificación

La clasificación de una tarea consiste en asignar automáticamente una o más categorías pre-definidas a elementos como textos. Para realizar una clasificación se suele utilizar un solo modelo de inferencia. Sin embargo, en algunos casos los modelos por sí solos no suelen presentar un buen rendimiento frente a una tarea por diversas cuestiones, y así como humanos solemos pedir opiniones sobre un tema antes de tomar una decisión, se pueden utilizar más de un modelo en conjunto que permita mejorar el rendimiento para resolver la tarea.

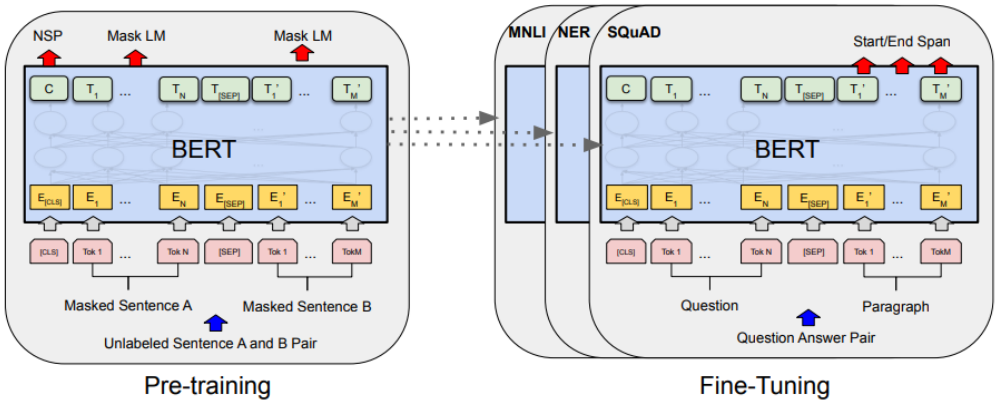


Figura 7. Proceso de ajuste fino del modelo BERT para la tarea de " Pregunta-Respuesta" con los conjuntos de datos MNLI, NER y SQuAD. Imagen extraída de Devlin et al. (2019)

2.2.5.1. Técnicas de ensamble

Un método por ensamble en aprendizaje automático es una técnica que combina múltiples modelos de aprendizaje para mejorar el rendimiento y la precisión en las predicciones. En lugar de depender de un solo modelo, los métodos por ensamble aprovechan la diversidad de los modelos individuales y buscan combinar sus resultados para obtener una predicción más robusta y generalizable.

Bagging.

Bagging, del acrónimo *Bootstrap agregating*, construye el conjunto de clasificadores haciendo réplicas por re-muestreo con reemplazo (*bootstrap*) del conjunto de entrenamiento y usarlas para entrenar diferentes clasificadores y sus predicciones se utilizan para obtener un predictor agregado. El predictor agregado se puede definir como la combinación de las predicciones de todos los modelos base. La combinación de las predicciones se puede realizar promediando las salidas en regresión, por mayoría o votación ponderada en problemas de clasificación (Re & Valentini (2012)).

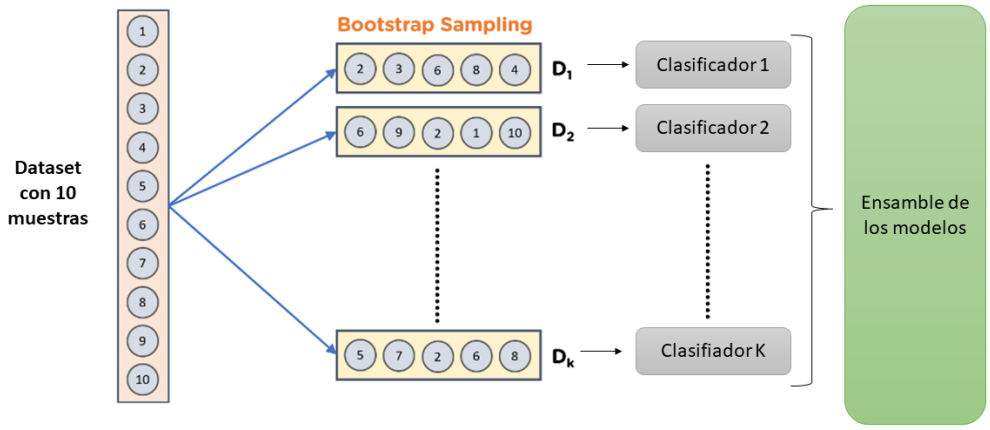


Figura 8. Técnica de ensamble *Bagging*.

Clasificación por método de cascada.

El enfoque en cascada es una forma específica de aprendizaje por ensamble en la cual se concatenan varios clasificadores, utilizando la salida de un clasificador como información adicional para el siguiente clasificador en la cascada. A diferencia de los métodos de ensamble por mayoría de votos o *stacking*, el método por cascada consiste en múltiples etapas. En la Figura 9 se observa el funcionamiento general de este tipo de ensamble.

2.2.6. Modelos de visión-lenguaje.

Un modelo de visión-lenguaje es una clase de modelos de inteligencia artificial que combina la capacidad de comprensión de imágenes (visión por computadora) con la comprensión del lenguaje natural (Alara & Sayak (2023)). Suelen construirse utilizando arquitecturas de redes neuronales que pueden procesar tanto imágenes como texto, fusionando estas modalidades en una representación compartida. Estos modelos pueden basarse en arquitecturas de codificador-decodificador o en arquitecturas basadas en atención como Transformers, los tres principales elementos son un codificador de imagen, de texto y una estrategia para combinar la información según la tarea deseada a resolver.

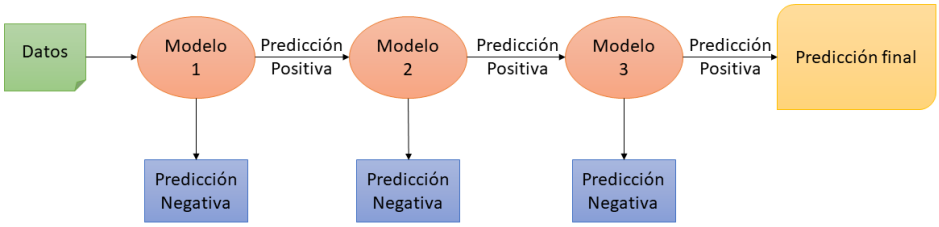


Figura 9. Diagrama de flujo cuando se sigue un enfoque de clasificación en cascada.

Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP).

BLIP (Li et al. (2022)) es un modelo de visión-lenguaje que ha sido preentrenado en tres tareas principales, donde dos de ellas se enfocan en la comprensión y la otra en la generación de contenido.

Su arquitectura se fundamenta en un Visual Transformer para el componente del codificador de imágenes. Este proceso implica dividir una imagen de entrada en parches y luego codificarlos en una secuencia de *embeddings*, a la cual se añade un token adicional [CLS] para representar las características globales de la imagen.

Para llevar a cabo las tres tareas objetivo, el modelo utiliza una mezcla multimodal de codificador-decodificador, un modelo multitarea que puede operar en una de las tres funcionalidades: (i) Codificador unimodal, que codifica por separado imagen y texto. El codificador de texto es el de BERT. (ii) Codificador de texto basado en imágenes, el cual proporciona información visual mediante la inserción de una capa adicional de atención cruzada entre la capa de *self-attention* y la capa *feed forward network* para

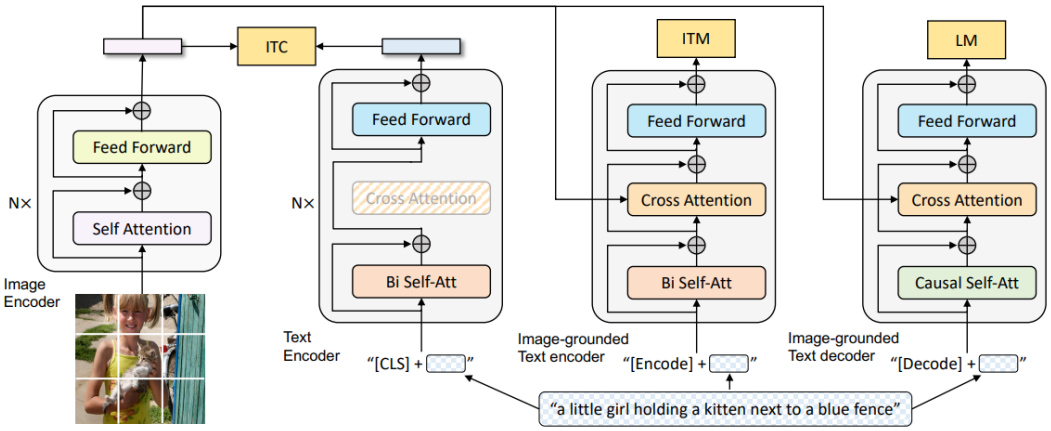
cada bloque transformador del codificador de texto. (iii) Decodificador de texto basado en imágenes, que reemplaza las capas bidireccionales de *self-attention* en el codificador de texto basado en imágenes con capas causales de *self-attention*.

Las primeras dos tareas relacionadas con la comprensión son *Image-Text Contrastive Loss* el cual activa el codificador unimodal y su objetivo es alinear el espacio de funciones de Visual Transformer y el Transformer para texto, alentando que los pares positivos de imagen y texto tengan representaciones similares en contraste con los pares negativos; *Image-Text Matching Loss* activa el codificador de texto basado en imágenes. Su objetivo es aprender la representación multimodal de texto e imagen que captura la alineación detallada entre la visión y el lenguaje. Y la tercera tarea *Language Modeling Loss*, activa el decodificador de texto basado en imágenes, cuyo objetivo es generar descripciones textuales dada una imagen.

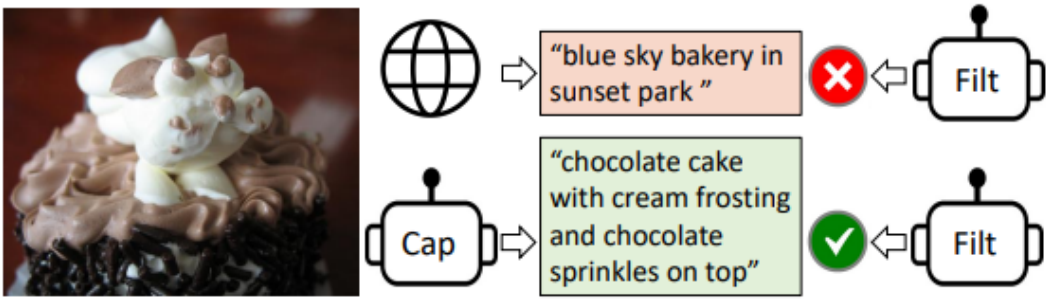
Adicionalmente, utiliza dos módulos, un descriptor para generar descripciones sintéticas a partir de imágenes web y un filtro para eliminar pares de imágenes y texto ruidosos. Dadas las imágenes web, el descriptor genera descripciones sintéticas con una descripción por imagen. El filtro es un codificador de texto basado en imágenes. Se afina con los objetivos *Image-Text Contrastive Loss* e *Image-Text Matching Loss* para saber si un texto coincide con una imagen. El filtro elimina los textos ruidosos tanto en los textos web originales como en los textos sintéticos, donde un texto se considera ruidoso si el *Image-Text Matching Loss* predice que no coincide con la imagen. En la Figura 10 se puede observar de manera general la arquitectura del modelo y un ejemplo de la función de filtrado.

2.3. Resumen

En este capítulo se definieron los diferentes conceptos, técnicas y herramientas que se utilizarán en los capítulos posteriores. Primero se definieron los conceptos generales que definen las clases en los conjuntos de datos utilizados y la tarea en que se basa. Después se mencionaron las técnicas estadísticas y computacionales utilizadas para llevar a cabo el presente trabajo de investigación, las cuales involucran la extracción de características, modelos de clasificación y técnicas de aumento de datos para el trabajo unimodal, utilizando solo texto, y multimodal, texto e imágenes.



(a) Arquitectura del modelo BLIP



(b) Ilustración del proceso de generación de descripciones sinteticas de imágenes y el modulo de filtrado.

Figura 10. Modelo BLIP y ejemplo al generar descripciones de imágenes (Li et al. (2022)).

Capítulo 3. Trabajo relacionado

En este capítulo se abordará el trabajo relacionado a la detección de eventos violentos en redes sociales. En la primera sección se presenta la relevancia que pueden tener las redes sociales para el análisis y predicción de patrones delictivos. En la segunda sección se abordan las técnicas y enfoques utilizados para la detección y clasificación de eventos violentos.

Para términos de este trabajo, un patrón delictivo se refiere a una serie o secuencia de delitos que comparten características similares en términos de modus operandi, ubicación, métodos utilizados o características de las víctimas. Por otro lado, un evento violento se definirá como un suceso el cual involucra algún tipo de violencia o repercusión física o psicológica en personas, grupos o incluso objetos de manera intencionada o no.

3.1. Detección de delitos en redes sociales

Varios estudios se han enfocado en la implementación y uso de la información proporcionada por redes sociales para el análisis y detección de delitos.

En el trabajo de Abbass et al. (2020), se desarrolló un marco de trabajo para predecir cibercrímenes en redes sociales. Recolecta 150 mil tweets relacionados con el acoso, intimidación, *hacking* y estafas dentro de las redes sociales. Realiza pruebas con los modelos Naive Bayes, kNN y SVM en conjunto con Bigramas como representación de los tweets en el espacio de características y la técnica TF-IDF para seleccionar las características más importantes. Sus resultados presentan una mayor exactitud que otros enfoques que utilizan redes neuronales para extraer características.

En Marivate (2015) realizan un análisis de datos, relacionado con delitos e incidentes de seguridad pública, en la región de Sudáfrica por medio de Twitter como fuente de información. Tras una recopilación de más de 60 mil publicaciones por medio de palabras clave y la extracción de temas por medio de la herramienta *Latent Dirichlet Allocation*, compararon y contrastaron cómo el monitoreo de cuentas oficiales sobre delincuencia y seguridad pública difiere del monitoreo de personas y organizaciones fuera del primer grupo. Entre los resultados que reportan, mencionan que los temas relacionados con accidentes de tráfico son los más reportados semana tras semana debido a la frecuencia e impacto que tiene en las personas, por ejemplo, causando retrasos en el transporte y siendo temas que las personas están más dispuestos a reportar.

Por otro lado, Piña-García & Ramírez-Ramírez (2019) realizaron un estudio en la Ciudad de México que busca comparar cómo las personas reportan un crimen de manera tradicional y en redes sociales por medio del tipo de crimen, definido en 13 categorías. Además, se incluyó la información geo espacial obtenida de los reportes del Departamento de Policía de la Ciudad de México y una recolección de publicaciones de la plataforma Twitter y Google Trends. Los resultados preliminares de esta investigación mostraron que la información de las redes sociales previamente analizadas puede ser integradas para mejorar los modelos basados en Big Data para predecir las tasas de criminalidad.

Asimismo, Mata et al. (2016) propone un enfoque para proporcionar estimaciones definidas por índices de criminalidad para generar rutas seguras en dispositivos móviles. Se utilizó un repositorio geo espacial para almacenar tweets relacionados con hechos delictivos de la Ciudad de México e informes oficiales que fueron geo codificados para obtener rutas seguras.

3.2. Detección de eventos violentos

3.2.1. Modalidad textual

Para tareas de clasificación, en (Sandagiri et al. (2020a)) utilizan la red social Twitter para detectar publicaciones relacionadas con delitos, en las categorías de robos, homicidios, ofensas sexuales, suicidio, relacionados con drogas y asaltos violentos. Los autores realizan su propia recolección de datos a través de la API de Twitter por medio de palabras claves que engloban los tipos de crímenes mencionados anteriormente dentro del idioma inglés. Proponen un enfoque de clasificación binario por medio del modelo BERT para identificar las publicaciones relacionadas con los crímenes. Al compararse con otros modelos de clasificación como SVM y dos redes neuronales con diferentes técnicas para la generación de vectores. Los resultados obtenidos fueron superiores con una precisión del 92,8 y F1-Score del 92.4.

Por otra parte, Sandagiri et al. (2020b) proponen un enfoque con aprendizaje máquina para detectar delitos (asaltos, robos, delitos sexuales, drogas, homicidio) y la ubicación de estos. Con una recolección de más de 10 mil tweets y aplicando técnicas de preprocesamiento a los textos, como TF-IDF, para extraer características, logran una exactitud del 88.52 % con SVM. Para obtener la ubicación que se menciona en las publicaciones, utilizan una extracción de términos relacionados a ubicaciones, donde finalmente eran enviados a una API para encontrar el país. A su vez, reporta problemas comunes en el área de PLN como faltas ortográficas y abreviaciones de ciudades, lo cual dificultaba su obtención.

Marivate & Moiloa (2017) exploran como diferentes tipos de características (de usuario, texto, tema y gráficos) afectan el rendimiento de clasificadores binarios de aprendizaje máquina con algoritmos como SVM, Random Forest y Regresión logística. La recolección de los tweets fue hecha de cuentas relacionadas con seguridad pública sudafricana, pero no relacionadas con cuentas de periódicos o similares. El total de tweets recuperados fue de 1299, relacionados con temas de robo, tiroteos y secuestros. Exploran *Self-Training* como aumento de datos, esta técnica consiste en entrenar los clasificadores con un conjunto de datos etiquetados y después clasificar tweets no etiquetados que terminan siendo anexados al conjunto de entrenamiento para ser entrenado nuevamente. Los mejores resultados, en términos de exactitud y F1-Score, fueron de 0.8 y 0.764 utilizando el clasificador de regresión logística.

3.2.2. Multimodal

La implementación conjunta de información textual y de imágenes puede mejorar los resultados para la clasificación en tareas como detección de violencia en fotos en redes sociales (Qin et al. (2022)), al brindar mayor contexto al modelo, permitiendo combatir la ambigüedad que puede presentarse en ciertos textos.

Un ejemplo de los beneficios de trabajar con información multimodal es el caso de Rodríguez-Bribiesca et al. (2021), donde hacen uso de un modelo adaptado de la versión Transformer multimodal para la clasificación de películas dentro de los géneros de comedia, acción, drama, familiar y horror. El modelo propuesto trabaja en conjunto con un módulo GMU (*Gated Multimodal Units*), las cuales son unidades de puertas altamente interpretables que deciden cómo cada modalidad influye en las unidades de activación de salida de la capa y, por lo tanto, deciden qué tan relevante es cada modalidad para hacer la predicción. Las modalidades que consideran son texto, audio, video, imagen y meta. El modelo fue comparado contra otros modelos como MulT-Concat y Fast Modal Attention. Los resultados mostraron que el modelo propuesto, Multimodal Transformer-GMU, presenta mejores resultados cuando las modalidades de texto y visuales son incorporadas.

Arriaga et al. (2017) presentan un conjunto de datos de detección de anomalías con el propósito de aplicaciones robóticas, como el patrullaje y servicios domésticos, así como el diseño e implementación de una arquitectura de aprendizaje profundo que clasifica y describe situaciones peligrosas utilizando solo una imagen como entrada. El conjunto de datos que armaron se encuentra formado por 1008 pares de imágenes y descripciones correspondientes a situaciones peligrosas que contengan ventanas rotas, gente

lastimada, peleas, accidentes de carros y fuego, armas y violencia doméstica. Para sus experimentos realizaron un clasificador binario al extender el modelo Neural Image Caption (NIC) al incorporar un CNN-LSTM y Multilayer Perceptron (MLP) y fue puesto a prueba mediante una partición del conjunto de datos de 80 % para entrenar y 20 % para pruebas, fue evaluado con las métricas de exactitud y METEOR obteniendo resultados del 97 % y 16.2 respectivamente.

Por otro lado, con un enfoque más cercano, Fatichah et al. (2020) busca resolver la detección de incidentes, no necesariamente de índole violento, publicados en redes sociales. Aprovecha la naturaleza multimodal de las redes sociales, enfocándose en la información textual y de imágenes por medio de redes neuronales profundas, específicamente, variantes de los modelos CNN y LSTM como C-LSTM para texto y la red preentrenada VGG16 para imágenes. Su propuesta fue comparada contra los modelos AlexNet, SqueezeNet y VGG19 para la sección de imágenes y para el caso de texto contra una arquitectura CNN. Los resultados fueron obtenidos por medio de dos modelos independientes, en el cual, el resultado final se obtuvo mediante el nivel más alto de confianza de predicciones de texto o imagen. Obteniendo resultados de 99.09 % y 99.08 % en exactitud para el modelo de texto e imágenes respectivamente en contraste con el 98.95 % por parte del modelo CNN y 98.33 % con SqueezeNet.

3.3. IberLEF: DA-VINCIS

Por otra parte, trabajos orientados en el idioma español para tareas de clasificación se encontraron aquellos relacionados con los participantes del evento organizado por IberLEF bajo el nombre de DA-VINCI. IberLEF es una campaña de evaluación compartida de sistemas de Procesamiento del Lenguaje Natural en español y otras lenguas ibéricas, y se celebra en el marco del congreso anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Su objetivo es promover la investigación en tareas de procesamiento, comprensión y generación de textos en al menos una de las siguientes lenguas ibéricas: Español, Portugués, Catalán, Vasco o Gallego[1].

Motivado por la falta de recursos en el idioma español, DA-VINCIS ("Detection of Aggressive and Violent Incidents from Social Media in Spanish") es una tarea para IberLEF la cual se enfoca en la detección de eventos relacionados con la violencia en Twitter, considerando solo texto(DA-VINCIS 2022) y texto e imágenes (DA-VINCIS 2023). Esta tarea se divide en dos subtareas:

[1] http://sepln2023.sepln.org/iberlef/, Accedido 26/Junio/2023

- Identificación de eventos violentos. Consiste en determinar si un tweet se encuentra asociado a un incidente violento o no (subtarea 1, clasificación binaria).
- Reconocimiento de categorías de eventos violentos. Reconocer la categoría del crimen al que pertenece un tweet dado (subtarea 2, clasificación multiclase).

3.3.1. DA-VINCIS 2022

Esta edición del evento se enfoca en la información textual proveniente de Twitter para resolver dos subtareas relacionadas con la detección de eventos violentos. La primera subtarea consiste en una clasificación binaria (violento y no violento) mientras que la segunda se enfoca a una clasificación multiclase, la cual, incorpora 5 categorías donde 4 de ellas se encuentran relacionadas con categorías específicas de violencia.

La edición de DA-VINCIS del año 2022 consiste en utilizar la información textual proveniente de los tweets para resolver alguna o ambas subtareas. En esta edición se cuentan con 5 clases para la subtarea 2 las cuales se definen a continuación:

- Accidente: Evento o acción eventual que tiene como resultado un daño involuntario a personas o cosas.
- Homicidio: Privación de la vida.
- Robo: Apoderamiento o destrucción dolosa de bienes ajenos sin derecho y sin consentimiento de la persona que legalmente puede disponer de ellos.
- Secuestro: Privación de libertad.
- Ninguno de los anteriores: Cuando no se reporta ningún delito en el tweet. Los tweets de esta categoría también se recuperaron usando palabras clave asociadas con eventos violentos.

En la tabla 1, se observa la distribución de los datos en cada una de las categorías correspondientes. Como se observa, la cantidad de datos es relativamente pequeña y cuando se toma en cuenta la subtarea 1 (clasificación binaria) no se presenta un desbalance de datos considerable. Sin embargo, cuando se aborda la subtarea 2 (clasificación multiclase) la distribución de los datos cambia considerablemente,

presentándose muy pocas instancias para cada nueva categoría que sustituyen a la anterior categoría "violento" que englobaba a todas estas y a su vez, desfavoreciendo algunas clases como secuestro y favoreciendo a otras como accidente. Esto puede ser causado dado lo mencionado en Prieto Curiel et al. (2020) y además de que los accidentes pueden reportarse con mayor frecuencia en las publicaciones por su naturaleza de llegar a afectar de manera más general a otras personas.

Tabla 1. Número total de instancias para cada categoría, utilizados en el entrenamiento

Subtarea	Categoría	# Instancias
1,2	No violento	1825
1	Violento	1587
2	Accidente	1137
2	Homicidio	265
2	Robo	184
2	Secuestro	47

En el caso de los participantes del evento DA-VINCI, varios autores utilizaron técnicas de aumento de datos para abordar la problemática con la distribución de los datos mencionada anteriormente. La técnica más utilizada, como se puede ver en la tabla 2, fue *back translation*, la cual consiste en traducir un texto de idioma fuente a otro texto en diferentes idiomas objetivos, y eventualmente traducirlo al idioma original, de esta manera se obtienen textos parafraseados del original. Se destaca que todos los participantes usaron algún tipo de modelo Transformer preentrenado, el cual les permitió aprovechar el conocimiento generado en su pre-entrenamiento para cubrir hasta cierto punto el problema de la cantidad de datos para los modelos de aprendizaje profundo.

Los organizadores del evento establecieron la métrica F1-score para evaluar los modelos en ambas tareas. Como se observa en la tabla 2, Vallejo-Aldana et al. (2022) obtuvieron el mejor resultado para la tarea 1 (clasificación binaria) con un enfoque en aprendizaje multitarea y técnicas de ensamble. Los autores utilizaron una tarea secundaria diferente para cada uno de los 3 modelos entrenados y realizando la predicción por medio de un esquema de votación por mayoría. Mientras que Qin et al. (2022) obtuvieron el mejor resultado para la tarea 2, utilizando un enfoque en aprendizaje rápido (*Prompt learning*), en el que se transmite la información de un modelo de lenguaje previamente entrenado en la tarea de reconocimiento por categorías de eventos violentos (subtarea 2), construyendo una representación conjunta de textos y etiquetas. Dentro de los equipos que utilizaron técnicas de aumento de datos, Montañés-salas & Peña-larena (2022) usaron la técnica de *back translation* y emplearon una estrategia de ajuste de hiperparámetros, para así mejorar el rendimiento de los modelos. Se empleó un enfoque de

votación sobre las predicciones de cada modelo, obteniendo en ambas tareas el tercer mejor rendimiento.

Tabla 2. Técnicas de aumento de datos, modelos y resultados de los participantes en DA-VINCI 2022.

Publicación	Aumento de Datos	Modelo	F1-Score subtarea 1	F1-Score subtarea 2
Ta et al. (2022b)	Back Translation	BERT-Base-Multilingual-cased	0.748	0.392
Tonja et al. (2022)	Forward Translation	DistilBETO	0.7455	0.4903
Vallejo-Aldana et al. (2022)	N/A	BERT	**0.775**	0.4731
Turón et al. (2022)	Back Translation	BERT,BETO,MarIA	0.773	0.528
Montañés-salas & Peña-larena (2022)	Back Translation	BERT/BETO, Twitter-XLM-Roberta, BSC-Roberta	0.765	0.504.
Qin et al. (2022)	N/A	BERT/BETO	N/A	**0.554**
Ta et al. (2022a)	Back Translation	GAN-BERT	0.744	N/A
García-díaz et al. (2022)	N/A	BERT/BETO, MarIA	0.764	0.469
Baseline	-	BERT/BETO	0.75	0.57

Por otro lado, Turón et al. (2022) además de aplicar un aumento de datos sintético, realizaron una reducción de ruido en la base de datos original, llevando a cabo un proceso de reetiquetado de los datos de entrenamiento al considerar los votos de 5 sistemas aprendidos de la base de datos ruidosos originales. La corrección sobre las etiquetas se realizó si al menos 4 de los 5 sistemas estaban de acuerdo.

3.3.2. DA-VINCIS 2023

La edición de DA-VINCIS del año 2023 consiste en utilizar la información textual y visual como imágenes provenientes de los tweets para resolver alguna, o ambas, de las subtareas. En esta edición se cuentan con 4 clases para la subtarea 2 las cuales se definen a continuación:

- Accidente: Evento o acción eventual que tiene como resultado un daño involuntario a personas o cosas.
- Asesinato: Privación de la vida.

- Robo: Apoderamiento o destrucción dolosa de bienes ajenos sin derecho y sin consentimiento de la persona que legalmente puede disponer de ellos.
- Ninguno de los anteriores: Cuando no se reporta ningún delito en el tweet. Los tweets de esta categoría también se recuperaron usando palabras clave asociadas con eventos violentos.

En la edición más reciente del evento DA-VINCIS, se brindó la libertad de trabajar de manera unimodal, considerando solo texto o imágenes, o multimodal, abordando estrategias que permitan aprovechar la información proporcionada por el texto e imágenes de los tweets.

Vallejo-Aldana et al. (2023) abordan el problema mediante múltiples enfoques multimodales, el primero es combinando las salidas de los vectores de representación de los modelos por separado y el segundo uniendo descripciones textuales de cada una de las imágenes relacionadas con un texto. El primer enfoque consiste en la concatenación de la salida del modelo Inception-V3 con modelos de texto separado. El modelo Inception V-3 se encuentra pre-entrenado en el conjunto de datos de ImageNET y adaptado a esta tarea, el cual se encarga de extraer y clasificar las imágenes asociadas a cada tweet; para los modelos que procesan el texto utilizan dos modelos RoBERTa, uno para procesar el texto de los tweets y otro para procesar el texto de las descripciones de las imágenes obtenidas por el modelo BLIP. El vector de características del modelo de descripciones, el de texto y el de la salida del modelo Inception-V3 se concatenan en un solo vector que luego se pasa a una cabeza de clasificación que está constituido por un perceptrón multicapa cuyas salidas son las probabilidades de clase para cada tweet. El segundo enfoque consiste en usar las descripciones de imágenes obtenidas del modelo BLIP y el texto de los tweets en la misma oración para entrenar un solo modelo usando un separado especial para el modelo RoBEERTa. A su vez, utilizan aumento de datos basado en el sobre muestreo de las categorías menos representadas, replicándolas con base en el criterio establecido por los autores. Y finalmente buscaron mejorar el desempeño de los modelos utilizando técnicas de ensamble, el cual consistió en entrenar cinco modelos por separado para cada subtarea, variando la inicialización de los pesos y promediando las probabilidades de las clases obtenidas de las salidas del modelo. Mediante el enfoque por ensamble y utilizando las descripciones de las imágenes obtienen el mejor rendimiento en sus experimentos, además de conseguir el primer lugar para la subtarea 1 con un F1-Score de 0.9264 y 0.8421 para la subtarea 2, utilizando el ensamble de modelos.

Gutiérrez-Megías et al. (2023) utiliza una arquitectura que se basa en el uso de dos modelos pre-entrenados que tokenizan y extraen las características de las imágenes y textos por separado. Los resultados obtenidos se almacenan por separado para ser utilizados como entrada a los modelos pre-

entrenados donde las salidas de estos modelos se concatenarán en un solo tensor, que será la entrada para las capas de clasificación. La tarea de procesamiento de texto se basa en el modelo RoBERTa y BEiT para la extracción de características visuales. Adicionalmente, utiliza técnicas de aumento de datos para ambos tipos de información para abordar el problema de desbalance de clases. Para el texto utiliza *back translation* y transformaciones geométricas y de color para las imágenes, especialmente un giro horizontal seguido de un giro vertical, convertir la imagen a monocromática, lo que da como resultado una representación en escala de grises y aplicar un efecto de desenfoque a la imagen. Los resultados obtenidos fueron de 0.9165 y 0.8733 en F1-Score para la subtarea 1 y 2 respectivamente.

Hernández-Minutti et al. (2023) abordan el problema desde la perspectiva unimodal, considerando solo texto, con ayuda de diferentes algoritmos de aprendizaje máquina y profundo, como regresión logística, SVM, Naive Bayes, Perceptrón multicapa y XGBoost. Para la representación del texto en el espacio de características utilizaron varios enfoques, el primero basado en frecuencia, una bolsa de palabras con peso binario y TF-IDF. Después de una búsqueda exhaustiva de los mejores hiperparámetros, utilizaron un enfoque de ensamble basado en *soft voting* donde descartaron a los modelos Naive Bayes y al Perceptron multicapa debido a su bajo rendimiento comparado con los demás modelos. Los resultados obtenidos fueron de un F1-Score de 0.882 y 0.804 para la subtarea 1 y 2 respectivamente.

Zatarain Cabada et al. (2023) utilizan solo la información textual proveniente del tweet y exploraron diferentes representaciones de características para texto como bolsa de palabras, el cual se trabajó en conjunto con el modelo de EvoMSA y representaciones contextuales utilizando modelos basados en Transformers como BERT, para la subtarea 1, y RoBERTa, para la subtarea 2. Después de una optimización de hiperparámetros con los modelos BERT y RoBERTa, los resultados finales consistieron en un F1-Score para la subtarea 1 de 0.897 y para la subtarea 2 de 0.765. Por otro lado, Rubio et al. (2023) utilizan información textual, abordan el problema utilizando la versión en español de BERT llamado BETO, dentro de su preprocesamiento se encuentran la corrección de errores de ortografía y amplían las siglas y abreviaturas presentes en los textos. Con un ajuste de hiperparámetros obtienen el tercer lugar entre todos los participantes con un F1-Score de 0.918 y 0.869 para la subtarea 1 y subtarea 2.

Con base a lo mencionado anteriormente, se observa que la técnica de aumento de datos más utilizada para abordar el desbalance de clases en ambas ediciones de DA-VINCI fue principalmente *back translation*, dejando otras técnicas y enfoques sin explorar. Por lo tanto, el explorar otras técnicas y estrategias enfocándonos en los datos puede permitir una mejor generalización y un menor sobre ajuste al abordar este tipo de tareas. Por otro lado, observando la Tabla 3, la mayoría de los equipos, sin contar el baseline, optaron por un enfoque unimodal considerando la información textual de los tweets, si bien obtuvieron un

desempeño competitivo con los mejores modelos, el incluir ambos tipos de información puede permitirle al modelo ver otros puntos clave que no podría si solo considerará un tipo de información, por lo cual se considera relevante incluirla y buscar la mejor estrategia que permita aprovechar las diferentes naturalezas de la información de los tweets.

Tabla 3. Resultados de los participantes y modalidad utilizada en DA-VINCI 2023.

Publicación	**Aumento de Datos**	**Modalidad**	**Modelo**	**F1-Score subtarea 1**	**F1-Score subtarea 2**
Vallejo-Aldana et al. (2023)	Sobremuestreo de datos aleatorios	Multimodal	Ensamble: RoBERTa	0.926	0.842
Hernández-Minutti et al. (2023)	-	Unimodal: texto	SVM,Naive Bayes, Regresión logistica, MLP,XGBoost	0.882	0.8036
Zatarain Cabada et al. (2023)	-	Unimodal: texto	EvoMSA, BERT, RoBERTa	0.896	0.765
Gutiérrez-Megías et al. (2023)	Back Translation, transformaciones geometricas	Multimodal	RoBERTa,BEiT	0.917	0.873
Rubio et al. (2023)	-	Unimodal: texto	BERT/BETO, Twitter-XLM-Roberta, BSC-Roberta	0.919	0.869.
Baseline	-	Multimodal	BERT, ViT	0.895	0.843

Además, los conjuntos de datos presentados por el evento de DA-VINCIS 2022 y 2023, tienen un aporte considerable para la comunidad de habla español al existir una falta de recursos en el idioma que aborde dichas tareas, además de tener varios anotadores y criterios para su etiquetado y reflejar en cierta manera lo que se pública o habla, en este caso, en Twitter con respecto al reporte de hechos violentos.

3.4. Resumen

En este capítulo se presentaron estudios relacionados con el uso de la información proveniente de redes sociales para resolver o proponer soluciones que reduzcan la incidencia delictiva. A su vez, trabajos relacionados con la detección de eventos violentos, utilizando texto y de manera multimodal, fueron

presentados junto con su enfoque y resultados. Finalmente, se presentaron los trabajos relacionados con el conjunto de datos que se utilizaron en este trabajo que sirven para comparar nuestros resultados con otras estrategias utilizadas.

Capítulo 4. Métodos

En este capítulo se describen los dos métodos propuestos para llevar a cabo los experimentos pertinentes para el cumplimiento de los objetivos planteados en el capítulo 1. En primer lugar, se describe el método seguido en la primera fase de experimentación, que involucra solamente información textual. Finalmente, se describe el método para abordar el problema considerando información multimodal proveniente del texto de los tweets e imágenes asociadas a ellos.

4.1. Unimodal: Texto

El procedimiento de experimentación que se siguió durante esta fase se muestra en la Figura 11.

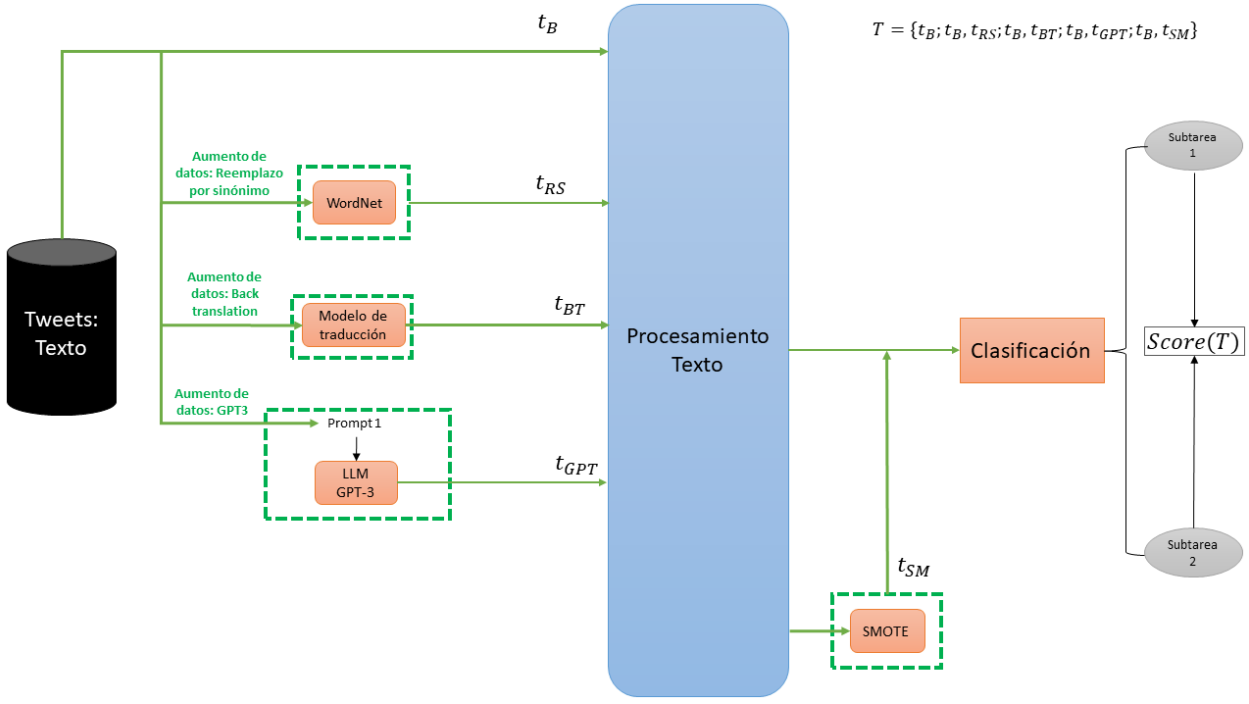


Figura 11. Proceso experimental seguido para resolver las subtareas propuestas en DA-VINCI 2022 y determinar la mejor técnica de aumento de datos para el dominio textual. Donde: T es la combinación de los textos; t_B es el texto original proveniente de los tweets; t_{RS} son los textos que resultaron de aplicar la técnica de reemplazo por sinónimo; t_{BT} son los tweets obtenidos al aplicar back translation; t_{SM} son las nuevas características obtenidas al aplicar la técnica SMOTE; t_{GPT} los tweets sintéticos obtenidos al utilizar el modelo GPT-3.

4.1.1. Procesamiento

El texto proveniente de redes sociales contienen muchas palabras y caracteres que no responden bien a los métodos típicos para la extracción de características, por ejemplo, *stop words*, puntuación, ortografía incorrecta, etc. y su presencia puede tener un efecto adverso en el desempeño de un algoritmo de inferencia en una tarea dada. La limpieza o procesamiento del texto que se realizó fue basada en Naseem et al. (2021), realizando un número de pasos mayor o menor según el enfoque de aprendizaje a seguir.

Para los algoritmos de aprendizaje máquina, el procesamiento de texto consistió en pasar todo el texto a minúsculas, remover puntuaciones, dobles espacios, stop words, números, emoticones, URLs, retweets, menciones de usuarios, caracteres especiales y finalmente, se procedió a lematizar las palabras. Por otro lado, para el enfoque en aprendizaje profundo se pasó todo el texto en minúsculas para evitar problemas por sensibilidad entre palabras en mayúsculas y minúsculas, se removieron URLs, ya que no presentan ninguna información relevante con el método que buscamos seguir, se eliminaron dobles espacios y el símbolo de menciones ('@'), caracteres especiales, emoticones, retweets y hashtags, este último solo fue el símbolo '#' puesto que se considera que puede tener información relevante. Esta diferencia en el procesamiento del texto para los diferentes tipos de aprendizajes se debe a la capacidad de los modelos, por ejemplo, en los modelos de aprendizaje profundo como redes neuronales recurrentes (RNN) o modelos basados en Transformers tienen la capacidad de aprender automáticamente representaciones complejas y jerárquicas, capturando relaciones sutiles y semánticas en el lenguaje. Por otro lado, los modelos de aprendizaje máquina es necesario utilizar técnicas para extraer y seleccionar características y patrones de los datos de texto. En la Figura 12 se observa el efecto a nivel instancia de realizar los diferentes tipos de limpieza en un tweet.

Extracción y selección de características

El siguiente paso fue construir la representación en el espacio de características del texto y, en el caso de aprendizaje máquina, realizar una selección de características debido a la alta dimensionalidad que resulta de utilizar las técnicas mencionadas en 2.2.1.1.

Para obtener la representación los algoritmos de aprendizaje máquina, se decidió trabajar con una bolsa de palabras basado en unigramas con un peso binario, indicando solamente si una palabra se encuentra en el texto o no, y además con la técnica TF-IDF. Después de obtener la representación en el espacio de características, se realizó una selección de características con la técnica de χ^2 conservando el 50 % de las características, donde se presentó un mejor rendimiento en general.

Para el caso de aprendizaje profundo, se generaron representaciones contextualizadas de los textos con el modelo BERT, específicamente BETO. Se utilizó los *embeddings* proporcionados por los encoders de este mismo modelo, los cuales ofrecen mayor información contextual que otro tipo de *embeddings*, por ejemplo Word2Vec, además se agregó el token especial '[CLS]' el cual es necesario para la tarea de clasificación.

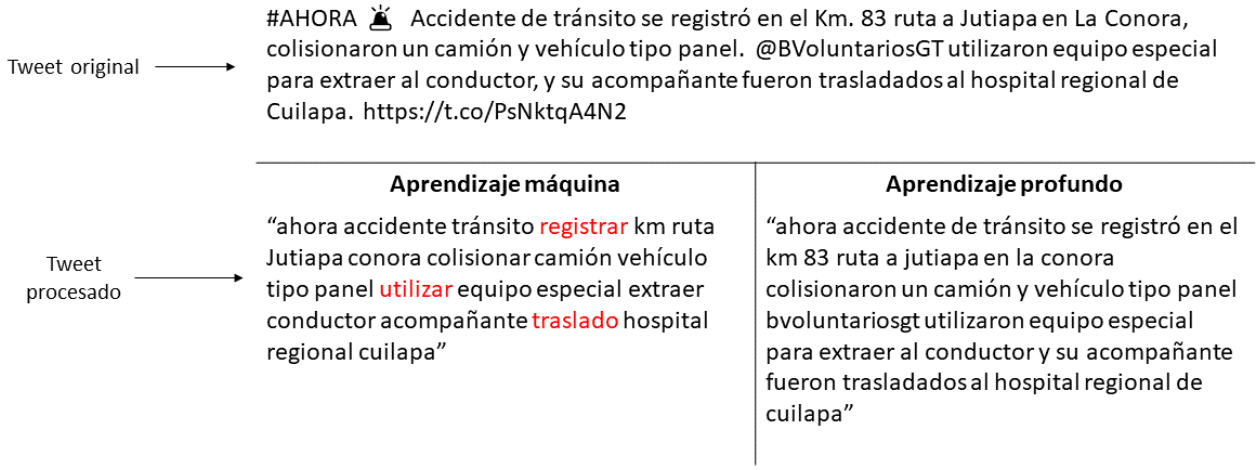


Figura 12. Ejemplo del resultado de aplicar el proceso de limpieza en los tweets cuando se utiliza un enfoque en aprendizaje máquina y profundo.

4.1.2. Aumento de datos

Debido al desbalance en los conjuntos de datos de ambas versiones de DA-VINCI, se decidió explorar diferentes técnicas de aumento de datos aplicadas a las categorías menos representadas. Primero, se exploró, con base en el conjunto de datos de DA-VINCI 2022, qué técnica permitía obtener un mejor rendimiento en el dominio de texto. Basándonos en los resultados obtenidos en la modalidad textual, se buscó la mejor alternativa para el aumento de imágenes que se asociara mejor para ser usada en el conjunto de datos de DA-VINCI 2023.

SMOTE.

Para la subtarea 1, se decidió igualar ambas clases, mientras que para la subtarea 2 se optó por realizar un aumento del 40 % para las clases de robo y homicidio y del 10 % para la clase de secuestro, basándose en la clase con mayor número de instancias que, en este caso, es la clase no violenta.

Para las siguientes técnicas de aumento de datos se decidió realizar el aumento de datos basándonos en

el número de instancias de las clases minoritarias de la subtarea 2. El aumento se realizó en cada clase por separado y finalmente se incorporaron al conjunto de datos base.

Reemplazo por sinónimo.

Para esta técnica se utilizó *Wordnet* en español como base de datos de referencia para los sinónimos a considerar. Se reemplazó una palabra por su sinónimo como mínimo y se duplicaron y triplicaron el número de instancias base para los experimentos.

Back Translation.

Al explorar esta técnica, se utilizó 2 idiomas, los cuales fueron el inglés y alemán. Se utilizó el modelo de traducción MarianMT[1] y se obtuvieron 3 grupos de datos sintéticos: las nuevas instancias provenientes del idioma en inglés, del idioma en alemán y la combinación de estas dos.

Modelos de lenguaje grande - Generative Pre-trained Transformers 3 (GPT-3).

Por otro lado, con el potencial de los modelos de lenguaje grande para resolver diferentes tareas en el área del procesamiento del lenguaje natural, se decidió hacer uso de esta herramienta para explorar su capacidad como técnica de aumento de datos. Otros autores han usado esta herramienta con este propósito, siguiendo un enfoque reformulando cada oración en el conjunto de entrenamiento en múltiples muestras conceptualmente similares, pero semánticamente diferentes (Dai et al. (2023)).

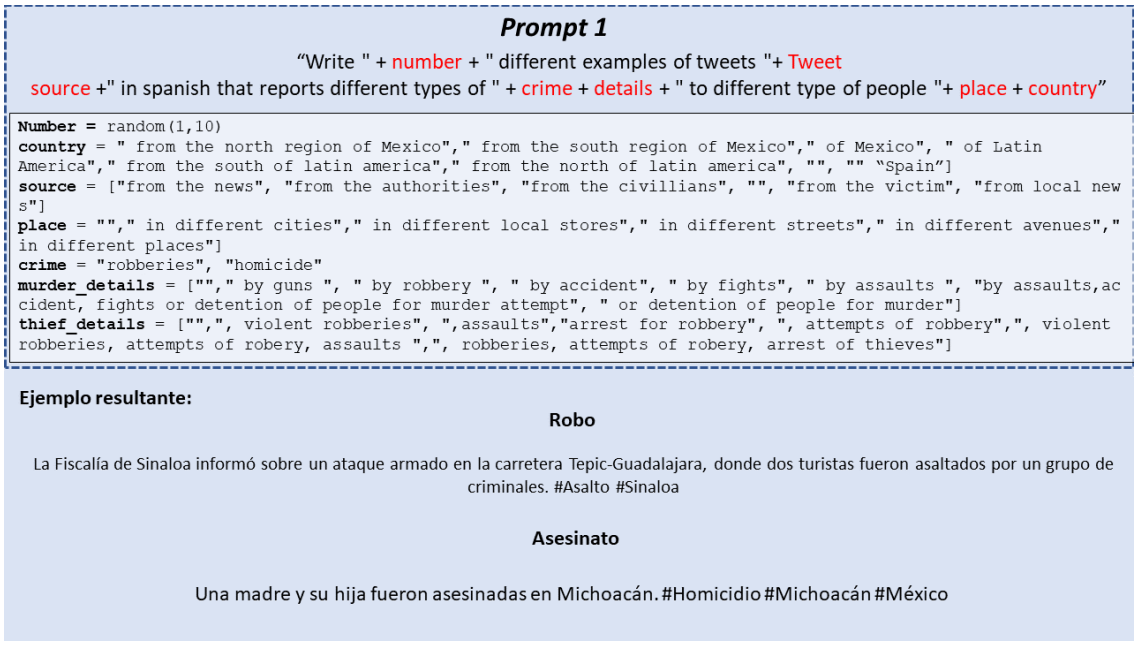


Figura 13. Prompt 1 utilizado para obtener nuevas instancias utilizando los modelos Davinci-003 de la familia GPT-3.

[1]https://huggingface.co/docs/transformers/model_doc/marian

En este trabajo se decidió explorar el aumento de datos utilizando el modelo de lenguaje grande de la familia GPT-3 llamado Davinci-003. A diferencia de Dai et al. (2023), el enfoque que se decidió utilizar fue a través de *prompt engineering* el cual consiste en diseñar, optimizar y refinar las indicaciones utilizadas para comunicarse con los modelos de lenguaje. El prompt que se utilizó consta de una estructura fija en la cual varias palabras clave van cambiando con cada petición, por ejemplo, el número de instancias, la ubicación, el tipo de crimen y posibles detalles relacionados con este. Alguno de los ejemplos que genera este prompt, así como la estructura base, pueden verse en la Figura 13.

4.1.3. Clasificación de eventos violentos

Para realizar la tarea de clasificación de hechos violentos, se utilizaron las representaciones en el espacio de características mencionadas en la sección 2.2.1 y dos modelos independientes, uno para cada tarea y para la tarea multiclase se utilizó una estrategia *OneVsRestClassifier* para los algoritmos de aprendizaje máquina, las consideraciones para cada modelo fueron las siguientes:

- **k-Nearest Neighbors (kNN)**

  Las características utilizadas fueron aquellas extraídas por medio de la técnica TF-IDF con la distancia de Minkowski considerando el número de vecinos igual a 5.

- **Naive Bayes(NB)**

  Se utilizó la versión multinomial, ya que es adecuada para características discretas como el recuento de palabras. Para este modelo se utilizaron las características extraídas por bolsa de palabras con un Alpha igual a uno.

- **Random Forest (RF)**

  Las características utilizadas fueron las proporcionadas por TF-IDF con un número de estimadores igual a 100.

- **Support Vector Machines (SVM)**

  Para este clasificador, se utilizaron dos tipos de características. El primero fue a través del uso de bolsa de palabras y el segundo utilizando los *embeddings* proporcionados por el modelo BERT al momento de explorar la técnica SMOTE, basándonos en Jiang et al. (2017). Esto último es debido a que el enfoque que seguimos con BERT es un proceso *end to end* y no es posible mantener las mismas condiciones si se desea aplicar la técnica SMOTE. Para este clasificador se hizo una

búsqueda del mejor kernel posible (sin cambiar los parámetros por defecto), para la clasificación binaria se utilizó un kernel sigmoide mientras que para la clasificación multiclase un kernel lineal.

- **Bidirectional Encoder Representations from Transformers (BERT)**

  Al utilizar un enfoque *end to end*, se utilizó la representación por vectores contextualizados generado por el modelo BERT y se conectó con una capa densamente conectada para realizar la tarea de clasificación. Los parámetros utilizados por este modelo estuvieron basados en Devlin et al. (2019), los cuales consistieron en entrenar por 3 épocas, utilizar un batch size de 32, learning rate de 4e-5, un optimizador Adam con una épsilon de 1e-6 y un learning reate scheduler tipo lineal.

4.2. Multimodal: Texto e imágenes

Por otra parte, el procedimiento que se siguió para considerar la información proveniente del texto del tweet y de las imágenes asociadas a este en la publicación original, puede ser observado de manera general en la Figura 14.

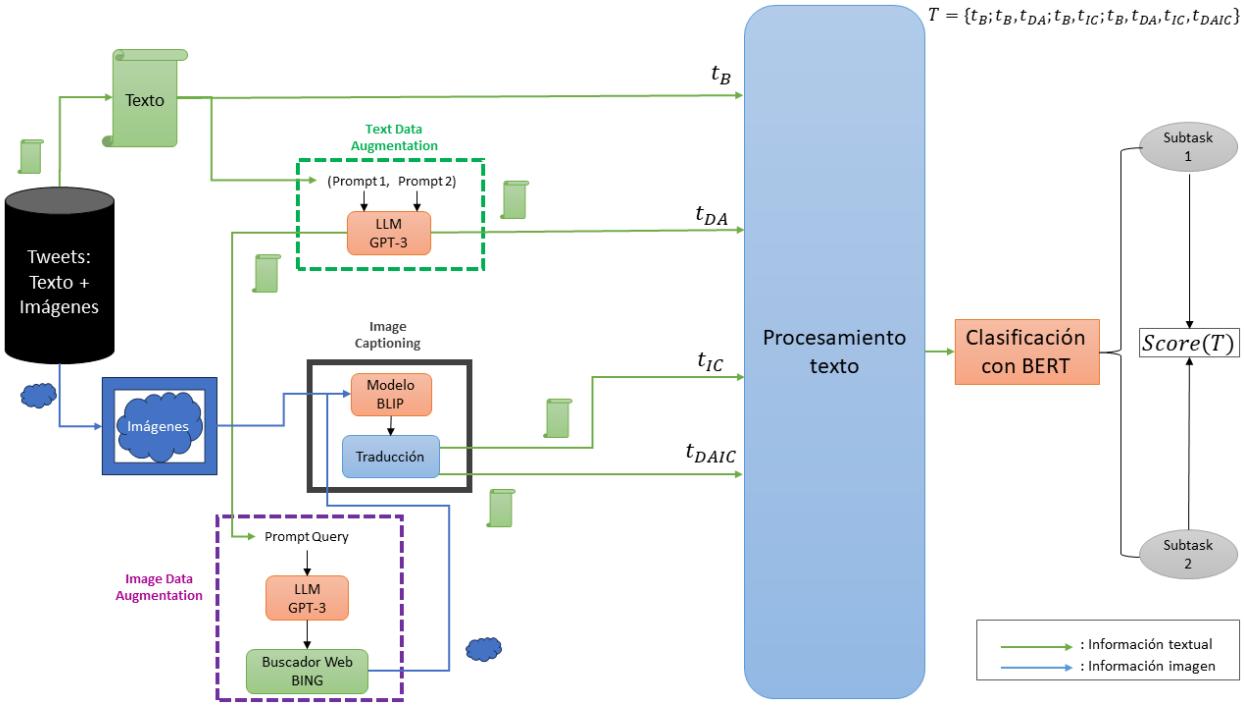


Figura 14. Proceso experimental seguido para resolver las subtareas propuestas en DA-VINCI 2023. Donde: T es la combinación de los textos; t_B es el texto original proveniente de los tweets; t_{DA} son los textos generados por GPT-3, t_{IC} son los textos obtenidos del proceso para obtener la descripción de imágenes y t_{DAIC} son los textos obtenidos del proceso de descripción de imágenes usando las imágenes recuperadas del paso de aumento de datos.

4.2.1. Procesamiento

Para el preprocesamiento de limpieza y construcción de la representación en el espacio de características se realizó el mismo procedimiento descrito en la sección 4.1.1 para el aprendizaje profundo.

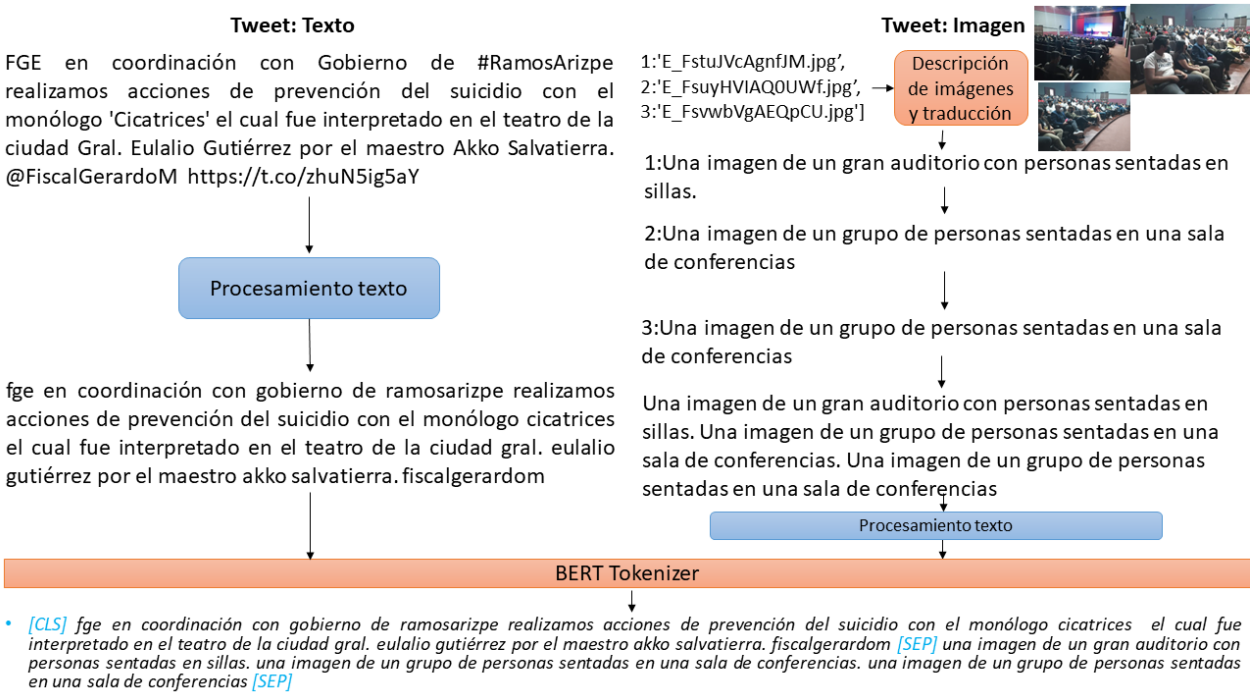


Figura 15. Ejemplo del resultado de aplicar el proceso de para obtener la descripción de imágenes.

Descripción de imágenes.

El método que se propone para tomar en cuenta la información de las imágenes es a través del uso del modelo de visión y lenguaje BLIP, el cual proporciona de manera automática descripciones textuales de las imágenes y con esto, trabajar todo desde el mismo dominio textual. Las descripciones proporcionadas por el modelo BLIP pasaron por un procesamiento previo al mencionado en la sección 4.1.1. Primero, debido a que se puede llegar a presentar una situación en que la relación entre el texto e imágenes de los tweets no sea 1 a 1, es decir, que existan varios tweets con más de una imagen asociada. El segundo inconveniente consiste en que las descripciones proporcionadas por el modelo se encuentran en el idioma inglés y último, algunas descripciones arrojadas por el modelo llegan a contar con algunos errores como la palabra 'arafed', presentándose mayormente en imágenes relacionadas con hechos violentos, o palabras repetidas. Para resolver lo anterior, el procesamiento consistió en dividir aquellas imágenes asociadas a un solo tweet y pasarlos por el modelo para obtener sus descripciones, para evitar los errores

mencionados anteriormente se condicionó la salida del modelo para que inicie con '*an image of...*'. Después, se realizó una traducción de todas las descripciones y se concatenaron en el mismo orden en el que se encontraban originalmente. En la Figura 15 se puede observar de manera visual este proceso con un ejemplo (No violento) con varias imágenes del corpus DA-VINCI 2023 y como es que la descripción de las imágenes se incorporan con el texto proveniente del tweet.

4.2.2. Aumento de datos

Grandes modelos de lenguaje - Generative Pre-trained Transformers 3 (GPT-3).

Para el aumento de datos utilizando el modelo de lenguaje de gran tamaño GPT-3, además de utilizar las instancias generadas por el prompt 1 con el modelo Davinci-003 mostrado en la Figura 13 se decidió explorar con una variante de este prompt utilizando el modelo Turbo 3.5 (ChatGPT), el cual nos referimos como prompt 2. Este nuevo prompt se considera una versión relativamente más estática en el cual se agregaron más instrucciones y detalles al *prompt*, manteniendo variable ciertas palabras clave para mantener la diversidad de los tweets sintéticos, por ejemplo, solo cambiaban parámetros como el país, tipo de crimen y número de instancias. Ejemplos de instancias generadas por este nuevo prompt, así como su estructura base, pueden verse en la Figura 16.

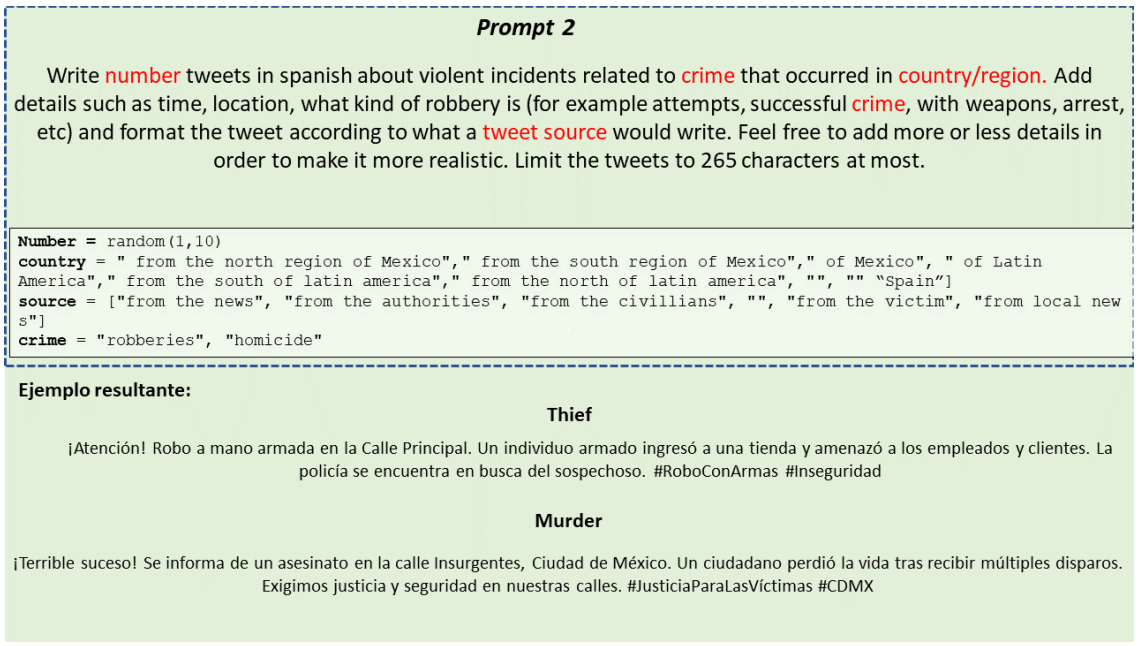


Figura 16. Prompt 2 utilizado para obtener nuevas instancias utilizando el modelo Turbo 3.5 de la familia GPT-3.

Imágenes.

Para mantener el mismo formato de texto-imagen(es) que se puede presentar al momento de trabajar con información multimodal, se decidió realizar un aumento de imágenes que permitiera acoplar a los tweets sintéticos generados por los modelos de lenguaje grande. Para lo anterior, se trabajó con un enfoque de recuperación de imágenes por la web, ya que los tweets generados no se encuentran directamente relacionadas con las instancias originales, por lo que trabajar con las imágenes base y realizar alguna transformación sobre ellas no proporcionaría una mayor diversidad de palabras.

Para obtener las imágenes de la web se utilizó el buscador Bing[2]. Para realizar las búsquedas se extrajeron palabras clave del texto con ayuda del modelo davinci-003 y un prompt llamado *query*, y después se prosiguió a descargar la primera imagen del resultado mediante un procedimiento automatizado en Python, en la Figura 17 se pueden observar algunos resultados de este proceso.

Tweet: Texto	Query	Imagen
Una víctima fue dejada herida tras un asalto a una tienda en la Ciudad de México. #Asalto #Mexico	Asalto en la Ciudad de México	
La policía de Puebla reportó un ataque armado en el centro de la ciudad, donde dos hombres armados asaltaron a varios transeúntes. #Asalto #Puebla	Ataque armado Puebla	
¡Alerta! Robo a transeúnte en la Avenida Central. Una mujer fue víctima de un robo violento por parte de un ladrón en motocicleta. Mantengamos la precaución en espacios públicos.#RoboEnVíaPública #Cuidado	Robo a transeúnte Avenida Central	
¡Atroz crimen en la tienda de conveniencia de la colonia Roma! Una mujer de mediana edad perdió la vida en un intento de robo violento. Exigimos justicia y mayor vigilancia para garantizar la seguridad de los ciudadanos. #ViolenciaEnLasCalles #México	Crimen tienda conveniencia colonia Roma México	

Figura 17. Ejemplo de imágenes obtenidas al realizar una recuperación de imágenes basada en palabras clave.

4.2.3. Clasificación de eventos violentos

Para desarrollar la tarea de clasificación mediante un enfoque multimodal se utilizaron dos modelos independientes, uno para cada subtarea, que trabajaran en el dominio textual (BERT/BETO con un

[2]https://pypi.org/project/bing-image-downloader/

enfoque end to end). Adicionalmente, para experimentos finales se trabajó con dos modelos relacionados mediante un método de ensamble por cascada, dónde en la primera etapa un modelo de clasificación binaria se encarga de clasificar las instancias consideradas como reportes de eventos violentos y otro modelo entrenado previamente en las clases de accidente, robo y homicidio, se encarga de discernir entre estas.

4.3. Resumen

En este capítulo se describió la metodología que se llevó a cabo para el procesamiento del conjunto de datos, la extracción de características, las técnicas de aumento de datos, las técnicas computacionales para el entrenamiento de los modelos de aprendizaje máquina y profundo, y finalmente, la estrategia para abordar la tarea de clasificación de eventos violentos con un enfoque multimodal.

Capítulo 5. Resultados

En este capítulo se presentan los resultados de los experimentos de este trabajo, así como los resultados oficiales en DA-VINCI 2022 y DA-VINCI 2023. Primero se muestran los resultados obtenidos durante la evaluación cruzada para cada una de las técnicas, después los resultados obtenidos utilizando el conjunto de prueba para realizar una comparación con los participantes y trabajo relacionado con esta tarea. Finalmente, se discuten los resultados obtenidos para cada una de las ediciones de DA-VINCI. Debido a que los organizadores definieron el puntaje de F1-Score, de la clase positiva para la subtarea 1 y macro F1-Score de la clase positiva para la subtarea 2, como la métrica principal para determinar el mejor modelo, a lo largo de este capítulo se le estará dando mayor énfasis a esta métrica en los gráficos presentados además de ser la media armónica de la precisión y recuerdo del modelo.

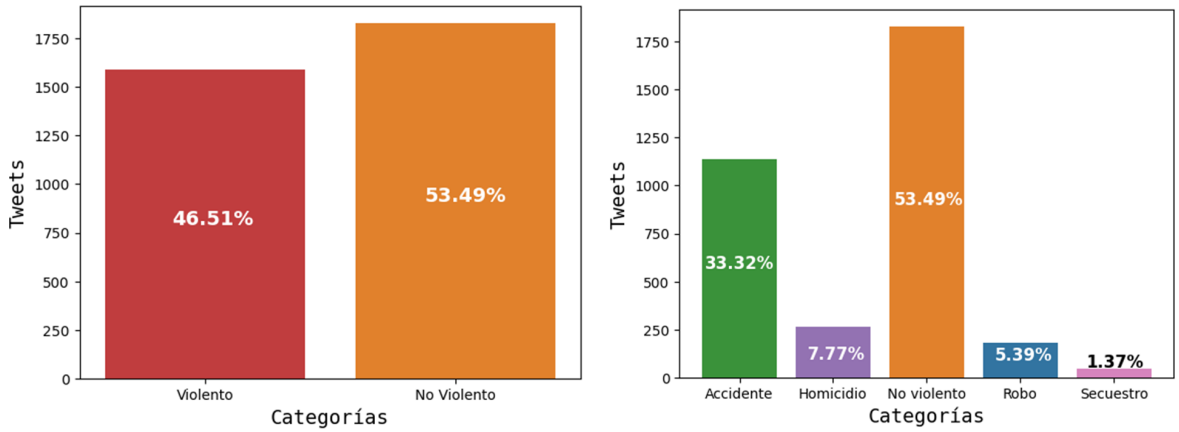
Para realizar las diferentes pruebas y determinar el mejor modelo y técnica de aumento de datos se decidió realizar una validación cruzada de 4-Folds para ambas versiones de DA-VINCI y de esta manera, observar el comportamiento general de los algoritmos de inferencia en este tipo de tareas en conjunto con las diferentes técnicas para la extracción de características y el aumento de datos.

5.1. DA-VINCI 2022.- Modalidad textual

El conjunto de datos base para esta edición consta de un total de 3412 instancias disponibles para el entrenamiento, en la Figura 18 se puede observar que para la subtarea 1 la distribución de los datos entre clases es cercana, mientras que para la subtarea 2 se puede ver un claro desbalance entre las clases que para la subtarea 1 conforman la categoría de violento, especialmente para la clase de secuestro, robo y homicidio.

Para poder obtener un mejor entendimiento del conjunto de datos, se obtuvo una representación visual del conjunto de datos utilizando la técnica T-SNE basada en uni gramas al ser el tipo de extracción de características base y otra utilizando los embeddings proporcionados por el modelo BERT. En la Figura 19a se puede observar que las clases no son fácilmente distinguibles, al estar varias de estas encimadas en el espacio de características. Esto se explica, ya que los organizadores para la categoría de 'No violento' recuperaron instancias usando palabras clave asociadas con eventos violentos, pero no se encuentran reportando un evento violento. Por otra parte, al utilizar embeddings proporcionados por el modelo BERT, se puede ver en la Figura 19b que existe una mejor agrupación de las características que, aunque aún existe un traslape entre las clases (especialmente entre No violento, 'Robo' y 'Homicidio'),

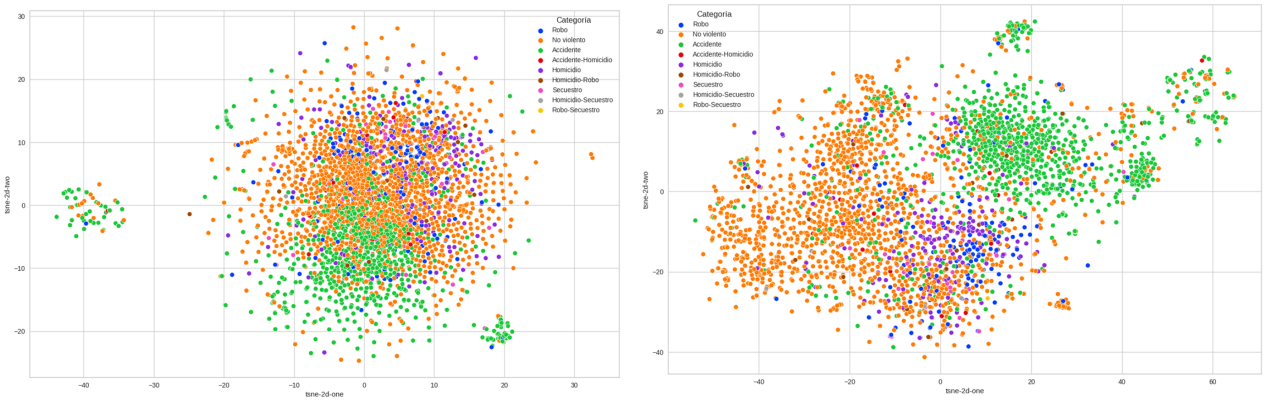
las categorías se encuentran con una mejor agrupación en comparación con usar características extraídas por uni gramas. Lo anterior se puede observar con mayor detalle entre las clases de 'Accidente' y 'No violento' mientras que 'Robo', 'Homicidio' con menor detalle.



(a) Proporción de las clases para la subtarea 1.

(b) Proporción de las clases para la subtarea 2.

Figura 18. Distribución de los datos para cada subtarea en DA-VINCI 2022



(a) Representación visual por TSNE al considerar un enfoque por uni gramas.

(b) Representación visual por TSNE al considerar embeddings proporcionados por el modelo BERT.

Figura 19. Representación visual del conjunto de datos por medio de la técnica TSNE para el conjunto de entrenamiento DA-VINCI 2022.

Se construyeron y evaluaron una serie de modelos basados en aprendizaje máquina y profundo descritos previamente en el capítulo 4.1.3. Los resultados obtenidos aplicando el conjunto de datos original a los diferentes modelos pueden ser observados en la Figura 20. Se observa que, en cuanto al criterio de los organizadores, el modelo BERT se posiciona entre los modelos con mejor rendimiento con un F1-Score de 0.777 (sd = 0.0149) para la subtarea 1 y 0.478 (sd = 0.008), seguido por Naive Bayes con 0.746 (sd = 0.005) en la subtarea 1 y 0.443 (sd = 0.024) en la subtarea 2, en la mayoría de las métricas

consideradas por los organizadores (precisión, recuerdo y principalmente F1-Score).

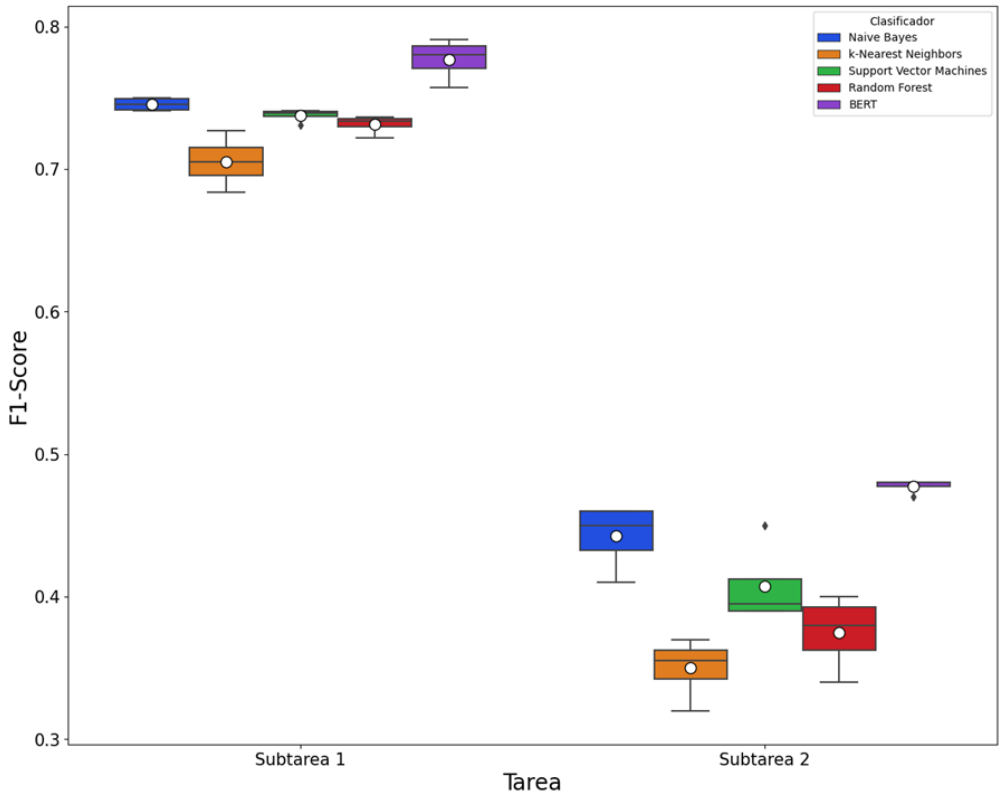


Figura 20. Resultados de los modelos utilizando el conjunto de datos sin aumento de datos.

Debido a que lo anterior se repite en la mayoría de los casos, en las siguientes secciones se reportan solo los resultados obtenidos por el modelo BERT y se retomaran los modelos de aprendizaje máquina que presentaron el mejor rendimiento en la sección 5.1.5. El rendimiento de los demás modelos puede ser consultado en Anexos .

5.1.1. Aumento de datos: SMOTE

Para esta técnica, como se mencionó en la sección 4.1.3, se decidió utilizar los embeddings generados por los encoders BERT y combinarlos con un modelo de aprendizaje máquina que en este caso se decidió por Support Vector Machines con un kernel RBF para la subtarea 1 y lineal para la subtarea 2 al ser los que mejores resultados presentaron. Los resultados obtenidos, utilizando solamente la técnica SMOTE y al

combinarla con una a nivel instancia (*back translation*), se compararon con los obtenidos con el modelo BERT utilizando el conjunto de datos base y pueden observarse en la figura 21.

En cuanto a la subtarea 1, los resultados al momento de solo aplicar la técnica SMOTE son de un F1-Score de 0.774 (sd = 0.022) con una precisión de 0.769 (sd = 0.0239) y recuerdo de 0.779 (sd = 0.029). Al combinar esta técnica con *back translation* se obtuvieron las métricas de F1-Score 0.769 (sd = 0.01), precisión de 0.746 (sd = 0.021) y recuerdo de 0.795 (sd = 0.016). En la Figura 21 se puede observar que el combinar las técnicas *back translation* y SMOTE no tuvo mejoras considerables con respecto al resultado base y utilizando solamente SMOTE como aumento de datos en F1-Score.

En la subtarea 2 se obtuvo un F1 Score de 0.413 (sd = 0.01) con una precisión de 0.41 (sd = 0.008) y recuerdo de 0.418 (sd = 0.015). Al incorporar la técnica *back translation* se obtuvo un F1-Score de 0.405 (sd = 0.017), precisión de 0.408 (sd = 0.017) y recuerdo de 0.408 (sd = 0.017).

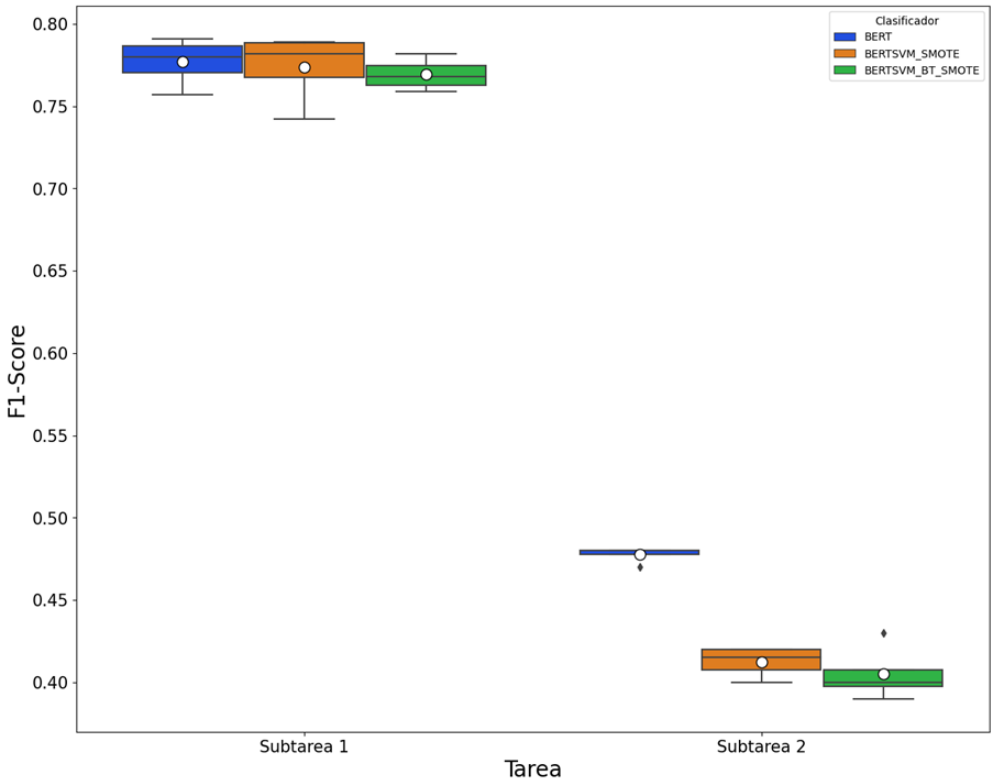


Figura 21. Resultados obtenidos en la validación cruzada utilizando SMOTE como técnica de aumento de datos.

5.1.2. Aumento de datos: Reemplazo por sinónimo

Se exploró el efecto de esta técnica duplicando ('SRx1') y triplicando ('SRx2') el número de instancias de las clases menos representadas, los resultados obtenidos para esta técnica pueden observarse en la Figura 22. Los resultados para la subtarea 1, duplicando las instancias, fueron F1-Score de 0.764 (sd = 0.011), precisión de 0.763 (sd = 0.004) y recuerdo de 0.766 (sd = 0.025). Al triplicar el número de instancias se obtuvo un F1-Score de 0.751 (sd = 0.024), precisión de 0.765 (sd = 0.01) y recuerdo de 0.738 (sd =0.047). Se observa una disminución del rendimiento F1-Score y recuerdo entre los resultados obtenidos al duplicar y triplicar el número de instancias con excepción de la precisión, aunque con una mayor dispersión.

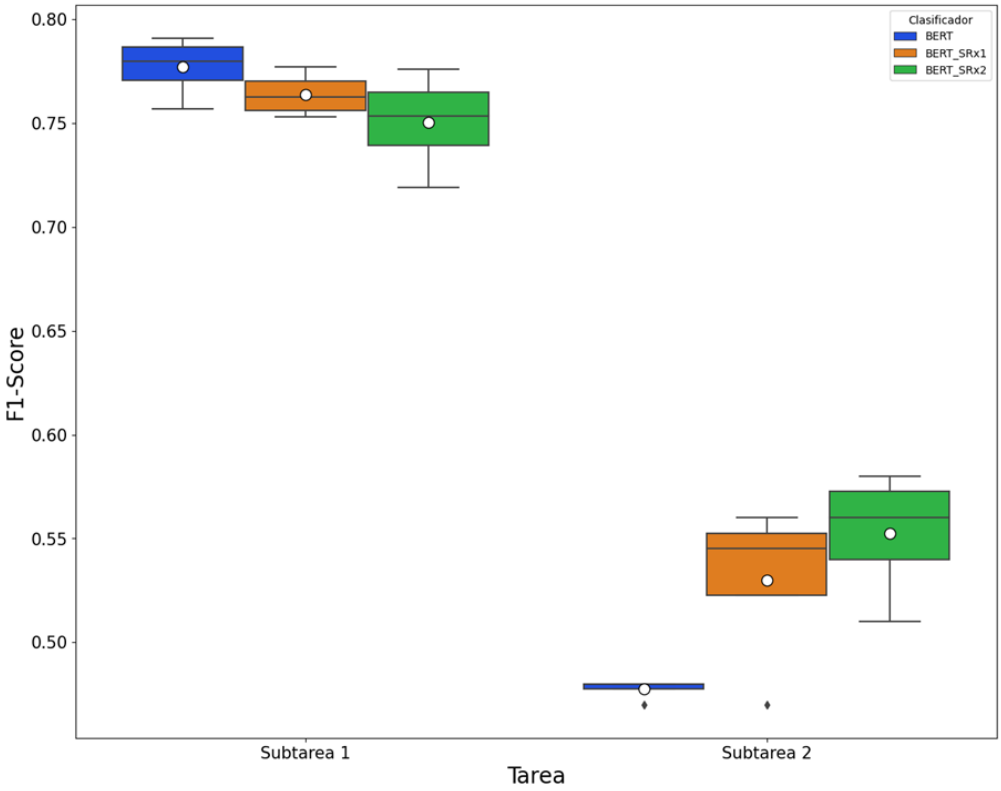


Figura 22. Resultados obtenidos en la validación cruzada al aplicar reemplazo por sinónimos como aumento de datos.

Por otro lado, para la subtarea 2 los resultados al momento de duplicar las instancias fueron de F1-Score 0.53 (sd = 0.041), precisión 0.625 (sd = 0.072) y recuerdo de 0.485 (sd = 0.031). Al triplicar las instancias minoritarias los resultados fueron de F1-Score 0.553 (sd = 0.031), precisión de 0.653 (sd = 0.022) y recuerdo de 0.513 (sd = 0.035). En este caso, se tiene un mejor desempeño para ambos

casos con respecto al resultado base del modelo, como se puede observar en la Figura 22, teniendo un desempeño superior al triplicar el número de instancias en las tres métricas.

5.1.3. Aumento de datos: Back translation

Al momento de explorar esta técnica se tomaron en cuenta los idiomas inglés y alemán. Se utilizaron dos combinaciones, las cuales consistieron en utilizar cada idioma por separado y finalmente combinar las instancias resultantes de ambos idiomas en un solo conjunto, de esta manera obteniendo una proporción dos y tres veces mayor a la original.

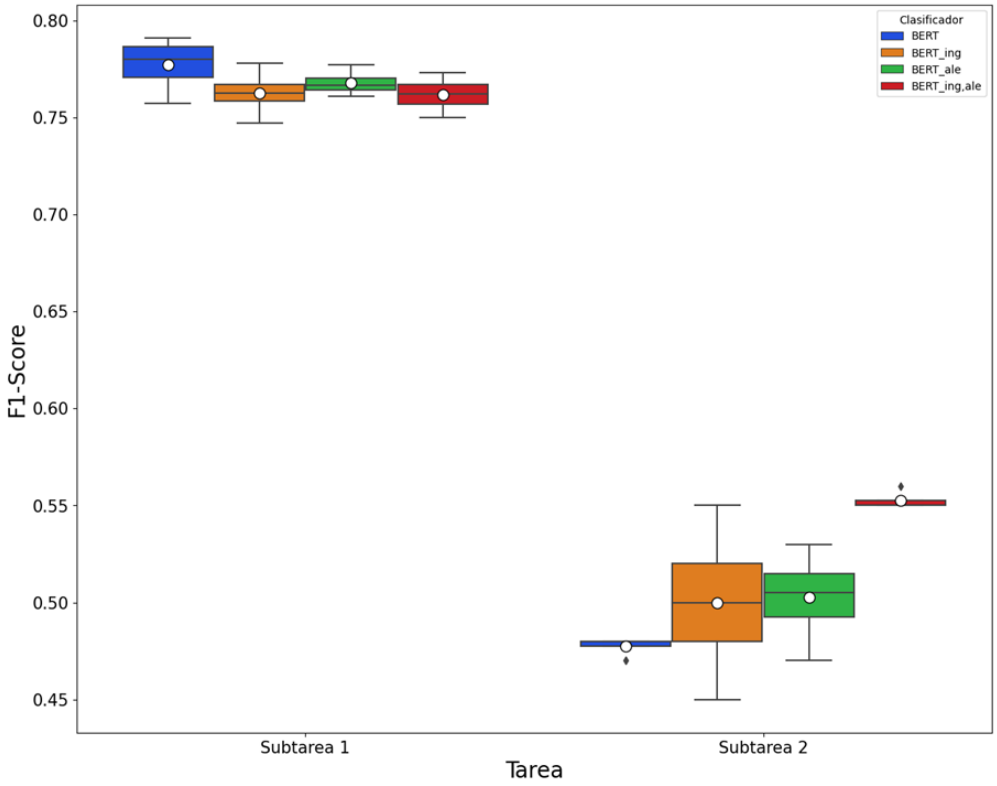


Figura 23. Resultados obtenidos en la validación cruzada utilizando *back translation* como aumento de datos.

Los resultados que se obtuvieron para la subtarea 1, considerando el aumento con el idioma inglés, fueron un F1-Score de 0.763 (sd = 0.013), precisión de 0.734 (sd = 0.011) y recuerdo con 0.795 (0.033). En cuanto al alemán, F1-Score de 0.768 (sd = 0.007), precisión de 0.742 (sd =0.009) y recuerdo de 0.795 (sd = 0.02). Al momento de considerar ambos idiomas en un solo conjunto, los resultados fueron de un

F1-score de 0.762 (sd =0.01), precisión de 0.73 (sd = 0.02) y recuerdo de 0.798 (sd = 0.025). Con base en los resultados anteriores y lo visto en la Figura 23, podemos observar que el idioma alemán para el aumento de datos proporciona un mejor rendimiento que utilizar solo inglés o combinar ambos conjuntos de datos sintéticos. En cuanto a la métrica de interés, no se obtuvo un desempeño superior, aunque se puede observar un aumento en la métrica de recuerdo al utilizar este tipo de aumento de datos pero con una mayor dispersión que el recuerdo base.

En la subtarea 2, los resultados considerando solo el idioma inglés fueron de 0.5 (sd =0.042) para F1-Score, 0.568 (sd = 0.073) en precisión y en recuerdo, 0.463 (sd = 0.033). En alemán se obtuvo un F1-Score de 0.502 (sd = 0.025), una precisión de 0.575 (sd = 0.068) y 0.465 (sd = 0.017) en recuerdo. Al momento de combinar las instancias generadas por ambos idiomas (y eliminar duplicados) se obtuvo un F1-Score de 0.553 (sd = 0.005), precisión de 0.605 (sd = 0.013) y 0.515 (sd = 0.01) en recuerdo. En la Figura 23 se puede observar de manera más detallada los resultados obtenidos, se observa una mejora en cada métrica con cada combinación de instancias sintéticas, obteniendo el mejor desempeño al combinar los dos idiomas tanto en promedio como dispersión.

5.1.4. Aumento de datos: GPT-3

Como se menciona en la sección 4.1.2, los resultados mostrados en esta sección son basándonos en el *Prompt 1* mostrado en la Figura 13 y donde algunos ejemplos de instancias generadas por este prompt pueden observarse en la Figura 24. Al utilizar el modelo GPT-3 como una herramienta de aumento de datos se exploró aumentar las instancias hasta un total de 7 veces, sin embargo, los mejores resultados se presentaron al realizar un aumento de 1 (GPTx1) y 3 (GPTx3) veces. Por lo tanto, los resultados obtenidos por estas últimas proporciones son los que se describen a continuación (ver Figura 25).

En la subtarea 1, los resultados al momento de duplicar las instancias minoritarias consistieron en un F1-Score de 0.783 (sd = 0.006), una precisión con 0.756 (sd = 0.006) y 0.812 (sd = 0.014) en recuerdo. Al evaluar el modelo considerando 3 veces más el número de instancias base, los resultados para F1-Score fueron de 0.772 (sd = 0.016), para precisión 0.74 (sd = 0.012) y para la métrica de recuerdo 0.807 (sd = 0.025). En la Figura 25 se observa como al utilizar esta técnica el recuerdo mejora para ambos enfoques (aumentando 1 y 3 veces más) mientras que en F1-Score permite al modelo mejorar en promedio y reducir su dispersión con respecto a los resultados base y al triplicar el número de instancias base cuando aumentamos una vez.

Clase	Instancia recuperada
Homicidio	"Informamos que en el estado de Baja California se registró un homicidio por arma de fuego donde fue víctima una mujer. #JusticiaParaTodos"
Secuestro	'Se persigue a los implicados en el secuestro de un bebé en una tienda local de Veracruz. ¡Pedimos la colaboración del público! '
Accidente	'Esta mañana tuve un accidente con mi bicicleta cuando salía para el trabajo. ¡Menos mal que no me pasó nada! #accidente #bicicleta'
Robo	'#Alerta: En la ciudad de #Caracas, se reportan varios intentos de robos a transeúntes y tiendas locales del norte de Latinoamérica. Mantenga la precaución y denuncie cualquier actividad sospechosa.'
No violento	'El gobierno debe moverse rápido para combatir el calentamiento global antes de que sea demasiado tarde #cambioclimatico '

Figura 24. Ejemplos de instancias generadas por GPT-3 para cada clase. A pesar de que se muestran ejemplos de las 5 clases, en este experimento solo se emplearon las pertenecientes a homicidio, robo y secuestro por tener menor representación en el conjunto de datos.

Los resultados para la subtarea 2 utilizando esta herramienta consistieron, al aumentar una vez, en un F1-Score de 0.578 (sd =0.01), una precisión de 0.643 (sd = 0.017) y con 0.54 (0.017) en recuerdo. Al realizar el aumento tres veces se obtuvo un F1-Score de 0.578 (sd = 0.01), 0.61 (sd = 0.005) en precisión y en recuerdo un puntaje de 0.563 (sd = 0.026). Se puede observar que al aumentar una y tres veces, el desempeño aumenta de 0.478 (sd = 0.008) a 0.578 (sd = 0.01) en F1-Score y donde se ve un aumento en la métrica de recuerdo pero una disminución de la precisión para ambos casos.

5.1.5. Comparativa mejores resultados

Como se mencionó anteriormente, en la mayoría de los casos el modelo BERT presentó el mejor rendimiento al momento de aplicar las diferentes técnicas de aumento de datos. En la Figura 26 se presenta el rendimiento del modelo con cada una de las diferentes técnicas. Se observa que al utilizar GPT-3 para aumentar el número de instancias, el modelo presenta el mejor desempeño para la subtarea 2 y para la subtarea 1 cuando consideramos aumentar 1 vez las instancias. En segundo lugar, se encuentra la técnica SMOTE para la clasificación binaria y reemplazo por sinónimo para la clasificación multiclase.

El uso de GPT-3 como técnica de aumento de datos tiene como ventaja sobre las otras técnicas de generar instancias nuevas y hasta cierto punto independientes de las originales, además, permite incorporar elementos o características nuevas fácilmente a través del prompt. Sin embargo, una de sus principales

desventajas son el hecho de que no son tan accesibles como las demás, es decir, posee un límite de créditos en cuanto al uso de su API. Otra de sus desventajas es el tiempo que se le debe dedicar al trabajar con el prompt considerando los diferentes parámetros, ya que si no se tiene cuidado puede terminar generando las mismas instancias después de cierto número de iteraciones. Las demás técnicas de aumento de datos basadas en el espacio de las instancias presentan una mayor facilidad de implementación, además de no tener limitantes económicas como al usar GPT-3. Una de sus desventajas es su dependencia de los datos originales y la cantidad de estos, es decir, si se cuentan con muy pocas instancias de una clase habrá un número muy reducido con respecto a las nuevas instancias que se quieran generar, además de estar limitadas por el vocabulario que se puede usar para seguir manteniendo el contexto. Por último, el uso de la técnica SMOTE presenta una ventaja sobre las demás en cuanto a tiempo de implementación y generación de instancias, sin embargo, se tiene poco control y conocimiento de las características que toma como base.

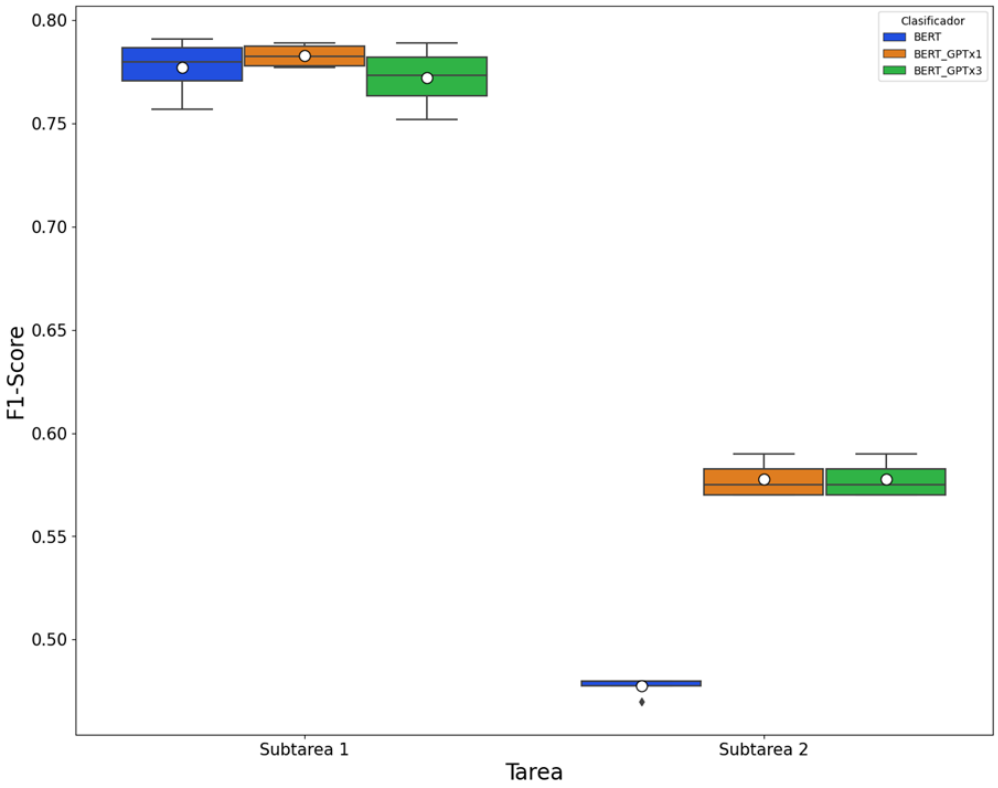


Figura 25. Resultados obtenidos al utilizar instancias generadas por GPT-3.

En la Figura 27 se puede observar el resultado de los modelos con mejor rendimiento en aprendizaje máquina y profundo, acompañados de los resultados obtenidos con solo el conjunto de datos base. En cuanto a la tarea de clasificación binaria, el modelo BERT presenta un rendimiento superior a

los mostrados con aprendizaje máquina. Por otro lado, en el caso de la tarea multiclase los modelos Naive Bayes y Random Forest logran igualar e incluso superar el rendimiento base del modelo BERT considerando el mejor caso al aplicar aumento de datos, pero al momento de compararlos con el mejor caso presentado del modelo BERT con aumento de datos su desempeño queda por debajo. Por otro lado, en la Tabla 4 se pueden observar los resultados promedio en cada una de las métricas (precisión, recuerdo y F1-Score) acompañados de su desviación estándar. La mejora en el desempeño para los casos mostrados se ven reflejados tanto en la métrica principal (F1-Score) como en el recuerdo para, traduciéndose como en una reducción de falsos negativos.

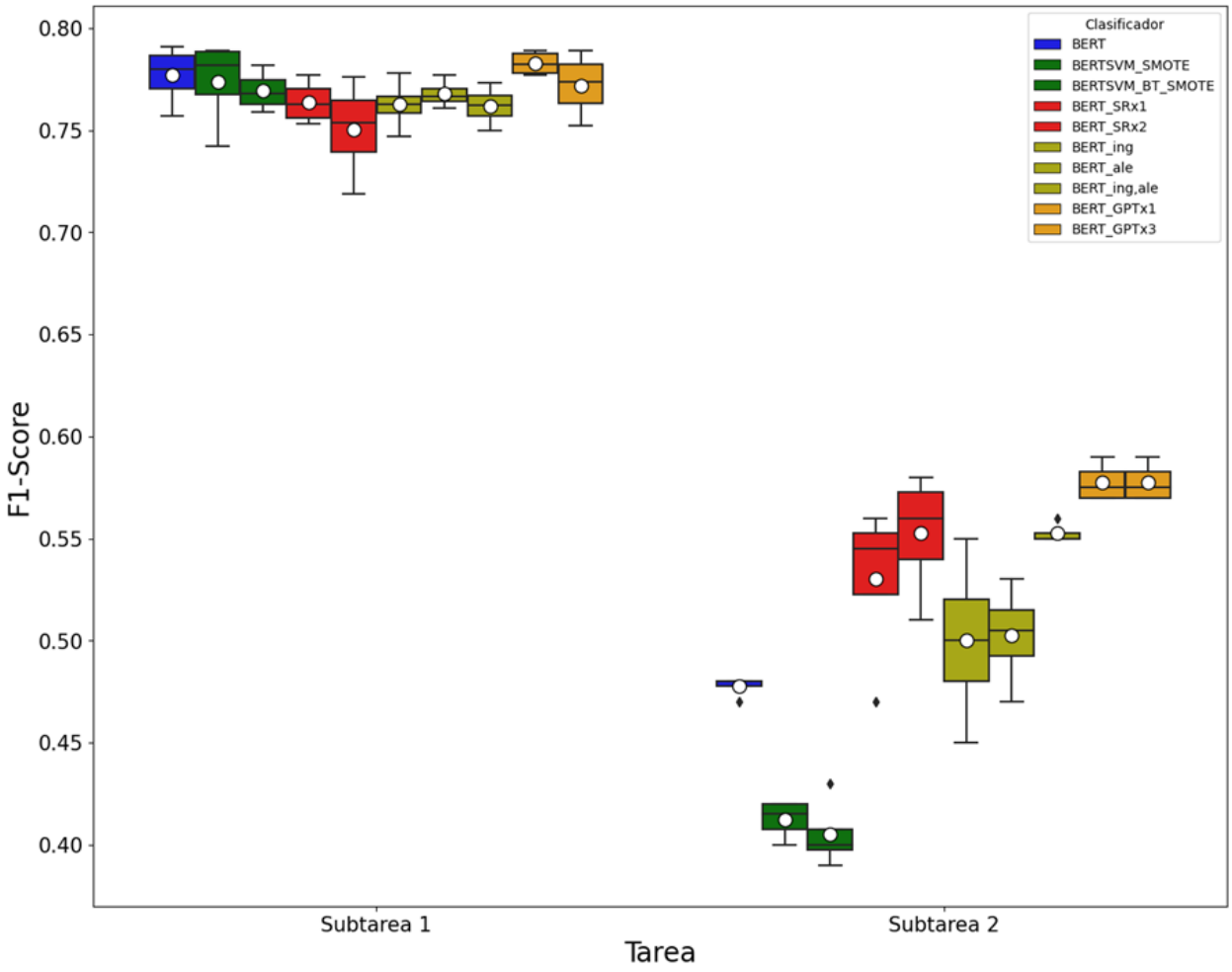


Figura 26. Resultados del modelo BERT al aplicar las diferentes técnicas de aumento de datos.

Tabla 4. Resultados experimentales utilizando el conjunto de entrenamiento DA-VINCI 2022 por validación cruzada.

(a) Resultados promedio sobre la clase positiva para la subtarea 1 evaluación local: Clasificación binaria

Modelo	Aumento	Representación	Precisión	Recuerdo	F1-Score
Naive Bayes	-	Unigramas	0.754 (0.004)	0.736 (0.009)	0.745 (0.005)
Naive Bayes	Reemplazo por sinónimo x2	Unigramas	0.714 (0.01)	0.815 (0.015)	0.761 (0.009)
SVM	-	Unigramas	0.752 (0.01)	0.724 (0.008)	0.737 (0.005)
SVM	Reemplazo por sinónimo x2	Unigramas	0.714 (0.01)	0.819 (0.015)	0.747 (0.005)
BERT	-	Vectores contextualizados	0.762 (0.0128)	0.789 (0.019)	0.777 (0.015)
BERT	GPT-3 x1	Vectores contextualizados	0.756 (0.006)	0.812 (0.014)	0.783 (0.006)

(b) Resultados promedio sobre la clase positiva, macro métricas, para la subtarea 2 evaluación local: Clasificación multiclase

Modelo	Aumento	Representación	Precisión	Recuerdo	F1-Score
Naive Bayes	-	Unigramas	0.523 (0.077)	0.42 (0.018)	0.443 (0.024)
Naive Bayes	GPT-3 x3	Unigramas	0.498 (0.03)	0.505 (0.037)	0.5 (0.032)
Random Forest	-	TF-IDF	0.526 (0.089)	0.353 (0.019)	0.376 (0.025)
Random Forest	GPT-3 x3	TF-IDF	0.479(0.026)	0.52 (0.019)	0.482 (0.022)
BERT	-	Vectores contextualizados	0.555 (0.029)	0.44 (0.008)	0.477 (0.005)
BERT	GPT-3 x1	Vectores contextualizados	0.642 (0.017)	0.537 (0.017)	0.577 (0.009)

5.1.6. Comparación con el trabajo relacionado: IberLEF DA-VINCI 2022

En la Tabla 5 y 6 se muestran los primeros 5 resultados oficiales de los demás participantes y el baseline establecido por los organizadores, además del resultado que se obtuvo con nuestro mejor modelo, BERT, con (GPT-3 x1) y sin aumento de datos. Se puede observar que el modelo BERT por sí solo obtiene un buen rendimiento obteniendo el segundo lugar con un F1-Score de 0.774, una precisión de 0.82 y 0.73 en recuerdo. Al aplicar el aumento de datos no hubo una mejora significativa obteniendo de 0.774 a 0.778 en F1-Score. En cuanto a la subtarea 2, el resultado obtenido sin aumento de datos fue de 0.4937 en F1-Score y 0.5496 con aumento, obteniendo una mejora tanto en precisión (+0.04) y recuerdo (+0.06).

Estos resultados, tanto para la subtarea 1 y 2, cumplen con el comportamiento observado durante el proceso de experimentación donde, en la subtarea 1 no se reflejó una mejora notable pero en la subtarea 2 sí.

Una clara ventaja que tiene nuestro enfoque sobre el primer (CIMAT-UG-UAM-IDIAP en la subtarea 1), segundo (VICOMTECH en la subtarea 1 y 2) y tercer lugar (ITAINNOVA en la subtarea 1 y 2) es su simplicidad en términos de pasos y procesos computacionales, es decir, los resultados se obtuvieron mediante un sólo modelo independiente e hyperparámetros fijos, mientras que los demás equipos siguieron enfoques de tipo ensamble y optimización de hyperparametros. Por otro lado, una desventaja de seguir un enfoque basado en aumento de datos es determinar la proporción de los datos final de las clases minoritarias, la cual hasta la fecha se logra mayormente de manera empírica.

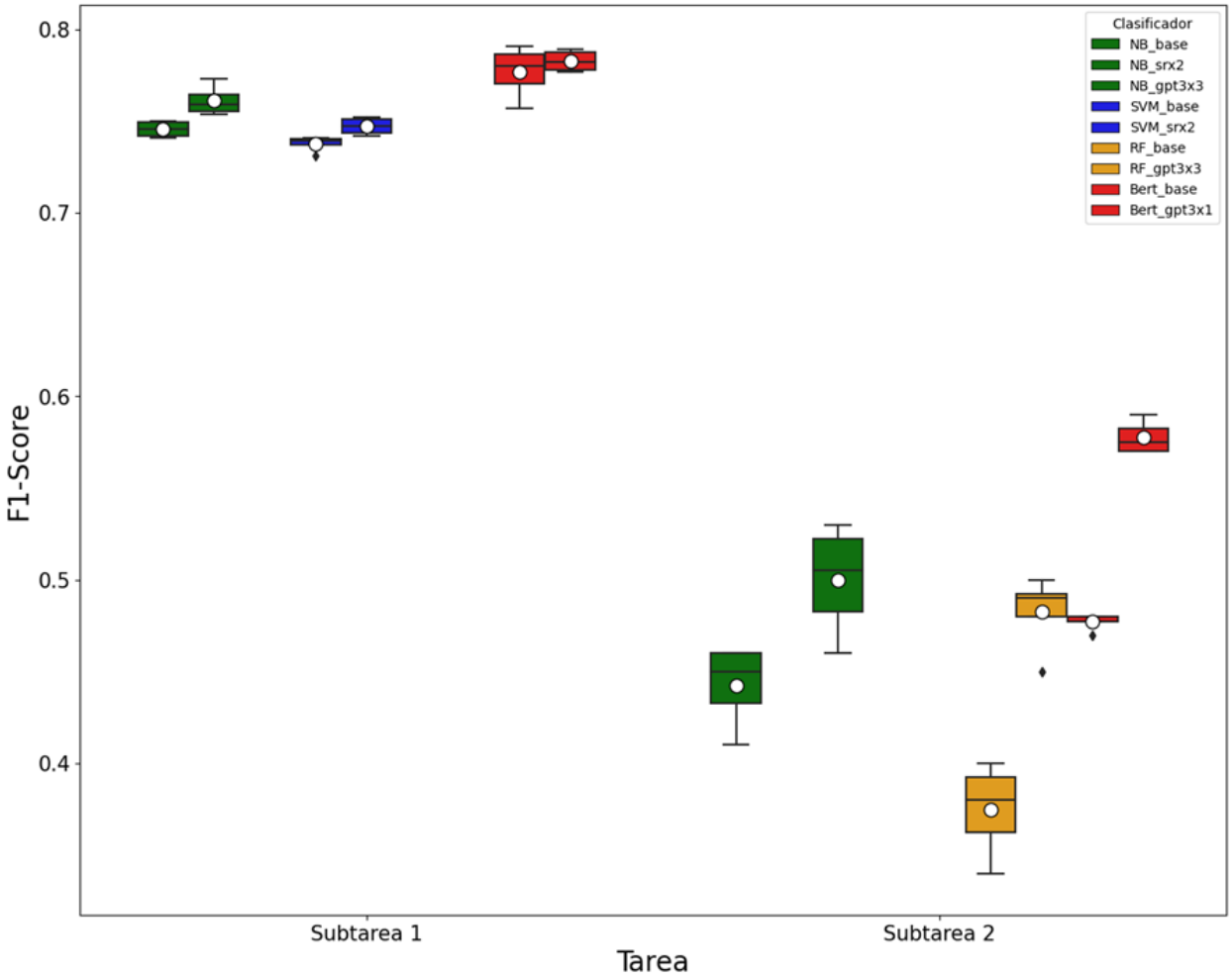


Figura 27. Mejores resultados obtenidos con Naive Bayes y Support Vector Machines con un enfoque en aprendizaje máquina y el modelo BERT con aprendizaje profundo.

5.1.7. Discusión

Desde los primeros resultados se puede observar que el modelo BERT y su arquitectura basada en Transformers supera a los modelos basados en aprendizaje máquina para esta y la mayoría de las tareas en el área de NLP. A pesar de esto, modelos como Naive Bayes y Support Vector Machines logran obtener un rendimiento bastante similar al de BERT al momento de aplicar aumento de datos, pero se debe considerar que ninguno de los modelos se les hizo una optimización de parámetros o incluyo una técnica complementaria para esta tarea como otros participantes en DA-VINCI 2022 hicieron, lo anterior debido a que se buscó abordar la problemática desde una perspectiva de los datos.

En cuanto a las técnicas de aumento de datos en nuestra evaluación cruzada, podemos observar que para la clasificación binaria los que mejores resultados se obtienen es al utilizar GPT-3 y embeddings del modelo BERT en conjunto con el modelo Support Vector Machines, aunque este último presenta una mayor dispersión en sus valores que el primero, mientras que para la clasificación multiclase se obtienen al utilizar reemplazo por sinónimo y GPT-3. Si bien, la técnica de back translation no es la peor en cada subtarea, se puede observar que tampoco es la mejor, esto puede deberse en parte a que los modelos de traducción utilizados para este tipo de aumento de datos pueden ofrecer desde resultados con grandes variaciones hasta resultados con una mínima variación, por ejemplo cambiar algún artículo en el peor de los casos. En términos de traducción es algo bueno, pero para esta tarea que buscamos variantes de las instancias originales no es muy favorable. Otro punto a considerar con esta técnica es que para mantener el número de instancias sintéticas creciendo es, en la mayoría de los casos, utilizar tantos idiomas como número de instancias nuevas se desee.

Tabla 5. Resultados oficiales utilizando el conjunto de datos 'Test' para la competencia (Arellano et al. (2022)) y el nuestro ('CICESE-DCC'), para la subtarea 1.- Clasificación Binaria.

Equipo	Precisión	Recuerdo	F1-Score
CICESE-DCC (GPT-3 x1)	**0.83**	**0.74**	**0.7804**
CIMAT-UG-UAM-IDIAP	0.803	0.750	0.775
CICESE-DCC (Sin aumento)	**0.82**	**0.73**	**0.774**
VICOMTECH	0.812	0.737	0.773
ITAINNOVA	0.779	0.751	0.765
UM-UJ-URJC	0.774	0.753	0.764
Sdamian	0.761	0.750	0.756
Baseline	0.763*	0.780*	0.750*

El efecto más notable en cuanto al uso de las técnicas puede ser observado en la subtarea 2 y esto es

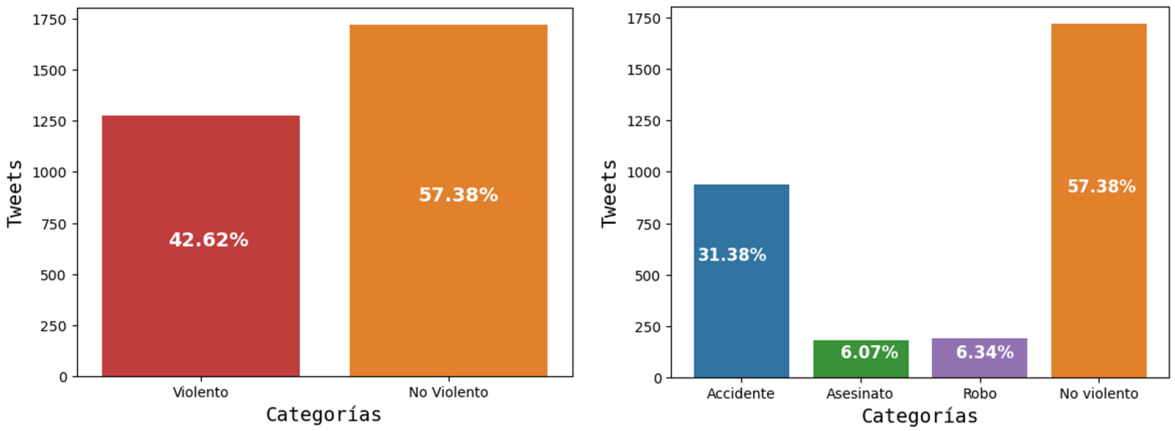
comprensible debido a que la estrategia de aumento de datos que se siguió fue aumentando las clases minoritarias en esta subtarea. Otro problema a considerar es que, se incluyeron instancias multietiqueta las cuales eran una proporción mucho menor que la clase de secuestro (siendo la de menor proporción entre las 5 clases), y esto si bien asemeja mucho la realidad al tener varios elementos en un solo tweet, sus proporciones no eran lo suficientemente grandes como para poder obtener una gran variedad de datos sintéticos de ellas. En cuanto a los modelos, se utilizaron estrategias de clasificación para poder abordar este problema y fueran consideradas durante el problema, aunque implicará problemas a los modelos de clasificación.

Tabla 6. Resultados oficiales utilizando el conjunto de datos 'Test' para la competencia (Arellano et al. (2022)) y el nuestro ('CICESE-DCC'), para la subtarea 2.- Clasificación Multiclase.

Equipo	Precisión	Recuerdo	F1-Score
GDUT	0.550	0.564	0.554
CICESE-DCC (GPT-3 x1)	**0.52**	**0.59**	**0.5496**
VICOMTECH	0.517	0.545	0.528
ITAINNOVA	0.509	0.503	0.504
CICESE-DCC (sin aumento)	**0.46**	**0.55**	**0.4937**
CIC-IPN	0.467	0.520	0.490
CIMAT-UJ-URJC	0.655	0.421	0.473
Baseline	0.498*	0.460*	0.570*

5.2. DA-VINCI 2023.- Multimodal: Texto e imágenes

Para esta edición de DA-VINCI, el conjunto de entrenamiento base consta de 2996 tweets con un total de 4267 imágenes aproximadamente. Como se puede observar en la Figura 28, existe un claro desbalance de datos entre la clase negativa (no violento/otro) y aquellas que constituyen a la clase de interés (violent, accidente, asesinato y robo), siendo la clase negativa se encuentra mayormente representada. En la subtarea dos se acentúa más este problema con la división de la categoría como violenta en 3 más, donde la categoría de asesinato y robo representan tan solo el 12.41 % del conjunto de datos. Por otro lado, en la Figura 28, para la subtarea 2 se observa un gran problema con las mismas clases mencionadas anteriormente, pero vemos que la clase secuestro se encuentra aún menos representada que estas últimas, representando tan solo el 1.37 % del total en el conjunto de datos, presentando un gran reto al momento de clasificar este tipo de instancias.



(a) Proporción de las clases para la subtarea 1.

(b) Proporción de las clases para la subtarea 2.

Figura 28. Distribución de los datos para cada subtarea DA-VINCI 2023.

Para abordar esta tarea se consideraron diferentes estrategias, las cuales se muestran en la Figura 29 y como primera intuición se buscó trabajar directamente con un enfoque 'texto-imagen' utilizando dos modelos independientes (un modelo para un tipo de información), sin embargo, resultados preliminares demostraron que este enfoque no era el apropiado. Por lo tanto, basándonos en los experimentos y resultados obtenidos en la sección 5.1 se decidió trabajar en el dominio textual. En consecuencia, se decidió utilizar el modelo BERT para procesar el texto y en el caso de las imágenes se utilizó el modelo descrito en la sección 2.2.6 al obtener descripciones de las imágenes más apropiadas que con otros modelos.

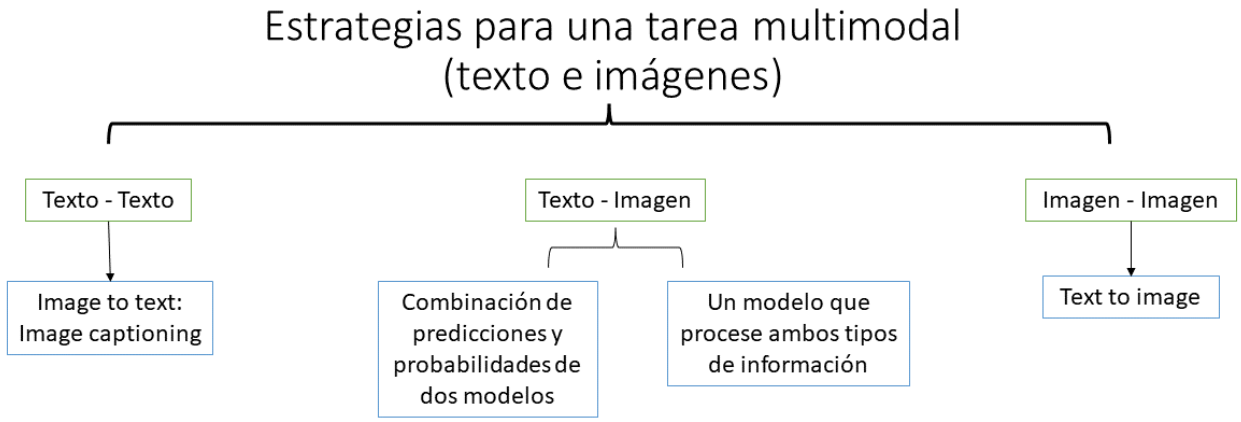
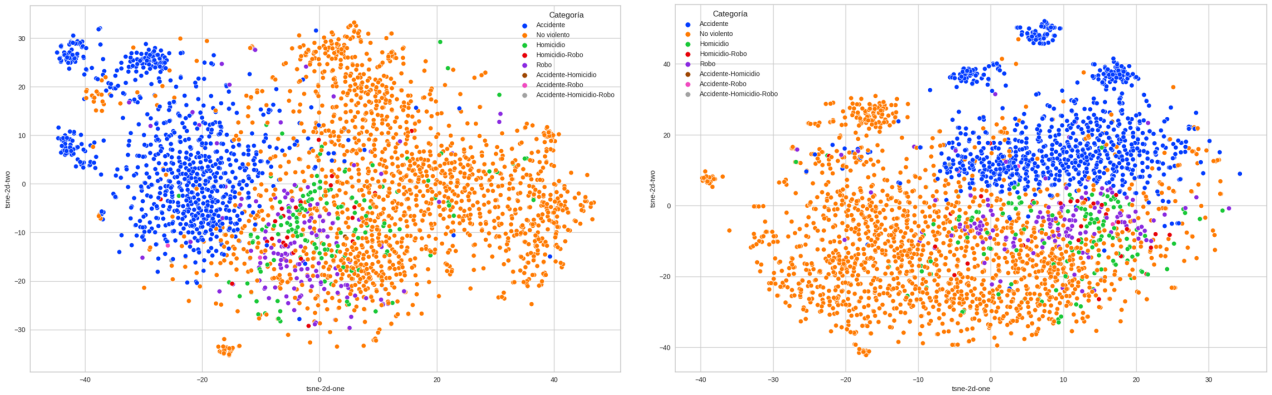


Figura 29. Estrategias consideradas para resolver la tarea propuesta en DA-VINCI 2023.

Al observar el conjunto de entrenamiento en el espacio de características por medio de la técnica TSNE (ver Figura 30) se observa un comportamiento similar que en los datos de la edición pasada, pero al considerar las descripciones de las imágenes podemos observar como algunas instancias que se mezclaban con la clase 'No violenta' se agrupan mejor con su respectiva clase. Demostrando, al menos de manera

visual, que seguir esta estrategia permite mejor agrupar las características de cada clase en una tarea multiclase.



(a) Representación visual por TSNE cuando no consideramos las descripciones de imágenes.

(b) Representación visual por TSNE cuando consideramos las descripciones de imágenes.

Figura 30. Representación visual del conjunto de datos por medio de la técnica TSNE para DA-VINCI 2023.

Los resultados utilizando el conjunto de datos base se pueden observar en la Figura 31. Se evaluó el modelo considerando solo el texto proveniente de los tweets y se comparó con el modelo considerando el texto tanto proveniente de los tweets como de las descripciones obtenidas de las imágenes. Los resultados muestran una mejora en promedio de la métrica F1-Score y recuerdo, pasando de F1-Score = 0.919 (sd =0.009) y 0.939 en recuerdo (sd =0.009) a un F1-Score de 0.923 (sd =0.0.014) y 0.944 (sd = 0.021) para la subtarea de clasificación binaria.

El mismo comportamiento pudo ser observado cuando se abordó para la tarea de clasificación multiclase, obteniendo un F1-Score de 0.868 (sd = 0.017) y 0.87 en recuerdo, considerando solo el texto de los tweets, a 0.883 (sd = 0.015) en F1-Score y 0.889 (sd = 0.017) en recuerdo.

A su vez, se decidió explorar el método de ensamble basado en cascada (sección 2.2.5.1) dónde en vez de trabajar con dos modelos independientes, uno para cada subtarea, se utilizó uno para distinguir entre instancias positivas y negativas (subtarea 1) y de las instancias clasificadas como positivas, se pasaron a otro modelo que fue previamente entrenado en las subclases de la clase positiva para distinguir entre estas, de esta manera en vez de clasificar 4 clases trabajó con 3. Los resultados con respecto a la métrica F1-Score se pueden observar en la Figura 31 donde se observa que no hay una mejora notable. Por esta razón, este enfoque se descartó para los experimentos posteriores.

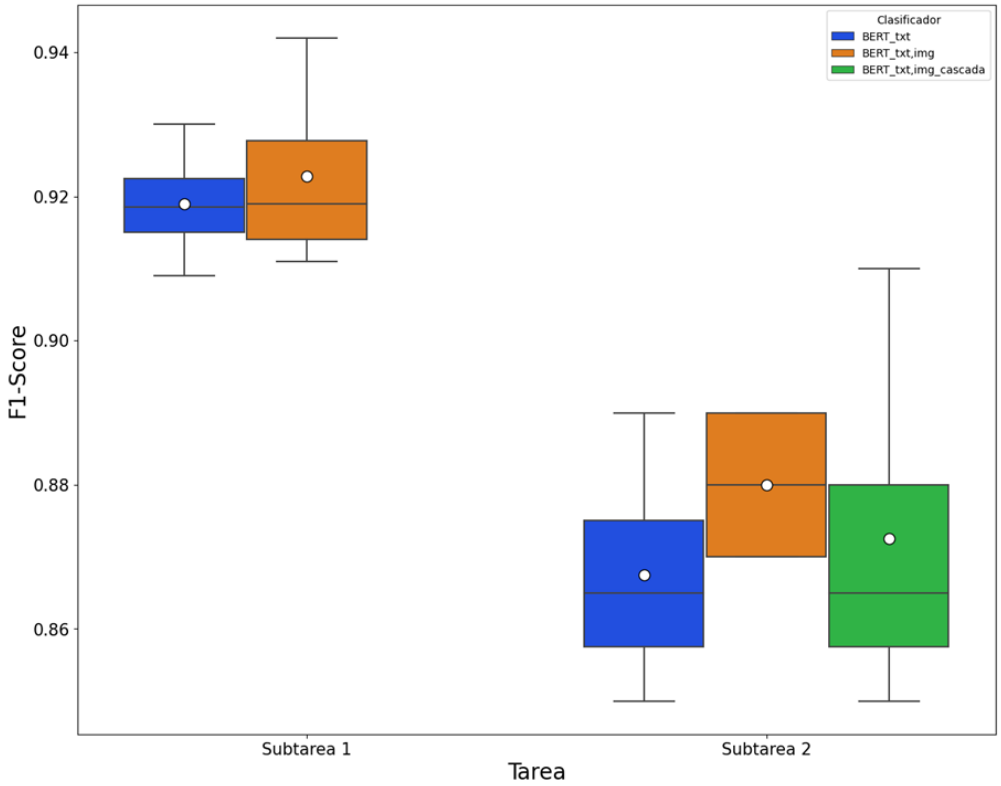


Figura 31. Boxplots que muestran los resultados en F1-Score, precisió y recuerdos obtenidos en la validación cruzada con el conjunto de datos original.

5.2.1. Aumento de datos

Como se mencionó anteriormente, el aumento de datos para el texto de los tweets fue considerando las instancias generadas por el primer prompt (prompt 1) y adicionalmente se trabajó con un segundo prompt (prompt 2), obteniendo instancias generadas por diferentes modelos de la familia GPT-3. Para poder asignarles una imagen a cada tweet, para mantener el formato de origen, se utilizó la web debido a las restricciones de tiempo en DA-VINCI 2023 y al observar que este método proporcionaba imágenes bastante relacionadas con el texto de los tweets sintéticos.

Prompt 1.

Las instancias obtenidas por el prompt 1, ver Figura 13, fueron evaluadas sin considerar las descripciones de las imágenes y después considerándolas. Los resultados pueden observarse en la Figura 32, para la subtarea 1 y considerando solo el texto proveniente de Tweets, se puede observar una mejora en cuanto a la métrica F1-Score, de 0.919 (sd = 0.009) a 0.923 (0.01), obteniendo un resultado similar a cuando

utilizamos las imágenes sin aumento de datos. Al aplicar el aumento de datos podemos observar una mejora en promedio, pero que no llega a ser considerablemente notable, pasando de 0.923 (sd =0.014) a 0.928 (sd = 0.011) en F1-Score. Sin embargo, se puede observar que el valor máximo de la validación cruzada al considerar texto, imágenes y aumento de datos es mayor a cuando solo se utiliza el texto de los tweets, pero no es mayor a cuando no se aplica aumento de datos.

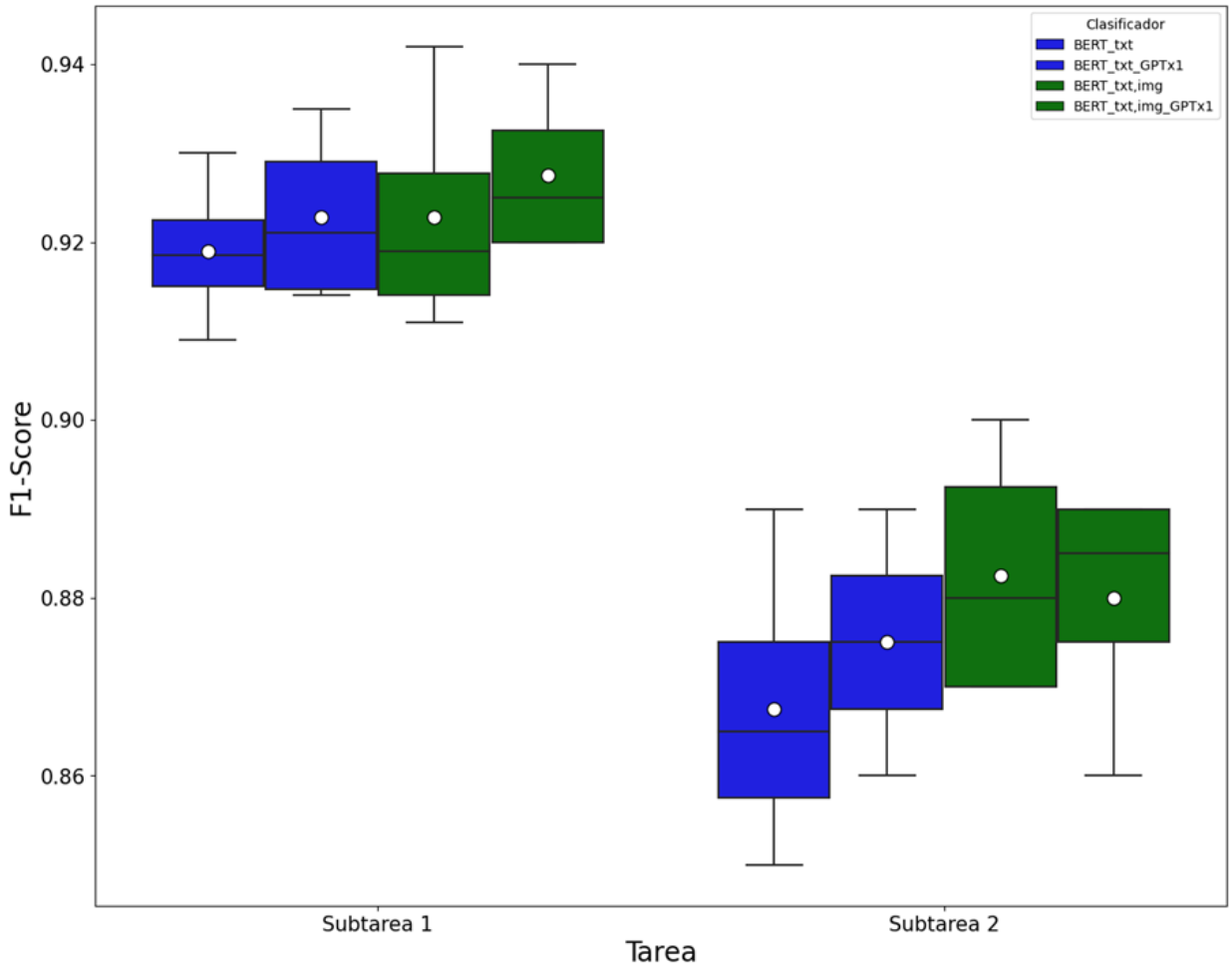


Figura 32. Resultados obtenidos con la validación cruzada considerando prompt 1 para el aumento de datos.

Para la subtarea 2, se obtiene una mejora que va de 0.868 (sd = 0.017) a 0.878 (sd = 0.013) en F1-SCORE con solo el texto de los tweets. Sin embargo, al utilizar las imágenes se tiene un rendimiento ligeramente peor en F1-Score, de 0.883 (sd = 0.015) a 0.88 (sd =0.015) con un valor mínimo de 0.86.

Prompt 2.

En la Figura 16 se puede observar con mayor detalle el último prompt (prompt 2) trabajado utilizando el modelo Turbo-3.5, los resultados fueron variables como se puede observar en la Figura 33. En promedio se

obtuvo un F1-Score de 0.921 (sd = 0.004) sin utilizar las imágenes y 0.919 (sd = 0.007) al considerarlas para la subtarea 1, lo que indica que no hubo una mejora considerable con respecto a no usar aumento de datos. En cuanto a la subtarea 2, se obtuvo un F1-Score de 0.863 (sd = 0.005) con solo el texto de los tweets y 0.878 (sd = 0.007), obteniendo una reducción en la dispersión de los resultados para ambas tareas pero obteniendo, en promedio, un rendimiento poco favorable.

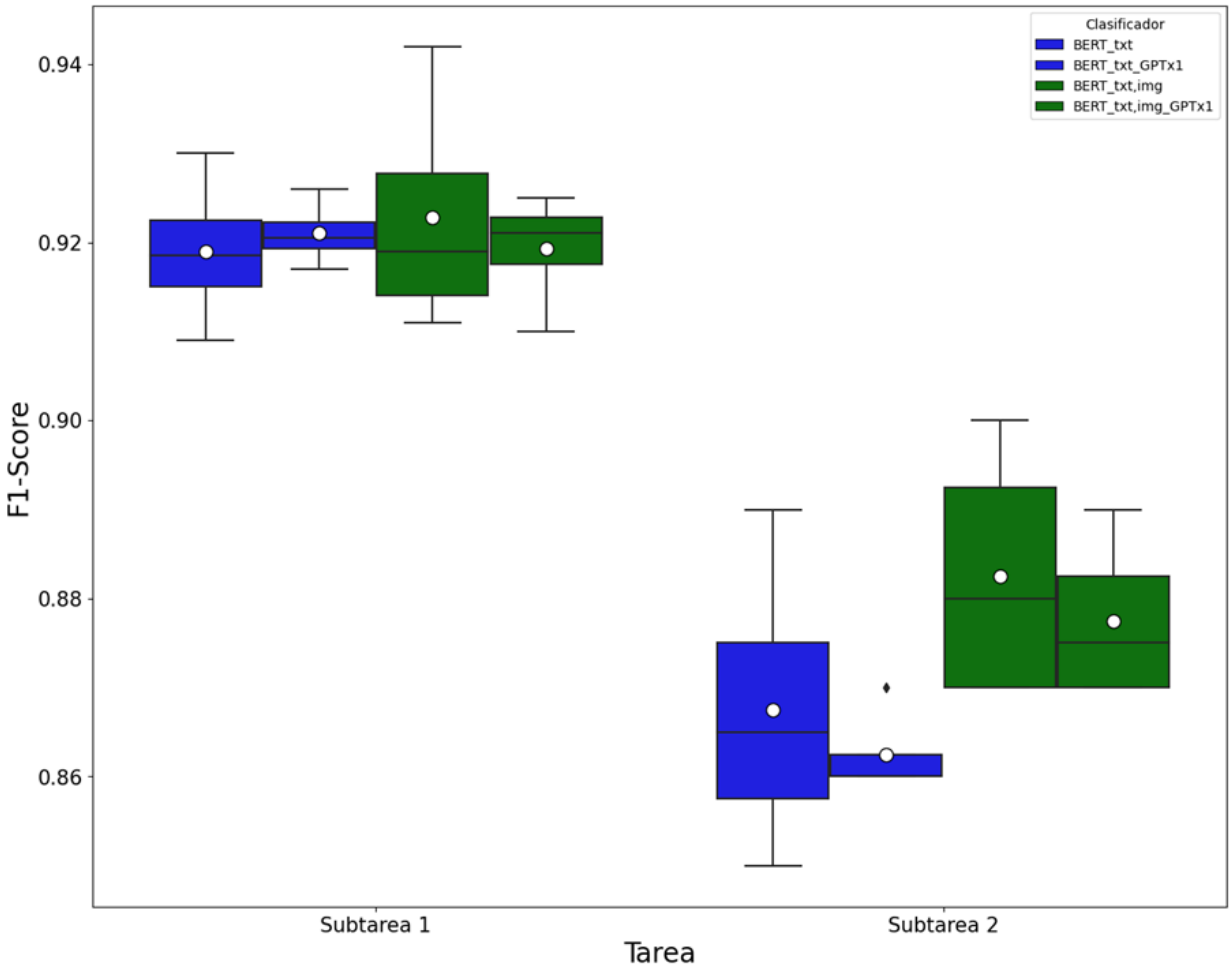


Figura 33. Resultados obtenidos con la validación cruzada considerando prompt 2 para el aumento de datos. En color azul se presentan los resultados considerando solo el texto proveniente de los tweets y en verde considerando tanto texto de los tweets como de las descripciones de las imágenes.

Comparación prompt 1 y prompt 2.

En la Figura 34 podemos observar los resultados de los modelos al considerar utilizar el aumento de datos generado por el prompt 1 y 2 para el texto de los tweets y de las imágenes obtenidas de la web. Se observa que las combinaciones que involucran aumento de dato con las instancias generadas por el prompt 1 presentan ligeras mejoras en promedio que cuanto se aplica aumento de datos. Para la subtarea

2, el prompt 1 genera mejores resultados que los obtenidos por el prompt 2 especialmente cuando solo se utiliza el texto proveniente de los tweets. Sin embargo, en ninguna de las subtareas se ve una mejora considerable, como se observó en la edición pasada cuando consideramos utilizar texto e imágenes. Esto implica una desventaja si se considera que se obtienen resultados muy cercanos cuando se utiliza y no se utiliza el aumento de datos, siendo más costoso en terminos computacionales.

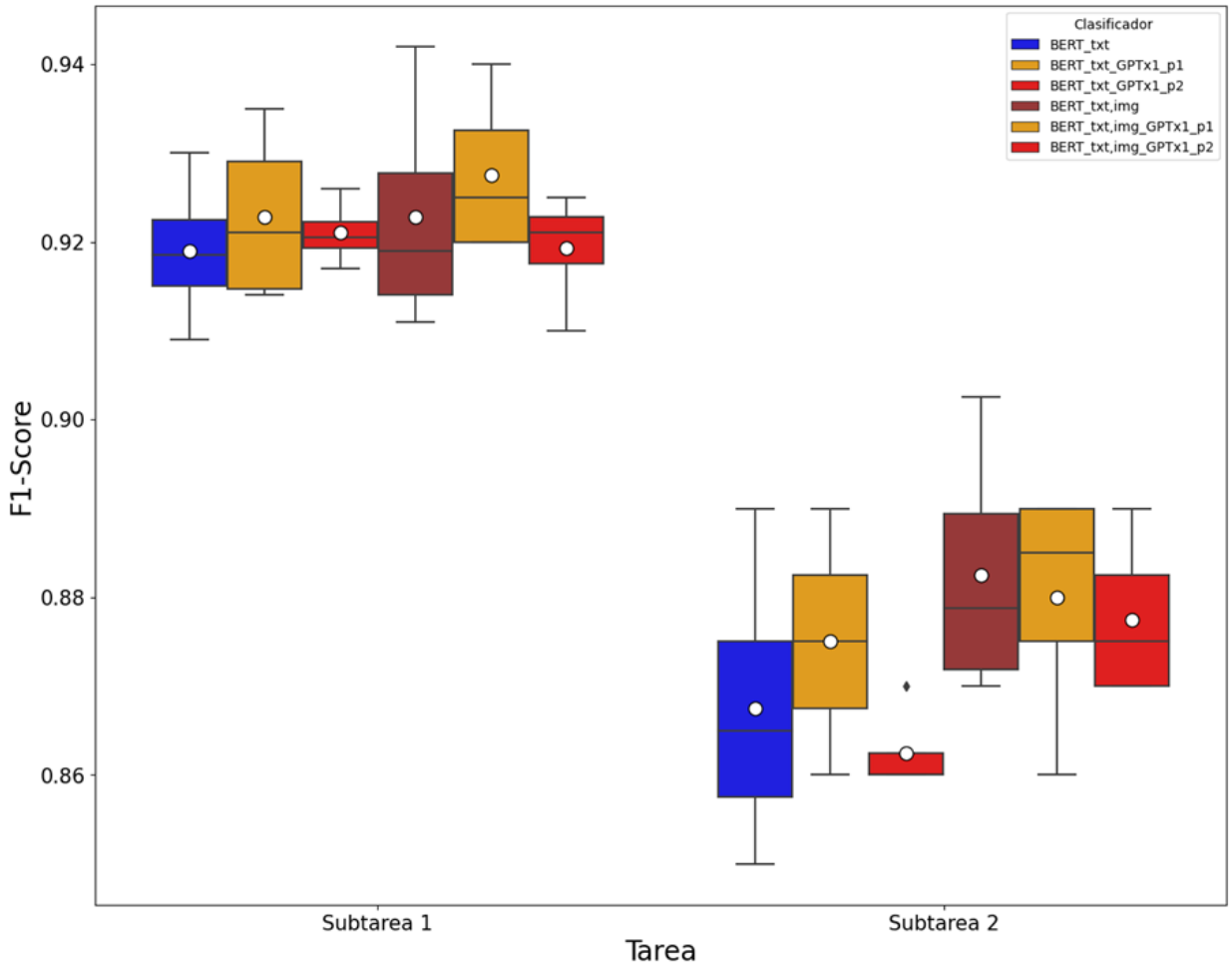


Figura 34. Resultados obtenidos con la validación cruzada considerando prompt 1 y 2 para el aumento de datos textual y recuperación de imágenes de la web. De color azul se presentan los valores base considerando solo texto de los tweets mientras que en café se presentan aquellos resultados considerando texto tanto de los tweets como de las descripciones de las imágenes. En color naranja se presentan los resultados utilizando aumento de datos generado por el prompt 1 y en rojo por el prompt 2.

5.2.2. Comparación con el trabajo relacionado: IberLEF DA-VINCI 2023

Los primeros 4 resultados en la evaluación final de la competencia DA-VINCI 2023 se muestran en la Tabla 7, donde se obtuvo, mediante el uso de texto y descripciones de imágenes, el segundo lugar y primer lugar para la subtarea 1 y 2 respectivamente sin utilizar aumento de datos. Lo primero que se puede observar es que la diferencia entre los primeros lugares, la diferencia es de unas milésimas. El resultado de nuestro modelo para la subtarea 1 fue de 0.92 en F1-Score, con una diferencia de 0.006 con el primer lugar, y con respecto a las demás métricas, una precisión y recuerdo de 0.901 y 0.941, mencionando que entre todos los participantes se obtuvo el primer lugar en la métrica de recuerdo. En cuanto a la subtarea 2, se obtuvo un F1-Score de 0.8797 y una precisión de 0.8737 con un recuerdo de 0.886 y, en este caso, obteniendo el primer lugar en precisión con respecto a los demás participantes.

Tabla 7. Resultados oficiales DA-VINCI 2023 de los participantes y el nuestro ('CICESE-DCC')

(a) Resultados para la subtarea 1: Clasificación binaria

Equipo	Precisión	Recuerdo	F1-Score
1er Mejor equipo	0.93	0.923	0.926
CICESE-DCC	**0.901**	**0.941**	**0.92**
3er Mejor Equipo	0.907	0.931	0.918
4to Mejor equipo	0.895	0.939	0.917
Baseline	0.946*	0.849*	0.895*

(b) Resultados para la subtarea 2: Clasificación multiclase

Equipo	Precisión	Recuerdo	F1-Score
CICESE-DCC	**0.8737**	**0.886**	**0.8797**
2do Mejor equipo	0.852	0.8973	0.873
3er Mejor equipo	0.862	0.878	0.869
4to Mejor equipo	0.844	0.858	0.849
Baseline	0.766*	0.941*	0.843*

Si bien nuestros experimentos locales (aquellos realizados y comparados mediante la validación cruzada) se enfocaron en utilizar un aumento de datos generados principalmente por dos prompts y dos modelos independientes (uno para cada subtarea), durante la etapa de evaluación final de la competencia se contó con un máximo de 10 envíos. Por lo tanto, se decidió enviar los resultados de las principales estrategias analizadas en secciones anteriores y además retomar otras estrategias como la clasificación en cascada y la incorporación de nuevas combinaciones. Estas estrategias y sus resultados pueden observarse en la Tabla 8, se puede observar que el segundo mejor resultado se obtuvo utilizando un aumento de datos con el prompt 1, con una diferencia en centésimas.

En cuanto a los diferentes enfoques para abordar los problemas, dentro de los primeros tres lugares en ambas tareas se puede observar el rendimiento y potencial que tiene el considerar diferentes tipos de información. Por ejemplo, el primer lugar en la subtarea 1 correspondiendo a Vallejo-Aldana et al. (2023), utiliza un método similar al nuestro al utilizar descripciones de imágenes obtenidas del mismo modelo BLIP, pero el resultado reportado es obtenido mediante un ensamble de modelos, mientras que el resultado obtenido en nuestro caso fue por medio de un solo modelo.

Mientras que Gutiérrez-Megías et al. (2023), quien obtuvo el segundo lugar en la subtarea 2, su modelo presenta ventajas en cuanto el manejo de los datos y el aumento de estos, al trabajar con dos modelos que procesen los distintos tipos de información permite mayor flexibilidad al realizar aumento de datos al no tener que estar directamente asociadas al texto del tweet como en nuestro enfoque, además de tener la posibilidad de realizar un aumento de imágenes basadas en el conjunto de datos original. Una de las ventajas de nuestro método es el uso de un solo modelo para generar las representaciones y la clasificación correspondiente al incorporar la capa densea de clasificación. Por otro lado, Rubio et al. (2023), acredor del tercer lugar para ambas subtareas, presentó un rendimiento muy cercano a los primeros lugares, utilizando solo información textual directamente proveniente de los Tweets originales. Por medio de la optimización de hiperparámetros y con un solo modelo preentrenado en el idioma español, logró obtener resultados competitivos con métodos que incorporan ambos tipos de información, algo que destaca es el proceso de limpieza que llevan a cabo a diferencia del nuestro, donde se encargan de procesar el texto de tal manera que extendieron las abreviaciones y siglas además de corregir los errores ortográficos.

Tabla 8. F1-Score para cada envío para la subtarea 1 (BIN) y 2 (MULT) en el conjunto 'Test'. P# hace referencia al aumento de datos con el prompt 1 o 2 según el número.

Estrategia	**F1-Score BIN**	**F1 Score MULT**
Tweets	0.9175	0.8575
Tweets + P1	0.9076	0.8619
Tweets + P2	0.9116	0.8536
Tweets + descripciones	**0.9203**	**0.8797**
Tweets + descripciones + P1 + descripciones	0.9187	0.8719
Tweets + descripciones + DA2 + descripciones	0.9151	0.8617
Tweets + descripciones + DA12 + descripciones	0.9127	0.8649
Tweets. Clasificación en cascada	–	0.8741
Tweets + descripciones. Clasificación en cascada	–	0.8724
Tweets + descripciones + P1 + descripciones. Clasificación en cascada	–	0.8778

5.2.3. Discusión modalidad texto e imágenes

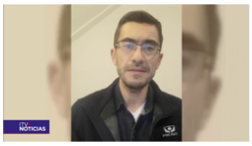
Tweet	Query	Imagen	Descripción
Homicidio ¡Terrible suceso! Se informa de un asesinato en la calle Insurgentes, Ciudad de México. Un ciudadano perdió la vida tras recibir múltiples disparos. Exigimos justicia y seguridad en nuestras calles. #JusticiaParaLasVíctimas #CDMX	asesinato Ciudad de México Insurgentes calle		Una imagen de una foto en blanco y negro de una calle con gente y caballos
Homicidio Lamentamos informar sobre un nuevo caso de homicidio en Guadalajara. Una mujer fue encontrada sin vida en su domicilio en la colonia Centro. Las autoridades se encuentran investigando el caso. #NoMásViolencia #JusticiaParaLasVíctimas	homicidio Guadalajara mujer		Una imagen de un hombre con gafas y una chaqueta negra
Robo Última hora: Intento de robo en el centro comercial local. Varios individuos intentaron ingresar a una joyería, pero fueron frustrados por la rápida respuesta de seguridad. #RoboFallido #SeguridadEfectiva	intento de robo centro comercial joyería seguridad efectiva		Una imagen de una mujer parada en un mostrador en una tienda

Figura 35. Imágenes con descripciones con poca información del evento violento.

Desde los primeros resultados se observó que el utilizar descripciones de imágenes en conjunto con el texto de los tweets permite mejorar el rendimiento de los modelos para clasificar reportes de eventos violentos. No obstante, al realizar un aumento de datos para ambas modalidades no se logró mejorar el desempeño base del modelo multimodal de manera considerable durante los experimentos con validación cruzada y esto se mantiene constante con los resultados oficiales. Teniendo los resultados más cercanos al utilizar ambos tipos de información con un método de ensamble por cascada y aumento de datos utilizando el prompt 1 pero requiriendo un mayor número de pasos para ello. Por lo tanto, ninguna estrategia logró superar, en términos de F1-Score, el resultado base. Sin embargo, al utilizar solamente el texto proveniente de los tweets se pudo observar una mejora en el rendimiento tanto en la validación cruzada como en la competencia, este último para la subtarea 2. A diferencia de la edición pasada, en este caso se contó con un tiempo más reducido, el cual no permitió experimentar con mayor detalle el aumento de datos para imágenes adecuado para asociarlo con el aumento de datos para texto. Al no presentar mejoras al utilizar aumento de datos multimodal, se le atribuye a ciertas imágenes que al momento de obtener su descripción pudo no agregar suficiente información e incluso agregar cierto ruido al modelo, algunos ejemplos de esto se pueden ver en la Figura 35. Además, se puede observar que

los resultados del prompt 2 no brinda los mejores resultados tanto en los resultados oficiales como en los obtenidos durante la experimentación, con excepción de la subtarea 1 y considerando solo texto de los tweets. Esto implica que generar nuevas instancias con variantes del mismo prompt permite obtener instancias que mejoren el rendimiento en comparación de un prompt relativamente más general para cada clase que se desee aumentar. Otro problema que se encuentra presente tanto en esta edición como la anterior, es la presencia de instancias multietiqueta las cuales para esta edición representaron el 1.13 % en el conjunto de entrenamiento. Para tratar el problema anterior, el modelo BERT se configuró para que tomara en cuenta este tipo de categorías, sin embargo, al ser muy pocas es natural que el modelo confunda este tipo de instancias, esto nos lleva a considerar que dentro de las instancias mal clasificadas se encuentran alguna de esta naturaleza.

5.2.3.1. Análisis de error

En la Figura 36 se puede observar con ayuda de la libreria *Transformer-interpret*[1] un aproximado de las predicciones hechas por el modelo y la importancia en las palabras basados en la atención proporcionada por el modelo. Las palabras resaltadas en rojo son palabras que no contribuyen a la clasificación de la categoría correspondiente, mientras que en verde sí lo son. Las etiquetas van en orden ascendente y corresponden a las clases de accidente, asesinato/homicidio, robo y no violento respectivamente.

Para la primera instancia (Figura 36a), se puede observar que a pesar de ser clasificada como accidente, las palabras 'muere' y 'accidente' proporcionan la mayor relevancia en todo el texto, sin embargo, palabras fuera de este contexto como 'rolling stones' y palabras incorporadas por las descripciones como 'una imagen' y 'brazos' llevan al modelo a clasificar esta instancia como no violento (LABEL_3). Por otro lado, en la segunda instancia (Figura 36b) se observa que se repite de manera similar el mismo error, el modelo centra su atención en palabras genéricas como 'hagamos', 'género' y las palabras que originalmente estaban acompañadas con hashtag (#). En estos casos, la incorporación de las descripciones de imágenes terminó incluyendo mayor ruido que información.

[1] https://github.com/cdpierse/transformers-interpret

Texto: "#VicioNoticias Muere el manager de giras de los Rolling Stones, Mick Brigden, en un accidente doméstico https://t.co/6MhZkh280i https://t.co/NsnBMvbiPA"

Clase: Accidente/LABEL_0

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
n/a	(0.07)	LABEL_0	-2.98	[CLS] vici ##ono ##ticia ##s muere el man ##ager de giras de los roll ##ing st ##ones mic ##k bri ##g ##den en un accidente doméstico [SEP] una imagen de un hombre en una bicicleta saluda ##n ##do con los brazos en el aire [SEP]
n/a	(0.01)	LABEL_1	-4.49	[CLS] vici ##ono ##ticia ##s muere el man ##ager de giras de los roll ##ing st ##ones mic ##k bri ##g ##den en un accidente doméstico [SEP] una imagen de un hombre en una bicicleta saluda ##n ##do con los brazos en el aire [SEP]
n/a	(0.00)	LABEL_2	-3.83	[CLS] vici ##ono ##ticia ##s muere el man ##ager de giras de los roll ##ing st ##ones mic ##k bri ##g ##den en un accidente doméstico [SEP] una imagen de un hombre en una bicicleta saluda ##n ##do con los brazos en el aire [SEP]
n/a	(0.88)	LABEL_3	2.57	[CLS] vici ##ono ##ticia ##s muere el man ##ager de giras de los roll ##ing st ##ones mic ##k bri ##g ##den en un accidente doméstico [SEP] una imagen de un hombre en una bicicleta saluda ##n ##do con los brazos en el aire [SEP]

(a) Texto original y su correspondiente clase seguidos de la importancia de cada palabra para las correspondientes clases.

Texto: "Pertenezco a un grupo de Gossip Girl en Facebook, hoy nos comentaron que una de las integrantes del grupo fue asesinada en su propia casa. Un feminicidio terrible. Hagamos visible la violencia de género y que nuestra voz sea escuchada #JusticiaParaDiana #NiUnaMas #NiUnaMenos https://t.co/uMusLhqizn"

Clase: Asesinato/LABEL_1

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
n/a	(0.01)	LABEL_0	-7.75	[CLS] perten ##ezco a un grupo de go ##s ##si ##p gir ##l en fac ##eb ##oo ##k hoy nos comentar ##on que una de las integrantes del grupo fue asesinada en su propia casa un femin ##ici ##dio terrible hagamos visible la violencia de género y que nuestra voz sea escucha ##da justicia ##para ##dia ##na ni ##una ##mas ni ##una ##men ##os [SEP] una imagen de un fondo rosa con las palabras justi ##ce parad ##ia una imagen de un mensaje de texto en un teléfono celular con una imagen de un hombre [SEP]
n/a	(0.03)	LABEL_1	-0.69	[CLS] perten ##ezco a un grupo de go ##s ##si ##p gir ##l en fac ##eb ##oo ##k hoy nos comentar ##on que una de las integrantes del grupo fue asesinada en su propia casa un femin ##ici ##dio terrible hagamos visible la violencia de género y que nuestra voz sea escucha ##da justicia ##para ##dia ##na ni ##una ##mas ni ##una ##men ##os [SEP] una imagen de un fondo rosa con las palabras justi ##ce parad ##ia una imagen de un mensaje de texto en un teléfono celular con una imagen de un hombre [SEP]
n/a	(0.01)	LABEL_2	-5.17	[CLS] perten ##ezco a un grupo de go ##s ##si ##p gir ##l en fac ##eb ##oo ##k hoy nos comentar ##on que una de las integrantes del grupo fue asesinada en su propia casa un femin ##ici ##dio terrible hagamos visible la violencia de género y que nuestra voz sea escucha ##da justicia ##para ##dia ##na ni ##una ##mas ni ##una ##men ##os [SEP] una imagen de un fondo rosa con las palabras justi ##ce parad ##ia una imagen de un mensaje de texto en un teléfono celular con una imagen de un hombre [SEP]
n/a	(0.97)	LABEL_3	5.68	[CLS] perten ##ezco a un grupo de go ##s ##si ##p gir ##l en fac ##eb ##oo ##k hoy nos comentar ##on que una de las integrantes del grupo fue asesinada en su propia casa un femin ##ici ##dio terrible hagamos visible la violencia de género y que nuestra voz sea escucha ##da justicia ##para ##dia ##na ni ##una ##mas ni ##una ##men ##os [SEP] una imagen de un fondo rosa con las palabras justi ##ce parad ##ia una imagen de un mensaje de texto en un teléfono celular con una imagen de un hombre [SEP]

(b) Texto original y su correspondiente clase seguidos de la importancia de cada palabra para las correspondientes clases.

Figura 36. Ejemplos de instancias mal clasificadas con el modelo basado en texto y descripción de imágenes.

Capítulo 6. Conclusiones y trabajo a futuro

La detección de eventos violentos a partir de modelos basados en inteligencia artificial puede implicar un gran reto cuando no se cuenta con la suficiente información que le permitan generalizar de manera adecuada. Cuando ocurre lo anterior, existen dos principales caminos los cuales se pueden seguir, estos consisten en realizar una optimización de los parámetros para que los modelos puedan adaptarse mejor a la temática y tipo de tarea, y la otra es proporcionar una mayor cantidad de datos. En este trabajo de investigación se decidió dar un énfasis en la cantidad de los datos, ya que los modelos de inferencia parten de estos para cualquier futura actividad, es decir, requieren datos para aprender a resolver una tarea e incluso, para realizar una optimización de parámetros, es necesario tener una cantidad de datos base que permitan un rendimiento mínimamente bueno para poder realizarlo y tener los efectos deseados. El problema surge cuando no se cuenta con los recursos suficientes (por ejemplo: personal, tiempo o la fuente para obtener la información) para realizar una recuperación adicional de datos y en estos casos, la generación de datos sintéticos resulta en una alternativa adecuada. Existen una gran variedad de técnicas como se mencionó en la sección 2.2.3, pero no significa que cualquiera sea efectiva para todo tipo de tareas, ya que no todas las técnicas llegan a preservar la naturaleza de la clase asignada. Cabe mencionar que, el uso de modelos preentrenados ha permitido incluir modelos de aprendizaje profundo para que puedan desempeñarse con conjuntos de datos no tan grandes y que, a su vez, estas técnicas de aumento para texto puedan aportar más con menor procesamiento.

El presente trabajo ha pretendido diseñar modelos que permitan tener un buen desempeño utilizando un conjunto de datos base extendido de manera sintética, al detectar eventos violentos dentro de un rango de categorías y poder distinguir entre ellas en el idioma español, esto último debido a que, como se mencionó anteriormente, la mayoría de los trabajos relacionados en este tema están enfocados en el idioma inglés. Dentro de los modelos trabajados, se mostró como es que los modelos de aprendizaje máquina pueden presentar un rendimiento aceptable, los modelos de aprendizaje profundo basados en la arquitectura Transformers BERT presentaron un mejor rendimiento general.

El objetivo general de esta investigación fue la de diseñar un método que permita la clasificación de publicaciones de eventos violentos en Twitter, utilizando técnicas de aumento de datos que mejoren el rendimiento a partir del uso de información multimodal. Por una parte, los métodos que involucraron solo información unimodal con ayuda del aumento de datos permitieron posicionarnos por encima de otras estrategias que involucraron técnicas de ensamble o aprendizaje multitarea tan solo con aumentar la cantidad de datos de entrenamiento, siendo una de las ventajas el permitir a los modelos obtener mayor generalización, solventar el inconveniente de escasez de datos públicos o su recolección y etiquetado,

ya que esto último puede afectar al modelo si la recolección de los nuevos datos y etiquetado no es la adecuada agregando ruido y la disminución del rendimiento del mismo modelo.

En este trabajo se exploraron 4 técnicas de aumento de datos como técnica complementaria para resolver la tarea de clasificación de eventos violentos en Twitter, de las cuales 3 se encuentran a nivel instancia y 1 en el espacio de características. Cada una de las técnicas cuenta con sus ventajas y limitaciones, por ejemplo, la técnica SMOTE presentó un buen desempeño en la clasificación binaria, donde él traslape de las clases es menor a comparación de cuando se aborda la tarea de clasificación multiclase donde presentó su mayor limitación. Por otro lado, las técnicas a nivel instancia como reemplazo por sinónimo y *back translation* permitieron un buen rendimiento de la mayoría de los modelos, principalmente en reemplazo por sinónimo, pero estos alcanzan su mejor rendimiento al aumentar las instancias en proporciones relativamente bajas (de 1 a 2 veces más el número de instancias originales). La ventaja de *back translation* es que permite en el mejor de los casos aplicar indirectamente otras técnicas a nivel palabra como el intercambio de posición de algunas palabras y aun así mantener el significado original de la instancia, pero en el peor de los casos cuando buscamos variedad en las instancias es cuando cambia un artículo por otro, los mejores resultados se obtuvieron para la subtarea 2 utilizando las instancias generadas por dos idiomas. La técnica de Reemplazo por sinónimo ofreció una mejora considerable, especialmente para la subtarea 2, siendo limitada por la cantidad de sinónimos de las palabras, ofrece una ventaja por su relativa simplicidad y, depende como se vea, el que no considere *stopwords*. La última técnica de aumento de datos para texto explorada fue el uso de modelos GPT-3 para generar nuevas instancias basadas en palabras claves extraídas del conjunto de datos original, ofreciendo el mejor desempeño general entre todas las técnicas al utilizar un *prompt* dinámico; sin embargo, al igual que las demás técnicas, el aumentar las instancias minoritarias en proporciones muy grandes no necesariamente implica una mejora garantizada en el rendimiento del modelo, incluso para GPT-3. Por lo tanto, la selección de las técnicas y la proporción a aumentar es un proceso iterativo y empírico para cada tarea.

La manera en que el ser humano adquiere información del medio es de manera multimodal a través de los sentidos, esto nos facilita tomar decisiones acordes a lo que se observa, sin embargo, en algunos casos la información captada por un sentido puede aportar más información que los demás. El caso anterior ocurre de igual manera en el campo de la inteligencia artificial, ya que al capturar información desde múltiples fuentes permite obtener una representación más rica y contextualizada de los datos, al capturar características y patrones que podrían perderse si se abordara cada modalidad por separado. En el caso que se presenta en este trabajo, combinar texto y el contexto visual de las imágenes puede mejorar la interpretación de la semántica. Como se pudo observar con el estado del arte, existen varias

estrategias para procesar la información multimodal y, con base en los resultados expuestos, el utilizar la información proveniente tanto del texto como las imágenes permitieron obtener los mejores rendimientos en ambas subtareas. Tratar la información de las imágenes por medio de la generación de descripciones automáticas, junto con la información textual de los tweets, permitió mejorar el rendimiento del modelo en comparación con solo utilizar información del texto o el texto y las imágenes por separado. Esto se puede deber a que permite destacar los elementos principales de las imágenes (como policías, tiendas y armas, etc.) a diferencia de si se extrae una representación de toda la imagen para que el modelo trate con ello. Esta estrategia (utilizando descripciones de imágenes), según resultados previos, parece favorecer al recuerdo o reducir los falsos negativos, el cual, dependiendo de la aplicación, puede ser favorable al permitir al modelo considerar la mayor cantidad de reportes reales y no pasar por alto un evento que requiera la atención de las autoridades, esto último pensando en aplicaciones del modelo.

Dado lo anterior, se pueden contestar las preguntas de investigación de la siguiente manera: (i) La técnica de aumento de datos que ha resultado ser la más efectiva en tareas de clasificación de eventos violentos con información multimodal ha sido utilizando modelos de lenguaje de gran tamaño para aumentar el texto para un problema multiclase en combinación con modelos de aprendizaje profundo basados en Transformer. En cuanto a imágenes, la recuperación de nuevas imágenes por web permitió obtener una ligera mejora en promedio a cuando no se utiliza. (ii) Utilizando los modelos de lenguaje grande para aumentar el número de instancias permitió una mejora considerable, particularmente para la clasificación multiclase. Sin embargo, presentó mejores rendimientos en los diferentes modelos cuando se aumentaron 1 y 3 veces el número de instancias minoritarias, un mayor aumento a este no presentó mejoras considerables.

6.1. Limitaciones

Durante el trabajo de investigación, una de las mayores limitantes fue el de no poder tratar las instancias multietiquetas por medio del aumento de datos, debido a que algunas de estas categorías contenían muy pocas como para generar una variedad de instancias considerable, por ejemplo en una categoría se presentó 1 sola instancia. Si bien el porcentaje de estas es muy bajo en comparación con las demás clases, no se puede negar que este problema de multietiqueta asemeja a como algunos problemas se presentan en la realidad. Este trabajo revelan que algunas clases contaban con muy pocas instancias como para realizar aumento de datos. Otro de los limitantes que se presentaron fue el tiempo que se

tuvo de experimentación entre la modalidad textual y visual. Esto último debido a que los experimentos realizados para la modalidad textual, la competencia DA-VINCI 2022, ya había finalizado, mientras que para DA-VINCI 2023 se contó con un tiempo más reducido para explorar y aplicar técnicas de aumento de datos para imágenes fuera de una imagen base. También se debe considerar que los modelos propuestos no son capaces de clasificar otras categorías de violencia fuera de las consideradas.

6.2. Trabajo a futuro

Los experimentos y resultados durante la competencia demostraron tener uno de los mejores rendimientos. Sin embargo, todavía existe la posibilidad de mejorarlo a través de la optimización de hiperparámetros, ya que mantuvieron los mismos durante todos los experimentos.

Realizar un estudio en profundidad sobre las características utilizadas para poder trabajar con modelos que sean explicativos y poder partir de nuevas bases para mejorar la calidad de las instancias con que se alimenta a los modelos.

Explorar diferentes métodos que permitan evaluar la calidad de las instancias sintéticas generadas por modelos de lenguaje grande, además de explorar otro tipo de modelos de lenguaje grande, por ejemplo, se podrían probar los *prompts* con versiones recientes de estos modelos como GPT-4 o incluso BART de Google para el aumento de datos de texto, así como mejorarlos. Para la información visual, experimentar con otras técnicas para recuperar información relevante de los tweets y mejorar la cadena de búsqueda que se utilizará, ya sea para generar imágenes con texto para modelos visuales como Midjourney o Dall-E, o incluso para explorar más el potencial de la web incorporando otras estrategias como más imágenes o filtros para reducir la cantidad de imágenes ruidosas recuperadas. En términos de modelos, explorar con mayor profundidad modelos de visión como Vision Transformer u otro modelo para procesar información visual para ver las capacidades y limitaciones al considerar dos tipos diferentes de datos al mismo tiempo y compararlos cuando solo se trabaja en un dominio.

Literatura citada

Aas, K. & Eikvil, L. (1999). Text categorisation: A survey. [Archivo PDF] https://www.cis.uni-muenchen.de/kurse/pmaier/ML_05/material/aas99text.pdf.

Abbass, Z., Ali, Z., Ali, M., Akbar, B., & Saleem, A. (2020). A framework to predict social crime through twitter tweets by using machine learning. *Proceedings - 14th IEEE International Conference on Semantic Computing, ICSC 2020*, 363–368. https://doi.org/10.1109/ICSC.2020.00073.

Aiken, M. & Park, M. (2010). The efficacy of round-trip translation for mt evaluation. [Hoja informativa en línea] Consultado el 15 de Septiembre 2022 en http://www.translationjournal.net/journal/51reverse.htm.

Alara, D. & Sayak, P. (2023). A dive into vision-language models. [Hoja informativa en línea] Consultado el 10 de Mayo 2023 en https://huggingface.co/blog/vision_language_pretraining.

Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, 164(February 2020), 114006. https://doi.org/10.1016/j.eswa.2020.114006.

Arellano, L. J., Escalante, H. J., Villaseñor-Pineda, L., Montes-Y-Gómez, M., & Sanchez-Vega, F. (2022). Overview of DA-VINCIS at IberLEF 2022: Detection of aggressive and violent incidents from social media in spanish. *Procesamiento del Lenguaje Natural*, 69(2), 207–215. https://doi.org/10.26342/2022-69-18.

Arriaga, O., Plöger, P., & Valdenegro-Toro, M. (2017). Image captioning and classification of dangerous situations. https://doi.org/https://doi.org/10.48550/arXiv.1711.02578.

Bayer, M., Kaufhold, M. A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7). https://doi.org/10.1145/3544558.

Breiman, L. (2001). Random Forest. *Machine Learning*, 5–32. https://doi.org/10.1109/ICCECE51280.2021.9342376.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR 2020*. https://doi.org/https://doi.org/10.48550/arXiv.2308.02976.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953.

Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Li, Q., Shen, D., Liu, T., & Li, X. (2023). ChatAug: Leveraging ChatGPT for text data augmentation. 1–12. https://doi.org/https://doi.org/10.48550/arXiv.2302.13007.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm), 4171–4186. https://doi.org/https://doi.org/10.48550/arXiv.1810.04805.

Fatichah, C., Sammy Wiyadi, P. D., Adni Navastara, D., Suciati, N., & Munif, A. (2020). Incident detection based on multimodal data from social media using deep learning methods. *7th International Conference on ICT for Smart Society: AIoT for Smart Society, ICISS 2020 - Proceeding*. https://doi.org/10.1109/ICISS50791.2020.9307555.

García-díaz, J. A., Jiménez-zafra, S. M., Rodríguez-garcía, M. Á., & Valencia-garcía, R. (2022). UMUTeam at DA-VINCIS 2022 : Aggressive and violent classification using knowledge integration and ensemble learning. *IberLEF 2022. La Coruña, Spain*, Vol-3202, 0–7. [Archivo PDF] https://ceur-ws.org/Vol-3202/davincis-paper8.pdf.

Gutiérrez-Megías, A. J., Stoia, S., Martínez-Santiago, F., Ureña-López, L. A., & Montejo-Ráez, A. (2023). SINAI Participation at DA-VINCIS task in IberLEF 2023: Data augmentation for multimodal classification. *CEUR Workshop Proceedings*, 71. Work submitted for publication.

Hernández-Minutti, B., Olivares-Padilla, J.-A., Valerio-Carrera, R., & Gambino, O. J. (2023). Detection of violent events in social media: DA-VINCIS 2023⋆. *CEUR Workshop Proceedings*, 71. Work submitted for publication.

Jiang, R., Chen, D., & Minn, D. (2017). Managing a Portfolio with Sentiment Analysis. [Archivo PDF] https://onedrive.live.com/?authkey=%21ACSly20TAD40ZFQ&id=D9E0D3D4DBC70585%2156648&cid=D9E0D3D4DBC70585&parId=root&parQt=sharedby&o=OneUp.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Nédellec, C. & Rouveirol, C., editors, *Machine Learning: ECML-98*, 137–142, Berlin, Heidelberg. Springer Berlin Heidelberg. https://doi.org/https://doi.org/10.1007/BFb00266.

Kaplan, A. M. & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information (Switzerland)*, 10(4), 1–68. https://doi.org/10.3390/info10040150.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, 591–600. https://doi.org/10.1145/1772690.1772751.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. https://doi.org/https://doi.org/10.48550/arXiv.2201.12086.

Ma, H., Huang, W., Jing, Y., Yang, C., Han, L., Dong, Y., Ye, H., Shi, Y., Zheng, Q., Liu, L., & Ruan, C. (2019). Integrating growth and environmental parameters to discriminate powdery mildew and aphid of winter wheat using bi-temporal landsat-8 imagery. *Remote Sensing*, 11, 846. https://doi.org/10.3390/rs11070846.

Marius-Constantin, P., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7), 579–588. https://www.researchgate.net/publication/228340819_Multilayer_perceptron_and_neural_networks.

Marivate, V. & Moiloa, P. (2017). Catching crime: Detection of public safety incidents using social media. *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference, PRASA-RobMech 2016*. https://doi.org/10.1109/RoboMech.2016.7813140.

Marivate, V. N. (2015). Extracting South African safety and security incident patterns from social media. *Proceedings of the 2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference, PRASA-RobMech 2015*, 106–111. https://doi.org/10.1109/RoboMech.2015.7359507.

Mata, F., Torres-Ruiz, M., Guzman, G., Quintero, R., Zagal-Flores, R., Moreno-Ibarra, M., & Loza, E. (2016). A mobile information system based on crowd-sensed and official crime data for finding safe routes: A case study of Mexico City. *Mobile Information Systems*, 2016. https://doi.org/10.1155/2016/8068209.

Montañés-salas, R. M. & Peña-larena, P. (2022). ITAINNOVA @ DA-VINCIS : A Tale of Transformers and simple optimization techniques. *IberLEF 2022. La Coruña, Spain*, Vol-3202, 7–8. [Archivo PDF] https://ceur-ws.org/Vol-3202/davincis-paper5.pdf.

Mumuni, A. & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, 16(November), 100258. https://doi.org/10.1016/j.array.2022.100258.

Naseem, U., Razzak, I., & Eklund, P. W. (2021). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, 80(28-29), 35239–35266. https://doi.org/10.1007/s11042-020-10082-6.

Piña-García, C. A. & Ramírez-Ramírez, L. (2019). Exploring crime patterns in Mexico City. *Journal of Big Data*, 6. https://doi.org/10.1186/s40537-019-0228-x.

Prieto Curiel, R., Cresci, S., Muntean, C. I., & Bishop, S. R. (2020). Crime and its fear in social media. *Palgrave Communications*, 6, 1–12. https://doi.org/10.1057/s41599-020-0430-7.

Qin, G., He, J., Bai, Q., Lin, N., Wang, J., Zhou, K., Zhou, D., & Yang, A. (2022). Prompt based framework for violent event recognition in spanish. *CEUR Workshop Proceedings*, 3202. [Archivo PDF] https://ceur-ws.org/Vol-3202/davincis-paper6.pdf.

Re, M. & Valentini, G. (2012). *Ensemble methods: A review*, (pp. 563–594). [Archivo PDF] https://www.researchgate.net/publication/230867318_Ensemble_methods_A_review.

Rodríguez-Bribiesca, I., López-Monroy, A. P., & Montes-Y-Gómez, M. (2021). Multimodal weighted fusion of Transformers for movie genre classification. *Multimodal Artificial Intelligence, MAI Workshop 2021 - Proceedings of the 3rd Workshop*, 1–5. https://doi.org/10.18653/v1/2021.maiworkshop-1.1.

Rubio, J. L. S., Almeida, A. V., & Segura-Bedmar, I. (2023). UC3M at Da-Vincis-2023: using BETO for detection of aggressive and violent incidents on social networks. *CEUR Workshop Proceedings*, 71. Work submitted for publication.

Sabharwal, A. & Selman, B. (2011). S. russell, p. norvig, artificial intelligence: A modern approach, third edition. *Artif. Intell.*, 175, 935–937. https://doi.org/10.1016/j.artint.2011.01.005.

Sandagiri, S. P., Kumara, B. T., & Kuhaneswaran, B. (2020a). Deep neural network-based approach to identify the crime related Twitter posts. *2020 International Conference on Decision Aid Sciences and Application, DASA 2020*, 1000–1004. https://doi.org/10.1109/DASA51403.2020.9317098.

Sandagiri, S. P., Kumara, B. T., & Kuhaneswaran, B. (2020b). Detecting crimes related Twitter posts using SVM based two stages filtering. *International Conference on Industrial and Information Systems*, (978), 506–510. https://doi.org/https://doi.org/10.1109/ICIIS51140.2020.9342698.

Ta, H. T., Bakar, A., Rahman, S., Najjar, L., & Gelbukh, A. (2022a). GAN-BERT : Adversarial learning for detection of aggressive and violent incidents from social media. *IberLEF 2022. La Coruña, Spain*, Vol-3202. [Archivo PDF] https://ceur-ws.org/Vol-3202/davincis-paper7.pdf.

Ta, H. T., Bakar, A., Rahman, S., Najjar, L., & Gelbukh, A. (2022b). Multi-Task learning for detection of aggressive and violent incidents from social media. *IberLEF 2022. La Coruña, Spain*, Vol-3202. [Archivo PDF] https://ceur-ws.org/Vol-3202/davincis-paper1.pdf.

Tonja, A. L., Arif, M., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2022). Detection of aggressive and violent incidents from social media in spanish using pre-trained language model. *IberLEF 2022. La Coruña, Spain*, Vol-3202. [Archivo PDF] https://ceur-ws.org/Vol-3202/davincis-paper2.pdf.

Turón, P., Perez, N., García-Pablos, A., Zotova, E., & Cuadros, M. (2022). Vicomtech at DA-VINCIS: Detection of aggressive and violent incidents from social media in spanish. *IberLEF 2022. La Coruña, Spain*, Vol-3202. [Archivo PDF] https://ceur-ws.org/Vol-3202/davincis-paper4.pdf.

Vallejo-Aldana, D., López-Monroy, A. P., & Villatoro-Tello, E. (2022). Leveraging events sub-categories for violent-events detection in social media. *CEUR Workshop Proceedings*, 3202. [Archivo PDF], https://ceur-ws.org/Vol-3202/davincis-paper3.pdf.

Vallejo-Aldana, D., López-Monroy, A. P., & Villatoro-Tello, E. (2023). Enhancing multi-modal classification of violent events using image captioning. *CEUR Workshop Proceedings*, 71. Work submitted for publication.

Van Der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(February). [Archivo PDF] https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

Wei, J. & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/d19-1670.

Zatarain Cabada, R., Barrón Estrada, M. L., Bátiz Beltrán, V. M., Camacho Sapien, A., Leyva López, N., Beltrán Ruiz, G. [U+FFFD], Cárdenas Sainz, A., & Cárdenas López, H. M. (2023). DA-VINCIS at IberLEF 2023: Detecting aggressive and violent incidents from social media in spanish using text information. *CEUR Workshop Proceedings*, 71. Work submitted for publication.

Anexos

A continuación se presentan los resultados obtenidos durante el proceso de experimentación utilizando validación cruzada con el conjunto de entrenamiento de DA-VINCI 2022.

SMOTE.

En la Figura 37 se pueden observar los resultados utilizando SMOTE como técnica de aumento de datos. Utilizando la técnica por sí sola ofrece una mejora para la subtarea 1 en la mayoría de los modelos y al combinarla con *back translation* el modelo Naive Bayes presenta la mejora más notable. Donde se obtiene el máximo de esta técnica es en la subtarea 2 donde Naive Bayes, kNN y SVM presentaron mejoras utilizando SMOTE y SMOTE con *back translation*. Sin embargo, modelos como Random Forest y SVM utilizando embeddings de BERT no presentaron una mejora significativa con respecto a su valor base.

Reemplazo por sinónimo.

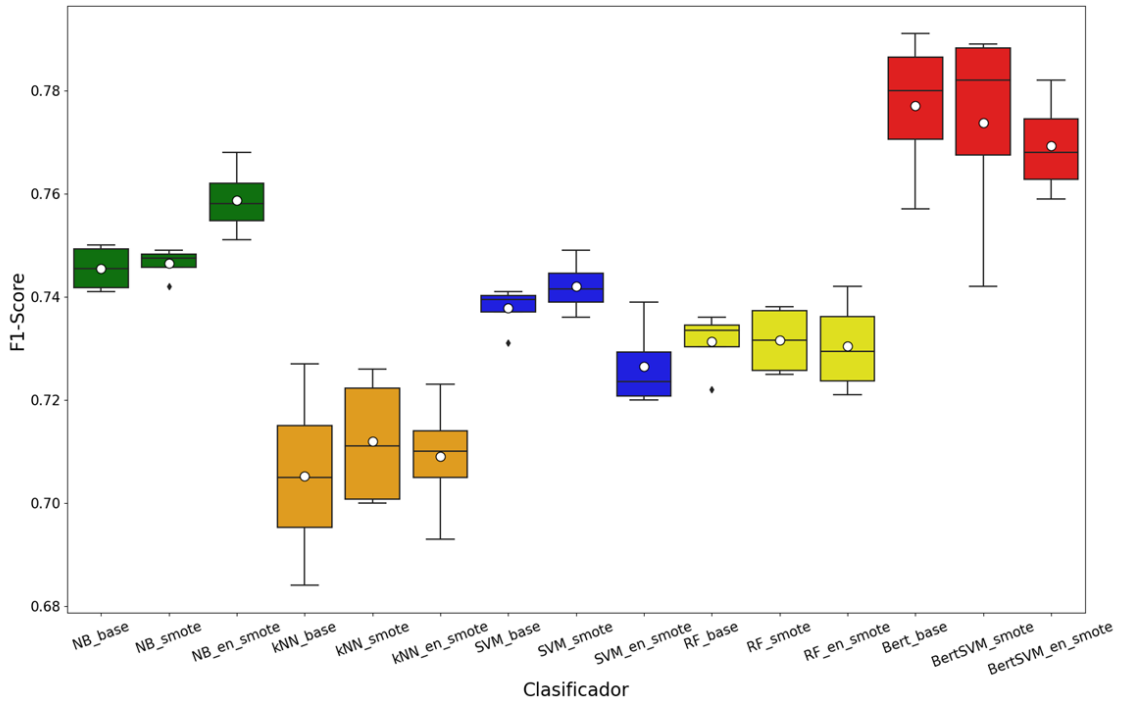
Los resultados con esta técnica se pueden observar en la Figura 38. Para la subtarea 1, los modelos que presentan un mejor rendimiento después del aumento de datos son Naive Bayes y SVM, mientras que para la subtarea 2 todos los modelos presentan mejoras con el aumento de datos, especialmente cuando se utiliza el aumento dos veces, siendo BERT, Random Forest, SVM y kNN los que presentan una mejora en promedio mayor a cuando se aplica sólo una vez.

Back translation.

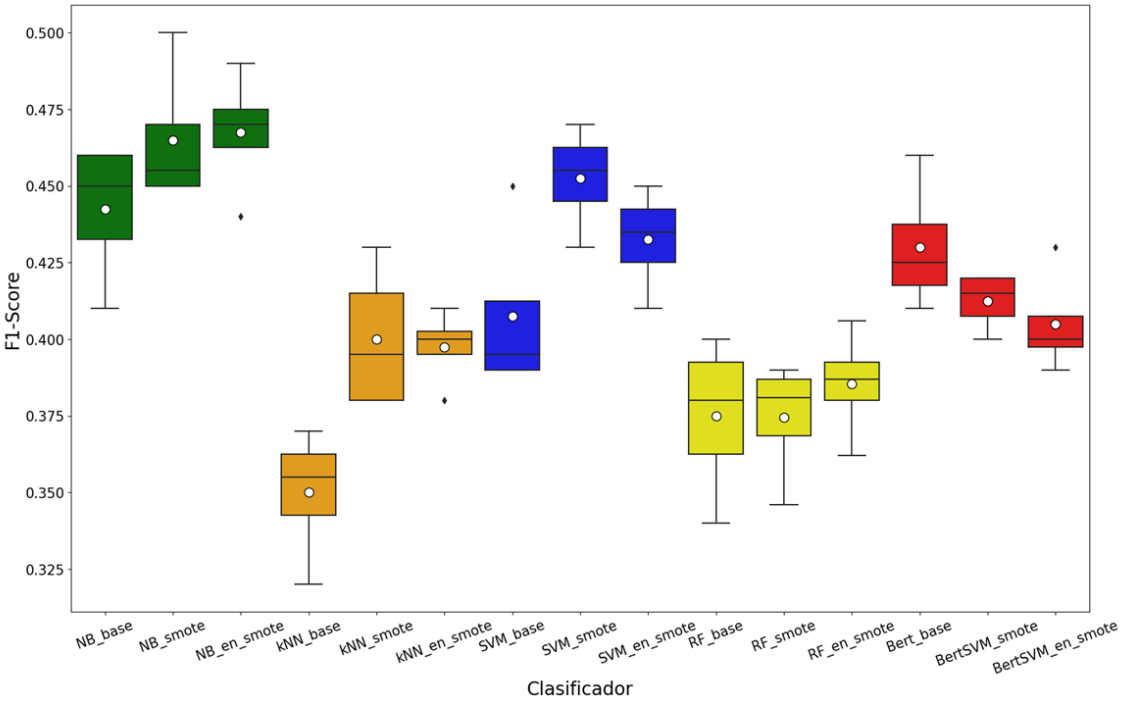
Estos resultados pueden ser observados en la Figura 39. Para la subtarea 1, Naive Bayes fue el modelo que mejor rendimiento con respecto a sus resultados base presentó en las tres combinaciones utilizadas en esta técnica. Seguido por kNN pero presenta una gran dispersión entre sus datos y en promedio compitiendo con Random Forest base. Para la subtarea 2, en promedio todos los modelos presentaron mejoras con las tres combinaciones de instancias.

Modelo de lenguaje grande: GPT-3.

Los resultados para esta técnica pueden observarse en la Figura 40. Para ambas subtareas se pueden observar mejoras en la mayoría de los modelos aunque para la subtarea 1 esto se presenta al utilizar un aumento de 1 vez mientras que para la subtarea 2 se presenta realizando el aumento 3 veces. En la subtarea 1, Naive Bayes y kNN muestran una mejora considerable con respecto a sus resultados sin aumento mientras que SVM incrementa ligeramente en promedio pero a costa de una mayor dispersión en sus resultados.

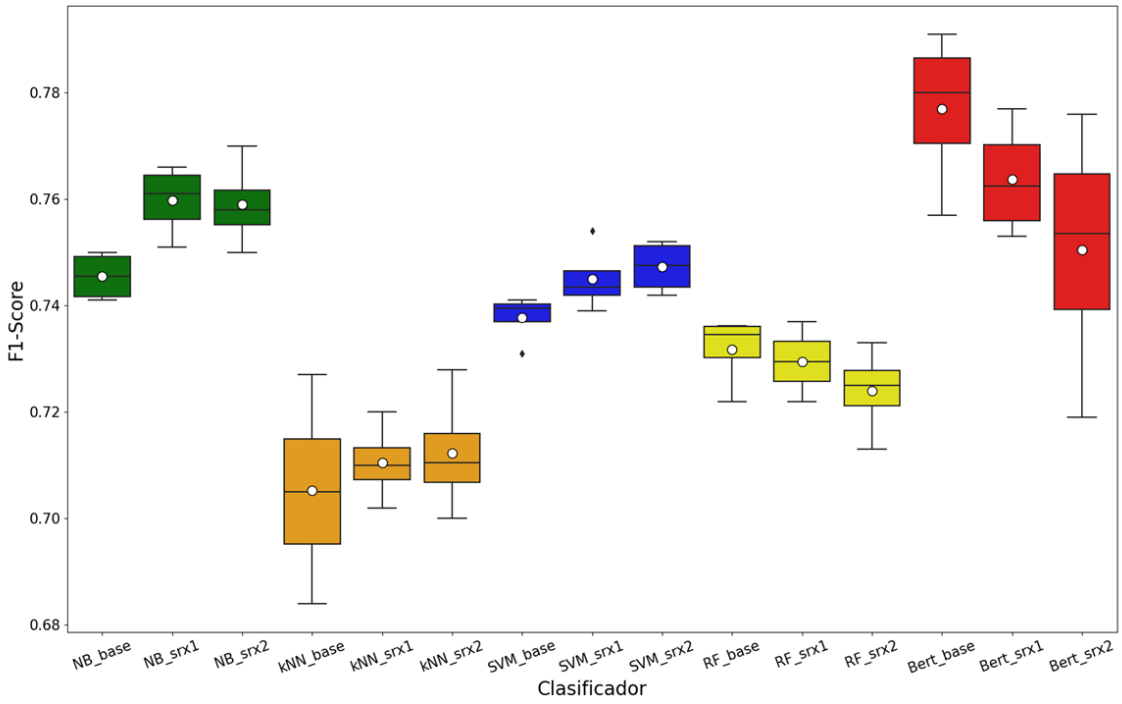


(a) Resultados de todos los clasificadores en validación cruzada al aplicar la técnica SMOTE para la subtarea 1 clasificación binaria.

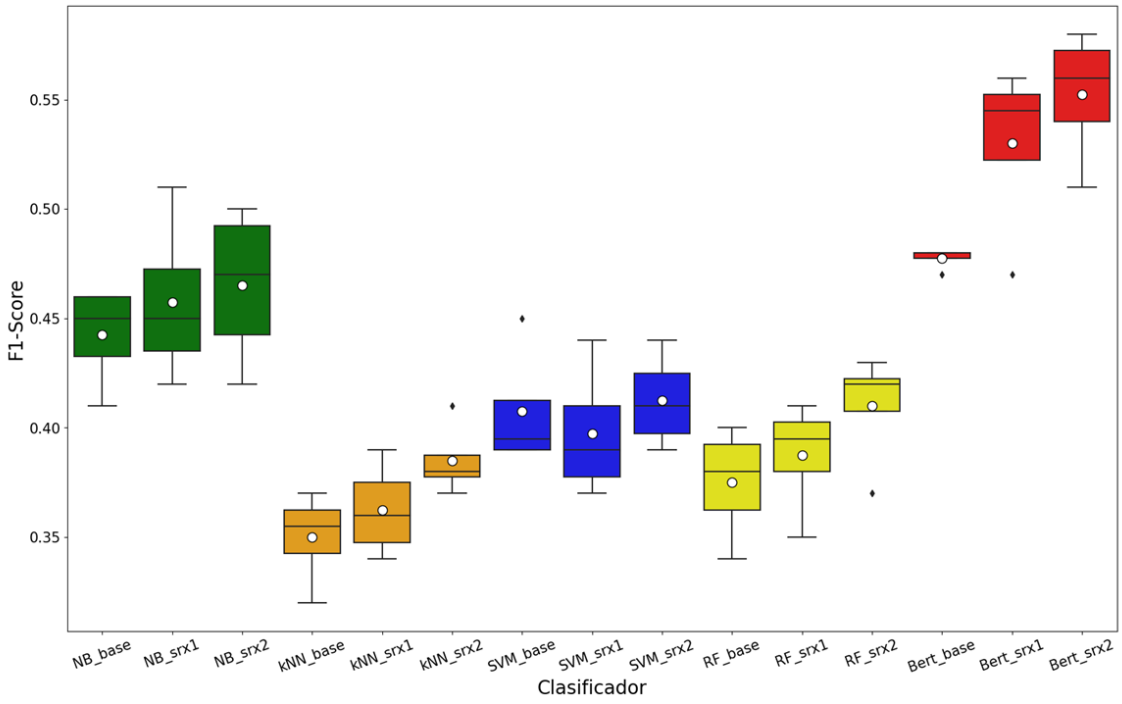


(b) Resultados de todos los clasificadores en validación cruzada al aplicar la técnica SMOTE para la subtarea 2 clasificación multiclase.

Figura 37. Resultados obtenidos al utilizar SMOTE como aumento de datos.

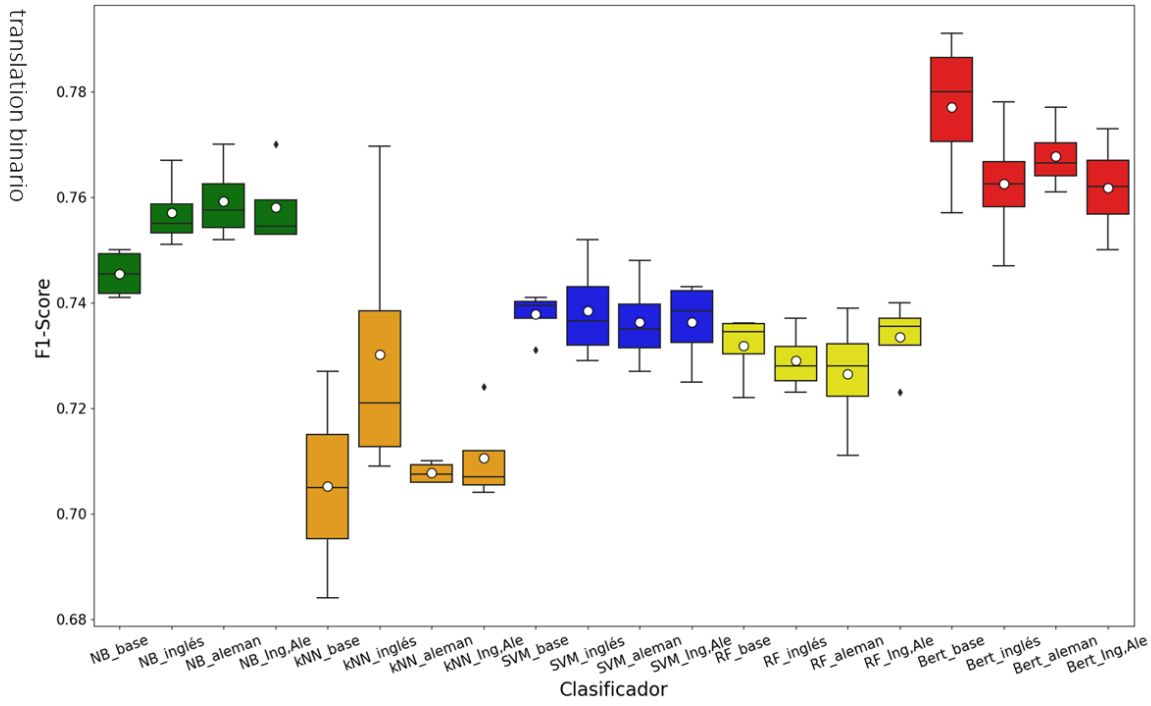


(a) Resultados de todos los clasificadores en validación cruzada al aplicar la técnica reemplazo por sinónimo para la subtarea 1 clasificación binaria.

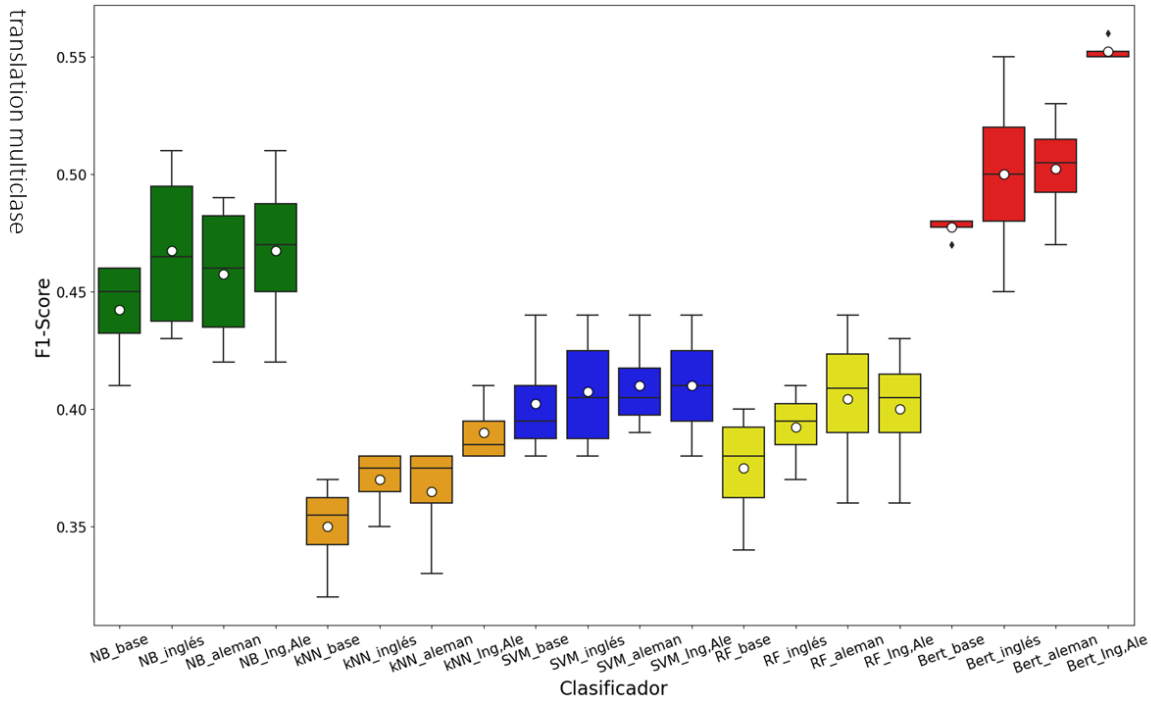


(b) Resultados de todos los clasificadores en validación cruzada al aplicar la técnica reemplazo por sinónimo para la subtarea 2 clasificación multiclase.

Figura 38. Resultados obtenidos al utilizar instancias generadas por reemplazo por sinónimo una y dos veces.

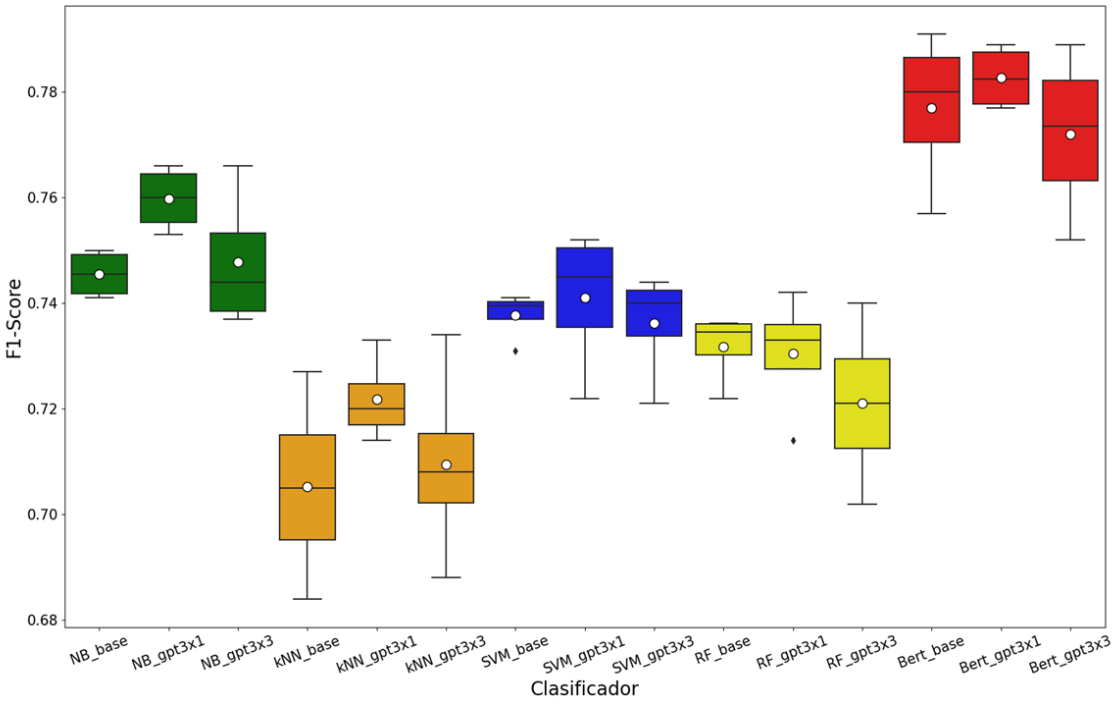


(a) Resultados de todos los clasificadores en validación cruzada al aplicar la técnica *back translation* para la subtarea 1 clasificación binaria.

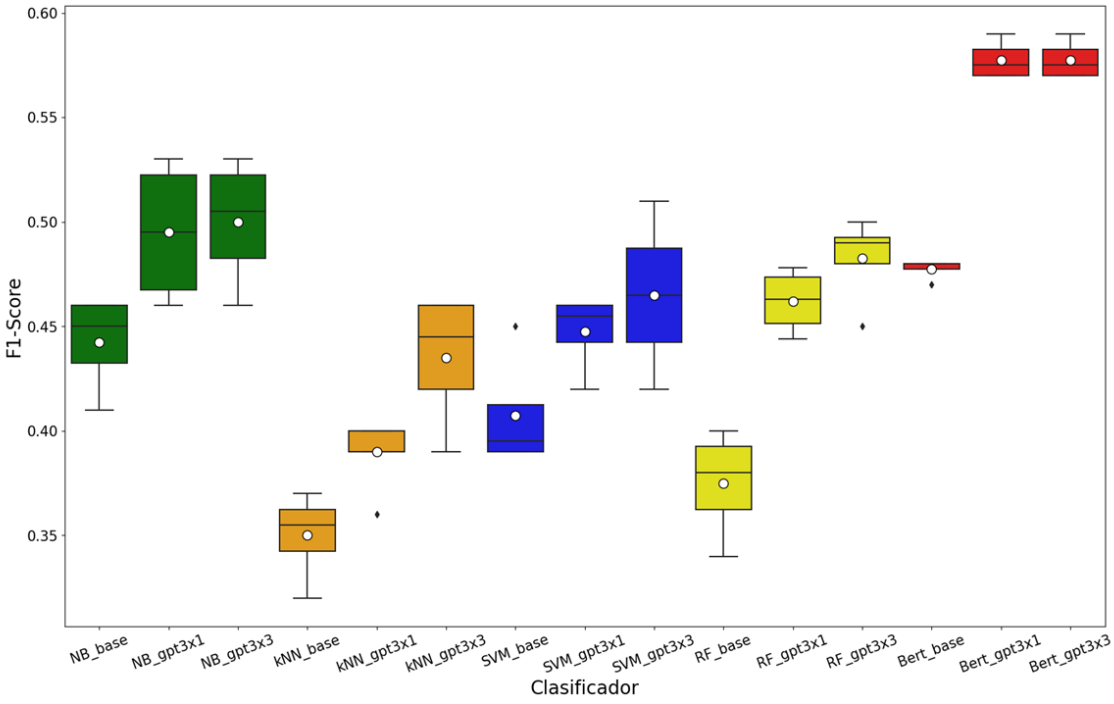


(b) Resultados de todos los clasificadores en validación cruzada al aplicar la técnica *back translation* para la subtarea 2 clasificación multiclase.

Figura 39. Resultados obtenidos al utilizar instancias generadas por *back translation*.



(a) Resultados de todos los clasificadores en validación cruzada al aplicar GPT-3 para generar nuevas instancias para la subtarea 1 clasificación binaria.



(b) Resultados de todos los clasificadores en validación cruzada al aplicar GPT-3 para generar nuevas instancias para la subtarea 2 clasificación multiclase.

Figura 40. Resultados obtenidos al utilizar instancias generadas por GPT-3.