



Universidad del  
**Rosario**

Escuela de Ingeniería,  
Ciencia y Tecnología

# **Visibilizar lo Invisibilizado: Herramienta de Procesamiento de Lenguaje Natural para Sistematizar Noticias de Homicidios de Población LGBTIQ+**

Presentado para obtener el título de

## **MAGÍSTER EN MATEMÁTICAS APLICADAS Y CIENCIAS DE LA COMPUTACIÓN**

Jenny Paola Rivera Burgos

Dirección:

Edgar José Andrade Lotero

Universidad del Rosario

Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Matemáticas Aplicadas y Ciencias de la Computación

## AGRADECIMIENTOS

ii

A Dios y a mi patria, por enseñarme a superar las adversidades y darme un propósito de vida.

A mi mamá, mi papá, Óscar y Sebastián, por ser mi todo: mi raíz, mi inspiración y mi fuerza para seguir adelante.

A Laura, porque más que nadie sabe cuánto ha costado llegar hasta aquí.

A Jerónimo y María, por enseñarme, con sabiduría y ejemplo, a comprender la violencia y el conflicto armado, a construir paz desde lo local y a nunca perder la curiosidad.

A Erika y Mildred, por abrirme las puertas de su hogar y mostrarme el valor de la amistad.

Al running, a Yeffy e Infinity, por enseñarme que soy capaz de hacer cosas difíciles, una zancada a la vez.

A Víctor, por escuchar pacientemente mis obsesiones con los homicidios en Colombia y con mi modelo.

A Edgar, por su guía, su tiempo y su acompañamiento generoso.

A Fabián Sánchez y Felipe Restrepo, jurados de este trabajo de grado, por su revisión y comentarios.

En Colombia, la violencia contra personas LGBTIQ+ continúa siendo una problemática sistemáticamente invisibilizada por las estadísticas oficiales, que carecen de variables diferenciales y de contexto. Diversas organizaciones sociales han recurrido a los medios de comunicación como fuente alternativa para documentar estos casos, enfrentándose al reto de sistematizar grandes volúmenes de información no estructurada. Este trabajo propone una herramienta de procesamiento de lenguaje natural (NLP) para automatizar la recolección, clasificación y extracción de variables clave a partir de noticias digitales sobre homicidios.

El sistema desarrollado abarca todas las etapas del ciclo de vida del aprendizaje automático. Primero, se implementó un proceso de web scraping para recolectar noticias desde Google Noticias. Luego, estas fueron etiquetadas manualmente mediante Label Studio, utilizando un esquema BIO con 24 entidades de interés. Se entrenaron dos modelos basados en BERT, una variante de BERT para español: uno para clasificación binaria (detectar si la noticia corresponde a un homicidio) y otro para reconocimiento de entidades nombradas (NER). La selección de hiperparámetros se realizó mediante búsqueda aleatoria y se incorporó una estrategia de aprendizaje activo basada en entropía y mínima confianza para priorizar ejemplos ambiguos durante la anotación.

El modelo de clasificación alcanzó un accuracy del 85 %, con un F1-score de 0.89 para la clase “relevante”. El modelo NER mostró buen desempeño en variables frecuentes como nombre de la víctima, edad, lugar y fecha del hecho, con F1-scores superiores a 0.70. Finalmente, se desplegó un producto viable mínimo (MVP) en forma de una aplicación web que permite a los usuarios ingresar un rango de fechas y obtener una tabla estructurada con los casos identificados y las entidades extraídas.

Este proyecto representa una contribución técnica, metodológica y política al análisis de violencia por prejuicio en Colombia. Al combinar NLP con una perspectiva de derechos humanos, ofrece una solución replicable para organizaciones sociales, periodistas e investigadoras interesadas en transformar grandes volúmenes de texto en evidencia estructurada para la exigibilidad de justicia.

In Colombia, violence against LGBTIQ+ individuals remains a systematically underreported issue in official statistics, which often lack contextual and differential variables. As a response, several civil society organizations rely on news media as alternative sources to document these cases, facing the challenge of manually processing large volumes of unstructured data. This project proposes a natural language processing (NLP) tool to automate the collection, classification, and extraction of key variables from digital news articles about homicides.

The proposed solution encompasses all stages of the machine learning lifecycle. First, a web scraping pipeline was implemented to collect news articles from Google News. These articles were manually annotated using Label Studio, following a BIO scheme with 24 target entities. Two models based on BERT—a Spanish-adapted version of BERT—were fine-tuned: one for binary classification (to detect whether an article reports a homicide) and another for named entity recognition (NER). Hyperparameters were selected via random search, and an active learning strategy based on entropy and minimum confidence was used to prioritize ambiguous examples for annotation.

The classification model achieved 85% accuracy, with an F1-score of 0.89 for the “relevant” class. The NER model performed well on frequently reported variables such as victim name, age, location, and date of the incident, with F1-scores above 0.70. A minimum viable product (MVP) was also deployed in the form of a web application, allowing users to input a date range and receive a structured table containing the identified cases and extracted entities.

This project constitutes a technical, methodological, and political contribution to the analysis of bias-motivated violence in Colombia. By combining NLP with a human rights perspective, it offers a replicable solution for civil society organizations, journalists, and researchers seeking to convert large volumes of text into structured evidence for justice advocacy.

# TABLA DE CONTENIDO

v

Capítulo 1 INTRODUCCIÓN .....	1
Capítulo 2 OBJETIVOS .....	4
2.1. Objetivo general.....	4
2.2. Objetivos específicos .....	4
Capítulo 3 PROBLEMA Y JUSTIFICACIÓN.....	5
3.1. Problema y contexto social .....	5
3.2. Contribución del proyecto.....	6
Capítulo 4 MARCO TEÓRICO Y ESTADO DEL ARTE.....	8
4.1. Procesamiento de Lenguaje Natural (NLP).....	8
4.2. Sequence Labeling y Named Entity Recognition (NER) .....	8
4.3. Modelos para Sequence Labelling.....	10
4.3.1. Modelos generativos: Hidden Markov Models (HMM).....	10
4.3.2. Modelos discriminativos: Conditional Random Fields (CRF) .....	11
4.3.3. Modelos neuronales: RNN y variantes .....	11
4.3.4. Modelos basados en transformers y LLM .....	13
4.3.4.1. Tipos de modelos .....	13
4.3.4.2. BERT y BETO.....	15
4.3.4.3. Fine-Tuning.....	16
4.3.4.4. Ajuste de hiperparámetros .....	17
4.4. Métricas de evaluación .....	18
4.5. NLP en contexto sociales y de violencia en Colombia.....	19
Capítulo 5 METODOLOGÍA .....	21
5.1. Ciclo de vida del aprendizaje automático aplicado al análisis de noticias sobre homicidios de personas LGBTIQ+ .....	21
5.2. Definición del problema .....	23
5.3. Recolección de datos.....	23
5.3.1. Recopilación de URLs .....	23
5.3.2. Extracción de contenido.....	25
5.4. Preprocesamiento de datos.....	25
5.4.1. Etiquetado manual de datos .....	25
5.4.2. Etiquetado BIO .....	27
5.5. Selección y entrenamiento del modelo .....	28
5.5.1. Modelo de Clasificación binaria .....	28
5.5.2. Modelo NER.....	28
5.5.3. Aprendizaje Activo .....	29
5.6. Evaluación del modelo.....	31
5.7. Despliegue y monitoreo .....	31
Capítulo 6 RESULTADOS Y DISCUSIÓN .....	32
6.1 Modelo de Clasificación Binaria .....	32
6.2. Modelo NER.....	33

6.3 Aplicación Web: Herramienta de Sistematización Automatizada.....	38vi
6.4 Discusión General.....	39
Capítulo 7 CONCLUSIONES Y RECOMENDACIONES.....	41
Recomendaciones .....	42
REFERENCIAS.....	45
ANEXOS .....	47
Anexo 1. Palabras claves para el proceso de webscrapping .....	47
Anexo 2. Palabras NER .....	48

## LISTA DE TABLAS

vii

Tabla 1. Estructura matriz de confusión .....	18
<i>Tabla 2. Métricas Clasificación Binaria</i> .....	33
Tabla 3. Métricas del modelo NER antes y después del entrenamiento .....	34
Tabla 4. Métricas del modelo NER por variables después del entrenamiento .....	36

## LISTA DE FIGURAS

viii

Figura 1. Diseño general de la arquitectura. ....	3
Figura 2. Ejemplo de etiquetado BIO. ....	10
Figura 3. Ejemplo de red neuronal recurrente. ....	12
Figura 4. Funcionamiento de un modelo bidireccional para el reconocimiento de entidades nombradas .....	15
Figura 5. Ejemplo excel de URLs recolectadas .....	24
Figura 7. Ejemplo csv de extracción de texto de URLs.....	25
Figura 8. Ejemplo de un texto etiquetado en Label Studio.....	26
Figura 9. Ejemplo del CSV exportado de Label Studio.....	27
Figura 10. Ejemplo del proceso de etiquetado BIO.....	28
Figura 11. Matriz de confusión – Clasificación Binaria.....	33
Figura 12. Matriz de confusión por variable modelo NER después del entrenamiento .....	37
Figura 13. Ejemplo base de datos resultado.....	39

## Capítulo 1

### INTRODUCCIÓN

En contextos marcados por la desigualdad y la violencia estructural, como el colombiano, el acceso a información confiable, desagregada y oportuna sobre crímenes contra poblaciones históricamente marginadas sigue siendo un desafío persistente. La invisibilización estadística de ciertos grupos, en particular de las personas LGBTIQ+, no solo limita el reconocimiento de las víctimas, sino que también impide la formulación de políticas públicas basadas en evidencia que aborden sus realidades específicas. Esta ausencia de datos diferenciados y sistemáticos en los registros oficiales [1] contrasta con el volumen creciente de información disponible en los medios digitales. Sin embargo, el análisis manual de estas fuentes resulta inviable dada su escala y diversidad.

En este escenario, el procesamiento de lenguaje natural (NLP, por sus siglas en inglés) se presenta como una herramienta potente para enfrentar estas brechas. Este proyecto propone una solución tecnológica para sistematizar noticias sobre homicidios que afectan especialmente a personas LGBTIQ+, mediante técnicas de NLP que automatizan el proceso de recolección, clasificación y extracción de variables clave a partir de textos periodísticos.

El proyecto abarca desde la recolección automatizada de noticias (web scraping), pasando por el etiquetado manual y el entrenamiento de modelos con arquitecturas tipo Transformer, hasta el despliegue de una aplicación web interactiva que permite a los usuarios ingresar un rango de fechas y obtener una tabla de datos estructurada. Esta herramienta busca no solo facilitar el trabajo de sistematización que realizan organizaciones sociales, sino también contribuir a la generación de evidencia para la exigencia de derechos y la formulación de políticas públicas con enfoque de derechos humanos y perspectiva de género y diversidad. El diseño general de la arquitectura se muestra en la Figura 1, la explicación del paso a paso se abordará en el capítulo 5.

El modelo de clasificación binaria, el cual decide si una noticia es sobre un caso de homicidio o no, alcanzó un *accuracy* del 85 %, es decir, en más de 8 de cada 10 casos predijo correctamente. El modelo NER, mostró éxito en aquellas variables que aparecen con mayor frecuencia en las noticias, como nombres y apellidos de la víctima, fecha del hecho, municipio, departamento y edad, con F1-scores superiores al 0.70. Se creó un producto viable mínimo en forma de una aplicación web, la cual permite a un usuario ingresar un rango de fechas y obtener una tabla de datos estructurada donde se relaciona cada una de las noticias encontradas con las variables identificadas.

La estructura del documento es la siguiente: el capítulo 2 presenta los objetivos del proyecto; el capítulo 3 describe el problema, su contexto social y la justificación del estudio; el capítulo 4 expone el marco teórico y estado del arte sobre NLP y NER; el capítulo 5 desarrolla la metodología empleada, basada en el ciclo de vida del aprendizaje automático; el capítulo 6 presenta los resultados y su análisis, destacando el desempeño de los modelos propuestos; y finalmente, el capítulo 7 formula las conclusiones generales y recomendaciones futuras.

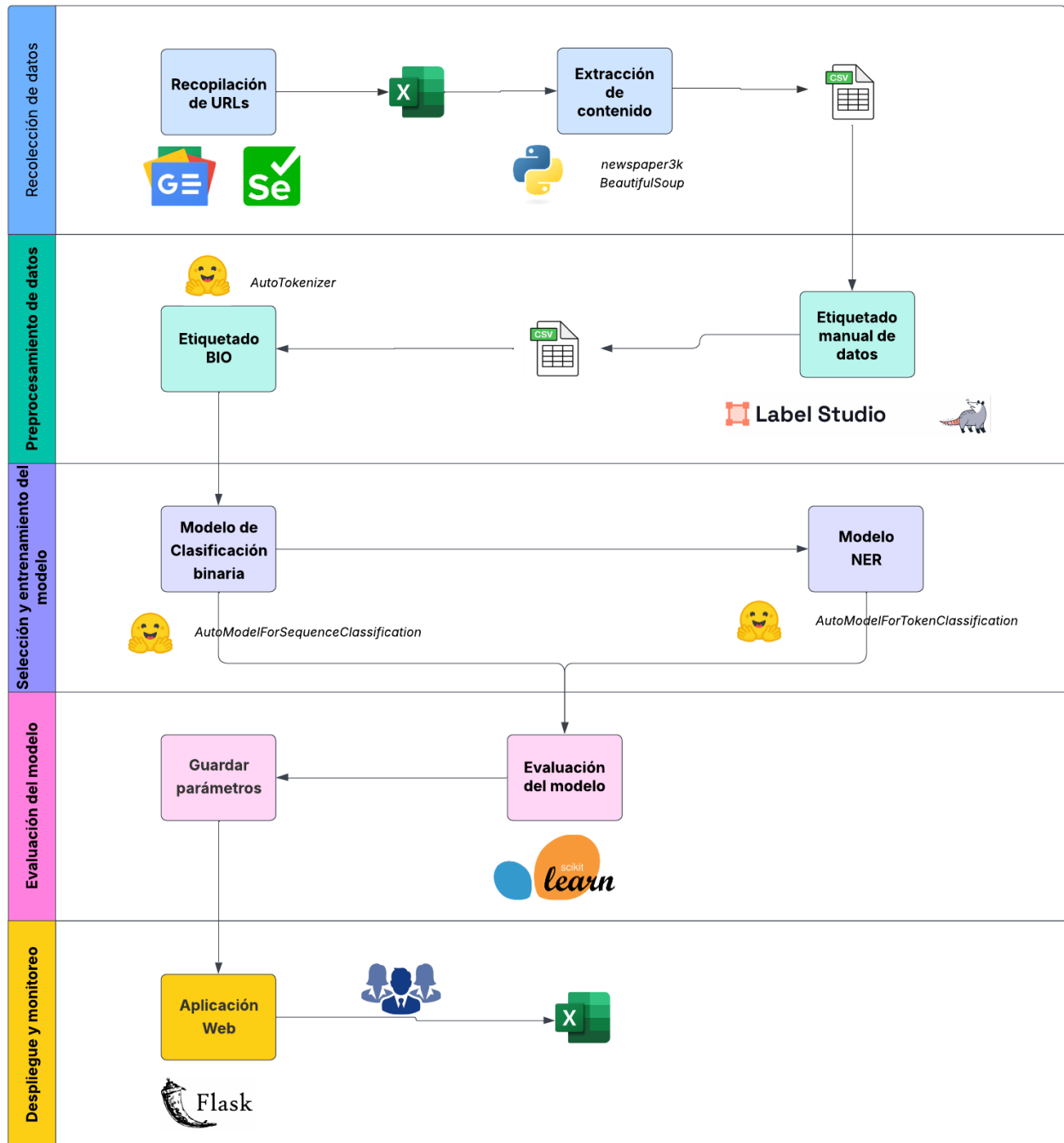


Figura 1. Diseño general de la arquitectura.

Fuente: elaboración propia

## Capítulo 2

### OBJETIVOS

#### 2.1. Objetivo general

- Diseñar, implementar y evaluar una herramienta computacional capaz de transformar una colección de noticias que documentan hechos de violencia contra la población LGBTIQ+ en una tabla de datos estructurada, con variables relevantes para su análisis sistemático y la generación de conocimiento en el ámbito de los estudios sociales y de violencia.

#### 2.2. Objetivos específicos

- Delimitar y operacionalizar las variables de análisis pertinentes para el estudio de noticias sobre homicidios de población LGBTIQ+, a partir de la revisión de literatura y el examen del corpus de datos.
- Construir, depurar y documentar los conjuntos de datos de entrenamiento y prueba, asegurando su representatividad y calidad para el modelo neuronal.
- Seleccionar y configurar el modelo de red neuronal más adecuado para la extracción automatizada de información, con base en criterios de desempeño y viabilidad técnica.
- Entrenar el modelo utilizando metodologías reproducibles y evaluar su rendimiento mediante métricas estándar en el procesamiento de lenguaje natural.
- Analizar e interpretar los resultados del modelo, identificando fortalezas, limitaciones y oportunidades de mejora.
- Implementar la herramienta en una plataforma web bajo la modalidad de *Producto Mínimo Viable (MVP)*, garantizando su funcionalidad básica

## Capítulo 3

### PROBLEMA Y JUSTIFICACIÓN

#### 3.1. Problema y contexto social

En Colombia, las causas de los homicidios son complejas y multifactoriales. Van desde la violencia interpersonal e intrafamiliar, el consumo de sustancias, la delincuencia organizada, la disputa por territorios y economías ilegales, entre otros. Esta violencia afecta de manera desproporcionada y diferencial a poblaciones históricamente marginadas y vulnerables como mujeres, personas LGBTIQ+, comunidades étnicas, líderes sociales, migrantes y jóvenes en entornos urbanos [2]. Pese a los esfuerzos institucionales por registrar y analizar estos crímenes, las estadísticas oficiales continúan siendo limitadas, fragmentadas y carentes de variables diferenciales y de contexto que permitan visibilizar adecuadamente a las víctimas y los motivos detrás de estos hechos [1].

Ante esta realidad, diversas organizaciones sociales han asumido la tarea de recopilar información sobre grupos poblacionales específicos para comprender las particularidades de las dinámicas de violencia. El Instituto de Estudios para el Desarrollo y la Paz (Indepaz) documenta violencia contra líderes sociales, defensores de derechos humanos y firmantes del Acuerdo de Paz; la Organización Nacional Indígena de Colombia (ONIC) contra comunidades indígenas; COALICO contra niñas, niños y adolescentes en contextos de conflicto armado; y Caribe Afirmativo junto con Colombia Diversa contra personas LGBTIQ+.

Buscando recopilar el universo de hechos de violencia contra su población de estudio, varias de estas organizaciones recurren a distintas fuentes de información, como lo son: reportes emitidos por parte de organizaciones sociales, víctimas, fuentes oficiales y medios de comunicación de carácter local, regional, nacional e internacional [3]. En este contexto, el monitoreo de prensa constituye un insumo clave al complementar y servir de soporte a las fuentes oficiales. Su principal ventaja radica en la disponibilidad inmediata de información, la cual en muchos casos aporta detalles adicionales sobre las víctimas y el contexto de los hechos. [4]

En sus informes anuales de violencia contra población LGBTIQ+, Colombia Diversa presenta de manera detallada la metodología utilizada para la recopilación de estos hechos de violencia. Ellos utilizan dos fuentes: registros oficiales y noticias periodísticas. Sin embargo, ambas presentan limitaciones significativas. Por un lado, los registros oficiales no están disponibles en tiempo real, suelen presentar rezagos de varios meses y carecen de variables diferenciales y de contexto. La ausencia de variables diferenciales, como la orientación sexual o identidad de género, en los sistemas de información oficiales contribuye a la invisibilización sistemática de las víctimas LGBTIQ+, dificultando el análisis de patrones delictivos y la formulación de respuestas institucionales con enfoque de derechos humanos [1].

Por otro lado, las noticias periodísticas, si bien ofrecen información más inmediata, se presentan en formatos no estructurados, es decir, tipo texto que requiere procesos manuales de monitoreo, recolección y sistematización de noticias. Este trabajo implica revisar diariamente múltiples fuentes periodísticas, extraer y codificar datos caso por caso, y contrastarlos con respuestas institucionales obtenidas a través de derechos de petición. A pesar de ser un proceso altamente demandante en términos de tiempo y recursos, ha sido esencial para suplir las falencias de información del Estado.

### **3.2. Contribución del proyecto**

Dado lo anterior, este trabajo de grado busca apoyar y facilitar el trabajo de estas organizaciones mediante la creación de bases de datos estructuradas a partir de noticias periodísticas, que incluya las variables de interés que se han identificado como pertinentes para el estudio sistemático de homicidios. Para lograr lo anterior, se propone el diseño, entrenamiento y despliegue de un modelo de clasificación y de reconocimiento de NER basado en arquitecturas tipo transformer, ajustado al idioma español, para estructurar automáticamente la información contenida en noticias sobre homicidios, con un énfasis particular en casos que involucran a personas LGBTIQ+, aunque aplicable a diferentes contextos. El modelo permite extraer entidades como nombre de la víctima, edad, género,

lugar, arma utilizada, motivo del crimen y relación con el perpetrador, entre otras variables relevantes.

La contribución de este proyecto se sitúa en varios niveles:

- **Técnico:** El proyecto plantea el desarrollo de un modelo de NER entrenado sobre noticias en español, utilizando la arquitectura BETO. La novedad radica en su aplicación a textos periodísticos para detectar múltiples variables en un solo pipeline, adaptado al contexto colombiano.
- **Social y político:** contribuye a la visibilización de las violencias por prejuicio, ofreciendo datos que permiten evidenciar patrones sistemáticos y apoyar procesos de denuncia y exigencia de justicia. Como destaca Colombia Diversa, la ausencia de datos completos impide garantizar los derechos de las víctimas LGBTIQ+.
- **Metodológico:** Se implementa una solución integral que abarca webscraping, anotación manual, entrenamiento supervisado y despliegue web. Esto permite una sistematización automatizada, reproducible y adaptable a otros contextos sociales.

En un país donde, según *Caribe Afirmativo*, al menos el 42 % de los datos carecían de información detallada y donde solo tres casos llegaron a la etapa de ejecución de penas en 2024 [5], una herramienta de NLP no solo representa una innovación técnica, sino también un acto de reparación y justicia simbólica: permite nombrar, sistematizar y visibilizar lo que ha sido históricamente silenciado.

## Capítulo 4

### MARCO TEÓRICO Y ESTADO DEL ARTE

#### 4.1. Procesamiento de Lenguaje Natural (NLP)

Los textos, y por tanto el lenguaje natural, han sido una fuente principal de información en las ciencias sociales. Hoy en día, la enorme cantidad de información disponible - noticias, registros administrativos, entrevistas, redes sociales - hace que su análisis manual sea prácticamente imposible. Además, la necesidad de presentar esta información de forma cuantitativa ha incrementado la demanda de técnicas que permitan extraer los datos más relevantes de los corpus disponibles y presentarlos de forma estructurada. Estos factores, junto con los avances en poder computacional, han dado origen a las ciencias sociales computacionales, una disciplina que aplica métodos computacionales para estudiar y comprender fenómenos sociales. [6]

Uno de los métodos computacionales más aplicados para este fin es el **procesamiento de lenguaje natural** (NLP, por sus siglas en inglés). El NLP es un campo interdisciplinar que combina la lingüística, la informática y la inteligencia artificial, y que tiene como objetivo lograr que las máquinas analicen, comprendan, generen e interactúen con el lenguaje humano de manera efectiva. A través de una colección de técnicas computacionales, el NLP busca automatizar tareas como la clasificación de textos, la extracción de entidades, la detección de sentimientos y la respuesta a preguntas, entre muchas otras. [7]

Las siguientes dos secciones de este capítulo, se basan en los capítulos 8, 9, 11 y 17 de *Speech and Language Processing* de Jurafsky y Martin [8].

#### 4.2. Sequence Labeling y Named Entity Recognition (NER)

Una de las aplicaciones más importantes del NLP es el **etiquetado de secuencias** (*sequence labeling*), una tarea que consiste en asignar a cada palabra  $x_i$  de una secuencia una etiqueta correspondiente  $y_i$ . De esta forma, se genera una secuencia de etiquetas del mismo tamaño que la secuencia de entrada, permitiendo identificar categorías gramaticales, entidades nombradas u otras características lingüísticas en contexto.

La asignación de etiquetas suele basarse en una función de probabilidad, siendo la más común el uso de la función *softmax*. Esta función transforma un vector de puntuaciones  $z = [z_1, z_2, z_3, \dots, z_K]$ , donde cada  $z_i$  representa la afinidad de la palabra con la  $i$ , y  $K$  es el número total de etiquetas posibles, en una distribución de probabilidad. El modelo predice la etiqueta más probable como:

$$y_i = \operatorname{argmax} P(y = 1|z)$$

Entre las tareas más frecuentes de *sequence labeling* se encuentran el **NER** y el **etiquetado gramatical** (*Part-of-Speech tagging*, POS).

- **POS** consiste en clasificar cada palabra según su categoría gramatical (sustantivo, verbo, adverbio, preposición, pronombre, etc.).
- **NER**, busca identificar y clasificar entidades específicas en un texto, como nombres de personas, lugares, organizaciones, entidades geopolíticas, fechas, entre otros.

Una estrategia común para abordar el etiquetado en NER es el uso del esquema **BIO** (**B**eginning, **I**nside, **O**utside) [8]. En este método:

- **B**: se asigna al primer token de una entidad,
- **I**: a los tokens subsecuentes dentro de la misma entidad, y
- **O**: a los tokens que no pertenecen a ninguna entidad.

Este esquema permite identificar no solo la clase de la entidad, sino también sus límites dentro del texto. El número total de etiquetas en este enfoque es  $2n + 1$ , donde  $n$  es el número de clases de entidades, dado que cada clase requiere una etiqueta B y una I, más la etiqueta O. La Figura 2 presenta un ejemplo de etiquetado BIO.

Palabra	Etiqueta
La	O
víctima	O
fue	O
identificada	O
como	O

Hugo	B-Víctima Nombre
Alexander	I-Víctima Nombre
Ramírez	B-Víctima Apellido
un	O
estilista	B-Víctima Profesión
de	O
49	B-Víctima edad
años	I-Víctima edad

Figura 2. Ejemplo de etiquetado BIO.

Fuente: elaboración propia

### 4.3. Modelos para Sequence Labelling

El etiquetado de secuencias puede abordarse con diferentes tipos de modelos, los cuales han evolucionado desde enfoques probabilísticos hasta arquitecturas basadas en modelos de lenguaje de gran escala (*Large Language Models, LLM*). Los enfoques más comunes incluyen modelos generativos, discriminativos, redes neuronales y, más recientemente, los modelos basados en *transformers*. A continuación, se enuncian algunos de los modelos más utilizados, junto con sus respectivas ventajas y limitaciones.

#### 4.3.1. Modelos generativos: Hidden Markov Models (HMM)

Los **modelos generativos**, como los Modelos Ocultos de Markov (HMM, por sus siglas en inglés), buscan entender como una clase específica podría haber generado una determinada observación. Para ello, modelan la distribución conjunta entre una secuencia de observaciones  $X = (x_1, x_2, \dots, x_n)$  y una secuencia de etiquetas  $Y = (y_1, y_2, \dots, y_n)$ . La probabilidad conjunta se expresa como:

$$P(X, Y) = P(y_1) \prod_{t=2}^n P(y_t | y_{t-1}) \prod_{t=1}^n P(x_t | y_t)$$

Este enfoque asume que la etiqueta actual depende únicamente de la anterior (propiedad de Markov) y que cada observación depende solo de su etiqueta correspondiente. Lo anterior implica que no se consideran independencias entre las observaciones. Aunque

eficaces en tareas sencillas, estos modelos presentan limitaciones en NLP, ya que no capturan adecuadamente el contexto completo del texto.

#### 4.3.2. Modelos discriminativos: Conditional Random Fields (CRF)

Los **modelos discriminativos**, como los Campos Aleatorios Condicionales (CRF, por sus siglas en inglés), en lugar de modelar cómo se generan los datos, se centran en aprender la probabilidad condicional  $P(Y|X)$ , es decir, la mejor secuencia de etiquetas dada una secuencia de entrada. A diferencia de los modelos generativos, los discriminativos permiten incorporar características contextuales de toda la secuencia, lo que los hace más efectivos en tareas como el reconocimiento de entidades y el análisis sintáctico.

#### 4.3.3. Modelos neuronales: RNN y variantes

Con el auge del aprendizaje profundo, los **modelos neuronales** como las Redes Neuronales Recurrentes (RNN, por sus siglas en inglés), se introdujeron para superar las limitaciones de los modelos tradicionales. Las RNN procesan secuencias de manera recurrente, manteniendo un estado oculto  $h_t$  que resume la información anterior:

$$h_t = RNN(x_t, h_{t-1})$$

La Figura 3 ilustra el funcionamiento de una RNN. Cada estado oculto  $h_t$  se obtiene a partir de la combinación de la entrada actual  $x_t$  y del estado oculto anterior  $h_{t-1}$ . Este mecanismo le permite al modelo retener información del pasado y actualizarla con nuevos datos en cada paso de la secuencia. A su vez, el estado oculto  $h_t$  genera una salida  $y_t$ , que

corresponde a la predicción asociada a dicho instante.

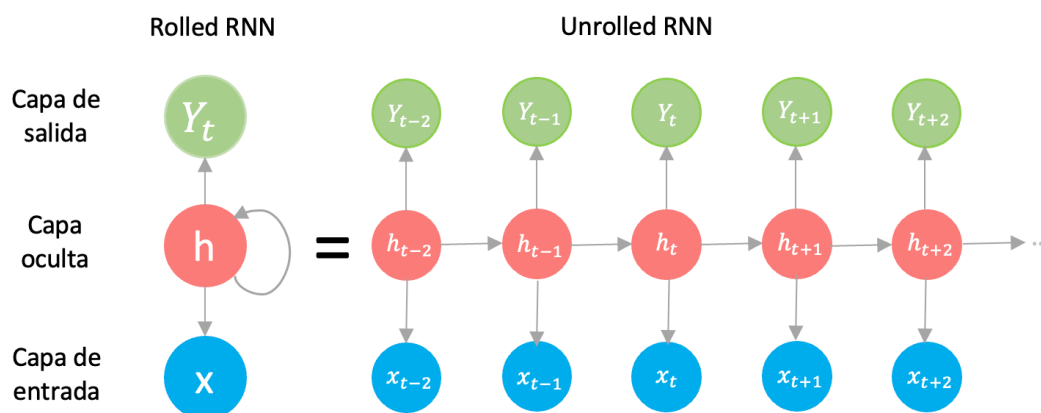


Figura 3. Ejemplo de red neuronal recurrente.

Fuente: elaboración propia

Este enfoque permite capturar dependencias a largo plazo en la secuencia. Una capa de clasificación *softmax* suele aplicarse sobre el estado oculto para predecir la etiqueta correspondiente a cada token. Adicionalmente, estas redes pueden combinarse con CRF para optimizar la secuencia de etiquetas considerando dependencias entre ellas.

Sin embargo, las RNN tradicionales enfrentan el problema del desvanecimiento del gradiente (*vanishing gradient problem*), que ocurre durante el entrenamiento con retropropagación a través del tiempo (*Backpropagation Through Time*, BPTT). En este proceso, los gradientes que se utilizan para ajustar los parámetros se multiplican repetidamente al retroceder en cada paso temporal. Cuando los valores son menores que 1, estas multiplicaciones sucesivas hacen que los gradientes se reduzcan exponencialmente, llegando a ser prácticamente nulos. Como consecuencia, la red tiene grandes dificultades para aprender dependencias a largo plazo, ya que las actualizaciones de los parámetros dejan de reflejar la influencia de entradas distantes en la secuencia. Para superar esta limitación, surgieron arquitecturas como LSTM (*Long Short-Term Memory*) y GRU (*Gated Recurrent Unit*), que incorporan compuertas de control del flujo de información y

permiten preservar gradientes estables, mejorando así la capacidad de retención de información relevante y mitigando el problema del desvanecimiento del gradiente.

#### 4.3.4. Modelos basados en transformers y LLM

Actualmente, los modelos de LLM basados en *transformers* se han convertido en los más relevantes y avanzados para el etiquetado de secuencias. La eficiencia de estos modelos y su adaptabilidad para diversas aplicaciones se debe a tres factores principales: el mecanismo de autoatención, el preentrenamiento y el gran número de parámetros [9].

A diferencia de las RNN, los *transformers* procesan toda la secuencia simultáneamente utilizando un **mecanismo de autoatención** (*self-attention*), que permite al modelo ponderar la relevancia de cada palabra en relación con las demás, sin importar su posición en el texto. La fórmula básica del mecanismo de atención es:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Donde Q, K y V son matrices que representan las consultas, claves y valores respectivamente, y  $d_k$  es la dimensión de las claves.

Adicionalmente, estos modelos se preentrenan sobre grandes corpus de texto sin etiquetas para aprender representaciones profundas del lenguaje, lo que se conoce como **preentrenamiento**. Por último, el gran número de parámetros, siendo estos los elementos internos del modelo que se ajustan durante el entrenamiento para aprender el patrón de los datos pueden ser adaptados a tareas específicas mediante un proceso de **ajuste fino** (*fine-tuning*).

##### 4.3.4.1. Tipos de modelos

Dentro de los modelos de LLM basados en *transformers*, es posible distinguir entre modelos según la direccionalidad con la que procesan la información textual.

- **Modelos unidireccionales o causales:** procesan el texto en una sola dirección, ya sea de izquierda a derecha o de derecha a izquierda. Estos modelos se utilizan principalmente en tareas de generación de lenguaje, donde cada palabra se predice en función de las anteriores, lo que corresponde al proceso de *decoding*. Su fortaleza radica en la capacidad de producir texto de manera secuencial y coherente, aunque su limitación es no poder aprovechar el contexto completo de la secuencia simultáneamente.
- **Modelos bidireccionales:** procesan el texto completa de manera simultánea, incorporando tanto el contexto precedente como el subsiguiente de cada palabra. Este enfoque, utilizado por BERT y sus variantes (como BETO en español), se ajusta a tareas de *encoding*, las cuales toman una secuencia de palabras y la transforma en una representación vectorial contextualizada de la misma, capturando el contexto de toda la secuencia. Esta bidireccionalidad hace que estos modelos sean particularmente efectivos en tareas de comprensión de texto, clasificación de secuencias y reconocimiento de entidades nombradas.

La Figura 4 ilustra el funcionamiento de un modelo bidireccional aplicado a la tarea de NER. Cada token de la oración, incluyendo los símbolos especiales como [CLS<sup>1</sup>], se transforma inicialmente en un vector de representación (*embedding*). Estos vectores son procesados por el *bidirectional encoder* mediante el mecanismo de *self-attention*, lo que permite contextualizar cada token utilizando la información de toda la secuencia. El

---

<sup>1</sup> El token [CLS] (*classification*) es un marcador especial de la arquitectura de BERT. Se coloca al comienzo de toda secuencia de entrada y cumple la función de generar una representación vectorial agregada del texto completo. Durante el proceso de encoding, este token se entrena para capturar información global de la secuencia, de manera que su embedding final pueda ser utilizado como una representación compacta de todo el documento o frase.

resultado son representaciones vectoriales enriquecidas con contexto  $h_i$  que capturan el significado de cada palabra en relación con las demás. Finalmente, estas representaciones son proyectadas a través de una capa de clasificación (*NER head*), que asigna etiquetas siguiendo el esquema BIO. En el ejemplo mostrado, el modelo identifica correctamente “Jane Villanueva” como una persona (B-PER, I-PER) y “United Airlines Holding” como una organización (B-ORG, I-ORG, I-ORG), mientras que tokens irrelevantes como “of” o “discussed” son etiquetados como O.

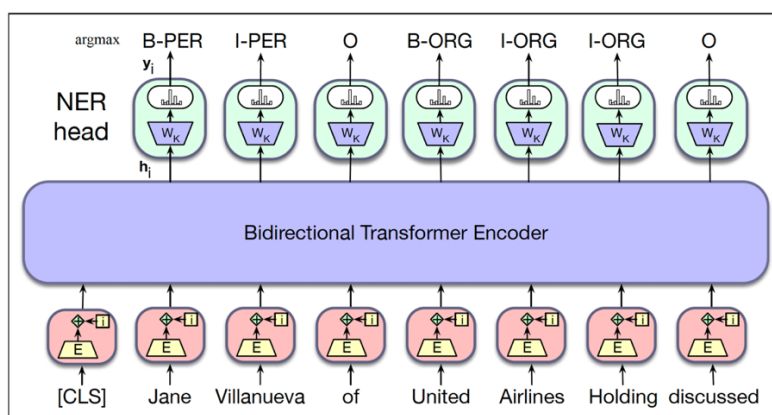


Figura 4. Funcionamiento de un modelo bidireccional para el reconocimiento de entidades nombradas

Fuente: Tomado de *Speech and Language Processing de Jurafsky y Martin* [8]

#### 4.3.4.2. BERT y BETO

Uno de los modelos más relevantes en NLP es **BERT** (*Bidirectional Encoder Representations from Transformers*) [10]. Este es un modelo bidireccional que fue preentrenado de manera no supervisada utilizando los corpus de *BooksCorpus* y *Wikipedia* en inglés, con el objetivo de resolver dos tareas principales:

- *Masked Language Modeling*: predicción de palabras ocultas dentro de una oración
- *Next Sentence Prediction*: predicción de la oración siguiente dada una oración anterior.

BERT ha sido ampliamente adoptado en distintas tareas y contextos, debido a su código abierto, lo que permite adaptarlo mediante *fine-tuning* sin necesidad de preentrenarlo desde cero.

Para el idioma español, una adaptación relevante es BETO [11], un modelo basado en BERT entrenado con un corpus en español y utilizando la técnica de *Whole Word Masking*. Existen dos versiones, *cased*, el cual considera las mayúsculas, y el *uncased*, el que no las considera. Con más de 109 millones de parámetros, BETO ha demostrado un alto desempeño en tareas de etiquetado de secuencias, como NER.

#### **4.3.4.3.Fine-Tuning**

A pesar de que los modelos de lenguaje de gran escala se entrenan con enormes volúmenes de datos provenientes de múltiples dominios, con frecuencia surge la necesidad de adaptarlos a contextos específicos que no estuvieron suficientemente representados durante el preentrenamiento. Este es el caso, por ejemplo, de aplicaciones en textos legales o médicos, de la adaptación de un modelo multilingüe a un idioma particular, o de la especialización en una tarea concreta. Para ello, se recurre al fine-tuning, que consiste en continuar el entrenamiento de un modelo previamente preentrenado utilizando datos adicionales del dominio, idioma o tarea de interés, con el fin de mejorar su desempeño en ese contexto específico. Existen distintos tipos de *fine-tuning*, y su uso depende del poder computacional y la tarea a realizar.

El de mayor demanda computacional es el denominado *continued pre-training*, el cual *consiste* en reentrenar todos los parámetros sobre un nuevo corpus, utilizando el mismo procedimiento y función de pérdida del preentrenamiento. Como alternativa, surge el enfoque de fine-tuning eficiente en parámetros (*Parameter-Efficient Fine-Tuning, PEFT*), en el cual se congelan la mayoría de los parámetros previamente preentrenados y se ajusta únicamente un subconjunto de ellos. De esta manera, se logra una adaptación al nuevo dominio o tarea con un uso más eficiente de los recursos computacionales, manteniendo al mismo tiempo la información general adquirida durante el preentrenamiento.

También hay procesos de fine-tuning para mejorar el rendimiento en la realización de tareas específicas (*Task-adaptive pre-training*), como clasificación y NER. En estos casos, el *fine-tuning* se implementa mediante la adición de capas de clasificación específicas (*classification head*) sobre el modelo preentrenado. En el caso de la clasificación de texto, la capa de clasificación suele ser una red totalmente conectada que toma como entrada la representación global de la secuencia, generalmente asociada al token especial [CLS] en modelos como BERT, y produce una etiqueta única para el documento. Por su parte, en el NER, la capa de clasificación opera a nivel de tokens: cada representación generada por el transformador se proyecta en un espacio de etiquetas correspondientes a las categorías definidas (por ejemplo, *B-PER*, *I-PER*, *O*). Posteriormente, se aplica una función de decisión como *softmax* o *CRF*. En ambos casos, el proceso de *fine-tuning* tiende a congelar parcial o totalmente los parámetros del modelo base, concentrando el entrenamiento en los parámetros recién introducidos y adaptados a la tarea específica.

#### 4.3.4.4. Ajuste de hiperparámetros

A diferencia de los parámetros regulares de un modelo, que son aprendidos automáticamente a partir de los datos durante el proceso de entrenamiento o preentrenamiento, los **hiperparámetros** constituyen parámetros especiales definidos por el investigador. Estos valores no se estiman directamente del conjunto de entrenamiento, sino que determinan la forma en que el algoritmo aprende y generaliza, influyendo en aspectos como la velocidad de convergencia, la capacidad de representación y la prevención del sobreajuste. Una configuración inadecuada puede conducir a un aprendizaje deficiente o a problemas de sobreajuste, comprometiendo la generalización del modelo. Los hiperparámetros más comunes son la **tasa de aprendizaje**, el **tamaño del lote** (*batch size*), el **número de épocas** y el **peso de regularización** (*weight decay*).

Existen diferentes estrategias para definir los hiperparámetros óptimos. Entre las más utilizadas se encuentran el **Grid Search**, que evalúa de manera exhaustiva todas las combinaciones posibles dentro de un espacio predefinido, y el **Random Search**, que selecciona aleatoriamente un subconjunto de configuraciones. Aunque el primero garantiza

una exploración completa, resulta computacionalmente costoso en problemas de alta dimensionalidad. En contraste, el segundo reduce los requerimientos computacionales y ha demostrado ser más eficiente en la práctica, especialmente cuando solo una fracción de los hiperparámetros tiene un efecto determinante en el rendimiento del modelo [12].

#### 4.4. Métricas de evaluación

Para evaluar el rendimiento de un modelo NER, se utilizan métricas derivadas de la **matriz de confusión**, la cual permite comparar las etiquetas reales con las etiquetas predichas por el modelo. Esta matriz se compone de las siguientes categorías:

*Tabla 1. Estructura matriz de confusión*

	<b>Predicción: Positiva</b>	<b>Predicción: Negativa</b>
<b>Real: Positiva</b>	Verdadero Positivo (VP)	Falso Negativo (FN)
<b>Real: Negativa</b>	Falso Positivo (FP)	Verdadero Negativo (VN)

*Fuente: elaboración propia*

- **Verdadero Positivo (VP):** El modelo identificó correctamente una entidad presente en el texto.
- **Falso Positivo (FP):** El modelo etiquetó como entidad un fragmento que no lo era.
- **Falso Negativo (FN):** El modelo no logró identificar una entidad que sí estaba presente.
- **Verdadero Negativo (VN):** El modelo correctamente no etiquetó como entidad un token irrelevante.

A partir de esta matriz se derivan tres métricas fundamentales para evaluar modelos NER:

- **Precisión (P):** Indica la proporción de entidades correctamente identificadas sobre el total de entidades predichas por el modelo. Es útil para evaluar la **exactitud** de las predicciones:

$$\text{Precisión} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}}$$

- **Exhaustividad o cobertura (Recall, R):** Mide la proporción de entidades verdaderas que fueron correctamente identificadas por el modelo. Evalúa la **capacidad de recuperación** de entidades relevantes::

$$\text{Recall} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

**Puntaje F1 (F1-score):** Es la media armónica entre la precisión y el recall. Es especialmente útil cuando se busca un balance entre ambas métricas, penalizando los casos en los que una es alta y la otra baja:

$$F1 = 2 * \frac{\text{Precisión} * \text{Recall}}{\text{Precisión} + \text{Recall}}$$

Estas métricas proporcionan una visión integral del desempeño del modelo, especialmente en contextos donde el desbalance entre clases (entidad vs. no entidad) es común, como ocurre frecuentemente en tareas de etiquetado de texto.

#### 4.5. NLP en contexto sociales y de violencia en Colombia

En Colombia, diversas entidades estatales y académicas han comenzado a aplicar técnicas de NLP para abordar problemáticas sociales, jurídicas y de violencia. Estas herramientas han demostrado ser valiosas para analizar grandes volúmenes de texto de forma sistemática, identificar patrones y facilitar la toma de decisiones informadas.

Un ejemplo es **PretorIA**, una herramienta desarrollada por la Corte Constitucional, que utiliza inteligencia artificial para clasificar y agrupar automáticamente las más de 2.500 sentencias que recibe diariamente la Corte. Su objetivo es organizar los fallos por similitud

de casos, facilitando así la identificación de precedentes y la priorización de decisiones relevantes [13].

Otro caso emblemático es el trabajo del Grupo de Procesamiento de Lenguaje Natural de la Comisión de la Verdad [14], que aplicó metodologías avanzadas de NLP al análisis de testimonios y entrevistas relacionadas con el conflicto armado colombiano. Entre las técnicas empleadas se destacan:

- **Análisis de redes semánticas**, que permitió construir grafos de relaciones entre entidades (personas, lugares, organizaciones, acciones). Estas redes fueron evaluadas mediante métricas estructurales como grado, centralidad de intermediación y cercanía, así como los algoritmos de hubs y authorities, con el fin de identificar actores clave y patrones de violencia recurrentes.
- **Análisis de coocurrencias**, que examinó términos frecuentemente mencionados en conjunto dentro de los relatos, revelando asociaciones temáticas como la conexión entre masacres y otras violaciones a los derechos humanos.
- **Modelos de representación semántica**, como *FastText*, que se utilizaron para agrupar términos con significados similares y explorar contextos temáticos complejos, por ejemplo, los relativos a ejecuciones extrajudiciales o desapariciones forzadas.
- **Análisis del lenguaje afectivo**, a través de lexicones emocionales, con el fin de capturar la carga emocional presente en los testimonios. Esto permitió clasificar términos asociados con emociones como miedo, tristeza o ira, proporcionando una dimensión subjetiva del relato que enriqueció la interpretación cualitativa.
- Finalmente, se diseñó un **sistema de recomendación de testimonios**, que, a partir de patrones lingüísticos, sugiere entrevistas similares dentro del corpus, facilitando el análisis comparativo y temático.

Estos avances demuestran el potencial del NLP no solo como herramienta técnica, sino también como instrumento ético y político en procesos de memoria histórica, justicia transicional y reparación simbólica.

## Capítulo 5

### METODOLOGÍA

#### 5.1. Ciclo de vida del aprendizaje automático aplicado al análisis de noticias sobre homicidios de personas LGBTIQ+

El desarrollo de soluciones basadas en aprendizaje automático (*Machine Learning, ML*), requiere seguir un conjunto de etapas bien definidas, conocido como el ciclo de vida del aprendizaje automático (*Machine Learning Lifecycle*), que proporciona una estructura metodológica para garantizar que los modelos sean efectivos, reproducibles y confiables.

A continuación, se describen las principales fases del ciclo [15], junto con su aplicación específica al presente trabajo:

1. **Definición del problema:** se establece el objetivo, las restricciones y los criterios de éxito del modelo.

*Aplicación:* Se definió como problema principal la falta de datos estructurados sobre homicidios, especialmente los que afectan a grupos poblacionales específicos como las personas LGBTIQ+. El objetivo del proyecto es automatizar el proceso de monitoreo de noticias para facilitar el análisis sistemático de estos.

2. **Recolección de datos:** se identifica y obtienen sistemáticamente datos relevantes y representativos.

*Aplicación:* Se seleccionó como fuente principal las noticias encontradas mediante Google Noticias. Se desarrolló un proceso de *web scraping* para recolectar automáticamente la URL y contenido de los artículos.

3. **Preprocesamiento de datos:** se limpian y transforman los datos para hacerlos aptos para el modelo. Esto puede incluir imputación de valores faltantes, normalización, detección de sesgos y codificación de variables.

*Aplicación:* Se realizó la limpieza de texto, tokenización y anotación manual. Se adoptó un esquema de etiquetas basado en el formato BIO.

4. **Selección y entrenamiento del modelo:** se elige el algoritmo más adecuado y se entrena usando datos etiquetados. Durante el entrenamiento se ajustan los parámetros internos para minimizar el error.

*Aplicación:* Se entrenó un modelo de clasificación binaria para filtrar noticias no relevantes, posteriormente, con las noticias relevantes, se entrenó un modelo NER. Para cada uno de los modelos se hizo una selección de hiperparámetros mediante una búsqueda aleatoria.

5. **Evaluación del modelo:** se mide el rendimiento del modelo utilizando métricas como precisión, *recall* y F1-score, comparando sus predicciones con un conjunto de prueba.

*Aplicación:* Se utilizaron métricas estándar derivadas de la matriz de confusión para evaluar tanto el modelo de clasificación como el modelo NER. Se analizaron los resultados por clase (por ejemplo, etiquetas de persona, lugar, arma) y se identificaron oportunidades de mejora, especialmente en entidades con baja frecuencia.

6. **Despliegue y monitoreo:** Si el modelo cumple con los criterios de rendimiento, se implementa como un producto mínimo viable (MVP) y se monitorea su uso.

*Aplicación:* Se desarrolló una aplicación web utilizando *Flask*, que permite a un usuario ingresar criterios de noticias y obtener automáticamente la información estructurada. Esta herramienta representa una solución reproducible y extensible, pensada para organizaciones sociales que monitorean violencia basada en género y orientación sexual.

A continuación, se detalla cada etapa, con los métodos y tecnologías utilizadas.

## 5.2. Definición del problema

El problema se definió mediante la revisión de literatura y revisión de casos de organizaciones que se dedican a visibilizar y denunciar violencia contra población LGBTIQ+ en Colombia.

## 5.3. Recolección de datos

La construcción del corpus de noticias comenzó con la recopilación de URLs y continuó con la extracción del texto de cada una de las noticias

### 5.3.1. Recopilación de URLs

Se utilizó *Selenium*, una herramienta que permite automatizar la interacción con navegadores web, para realizar búsquedas en Google Noticias. Para poder entrenar el modelo de clasificación de noticias, se ejecutaron consultas utilizando palabras claves asociadas a una variedad de temas (ver Anexo 1. Palabras claves para el proceso de webscraping), y así obtener, noticias no relevantes y relevantes. Este proceso da como resultado un Excel con el URL de cada noticia, la palabra clave y rango de fechas utilizadas en la búsqueda, y el estado de cada URL en cuanto a si ya se extrajo el texto correspondiente y si ha sido etiquetado (Figura 5)

Dado que los sistemas de búsqueda como Google pueden detectar comportamientos automatizados y restringir el acceso, se implementaron diversas estrategias para evitar la detección como bot:

1. **Navegación con modo *headless***: Se configuró el navegador en modo *headless*, lo que permite ejecutar el proceso sin interfaz gráfica, reduciendo consumo de recursos, pero manteniendo la funcionalidad completa del navegador.

2. **Rotación de user-agents:** Se incorporó un sistema para alternar *user-agents* que simulan diferentes navegadores y dispositivos, evitando patrones repetitivos que delaten la automatización.
3. **Tiempos de espera aleatorios:** Se programaron intervalos de espera con duraciones aleatorias entre acciones (como clics y desplazamientos), imitando el comportamiento humano.
4. **Manejo robusto de excepciones:** Se implementaron bloques de captura de errores para gestionar caídas del servidor, tiempos de espera excedidos o cambios en la estructura de la página, permitiendo continuar la ejecución sin perder información.
5. **Ejecución paralela con control de carga:** Se utilizó *ThreadPoolExecutor* para procesar múltiples páginas en paralelo, pero limitando el número de hilos activos para no generar un tráfico anormal que pudiera activar bloqueos.
6. **Filtrado por rango de fechas y métricas:** se adaptó el código para permitir búsquedas dentro de intervalos temporales específicos, optimizando la recolección de datos relevantes y evitando solicitudes innecesarias.

keyword	start_date	end_date	url	date_scrape	Scrapped	Labeled
asesinato	01/01/2025	06/14/2025	<a href="https://www.mincultura.gov.c">https://www.mincultura.gov.c</a>	2025-06-14 00:00:00	yes	
asesinato	01/01/2025	06/14/2025	<a href="https://www.cundinamarca.g">https://www.cundinamarca.g</a>	2025-06-14 00:00:00	yes	yes
desaparecieron	01/01/2025	06/14/2025	<a href="https://www.infobae.com/col">https://www.infobae.com/col</a>	2025-06-14 00:00:00	yes	yes
desaparecieron	01/01/2025	06/14/2025	<a href="https://www.elspectador.com">https://www.elspectador.com</a>	2025-06-14 00:00:00	yes	
delitos	01/01/2025	06/14/2025	<a href="https://www.cali.gov.co/bolet">https://www.cali.gov.co/bolet</a>	2025-06-14 00:00:00	yes	yes
delitos	01/01/2025	06/14/2025	<a href="https://www.dw.com/es/2%3">https://www.dw.com/es/2%3</a>	2025-06-14 00:00:00	yes	
delitos	01/01/2025	06/14/2025	<a href="https://nsp.org.co/revista-el-c">https://nsp.org.co/revista-el-c</a>	2025-06-14 00:00:00	yes	
delitos	01/01/2025	06/14/2025	<a href="https://www.procuraduria.gov">https://www.procuraduria.gov</a>	2025-06-14 00:00:00	yes	
delincuentes	01/01/2025	06/14/2025	<a href="https://www.eltiempo.com/cu">https://www.eltiempo.com/cu</a>	2025-06-14 00:00:00	yes	yes
violenta	01/01/2025	06/14/2025	<a href="https://www.infobae.com/col">https://www.infobae.com/col</a>	2025-06-14 00:00:00	yes	
violenta	01/01/2025	06/14/2025	<a href="https://www.eltiempo.com/ps">https://www.eltiempo.com/ps</a>	2025-06-14 00:00:00	yes	
violenta	01/01/2025	06/14/2025	<a href="https://www.pares.com.co/so">https://www.pares.com.co/so</a>	2025-06-14 00:00:00	yes	
violenta	01/01/2025	06/14/2025	<a href="https://caracol.com.co/2025/f">https://caracol.com.co/2025/f</a>	2025-06-14 00:00:00	yes	
violenta	01/01/2025	06/14/2025	<a href="https://www.infobae.com/col">https://www.infobae.com/col</a>	2025-06-14 00:00:00	yes	yes
violenta	01/01/2025	06/14/2025	<a href="https://www.eltiempo.com/co">https://www.eltiempo.com/co</a>	2025-06-14 00:00:00	yes	
violenta	01/01/2025	06/14/2025	<a href="https://www.infobae.com/esp">https://www.infobae.com/esp</a>	2025-06-14 00:00:00	yes	
tráfico de drogas	01/01/2025	06/14/2025	<a href="https://news.un.org/es/story/2">https://news.un.org/es/story/2</a>	2025-06-14 00:00:00	yes	
tráfico de drogas	01/01/2025	06/14/2025	<a href="https://elpais.com/america/20">https://elpais.com/america/20</a>	2025-06-14 00:00:00	yes	
tráfico de drogas	01/01/2025	06/14/2025	<a href="https://www.seguritatam.com">https://www.seguritatam.com</a>	2025-06-14 00:00:00	yes	
identidad de género	01/01/2025	06/14/2025	<a href="https://www.ambitojuridico.c">https://www.ambitojuridico.c</a>	2025-06-14 00:00:00	yes	yes
identidad de género	01/01/2025	06/14/2025	<a href="https://www.hrw.org/es/news">https://www.hrw.org/es/news</a>	2025-06-14 00:00:00	yes	
homosexual	01/01/2025	06/14/2025	<a href="https://www.senalmemoria.co">https://www.senalmemoria.co</a>	2025-06-14 00:00:00	yes	
homosexual	01/01/2025	06/14/2025	<a href="https://www.infobae.com/esp">https://www.infobae.com/esp</a>	2025-06-14 00:00:00	yes	
homosexual	01/01/2025	06/14/2025	<a href="https://flipro.org/es/quienes-s">https://flipro.org/es/quienes-s</a>	2025-06-14 00:00:00	yes	
homosexual	01/01/2025	06/14/2025	<a href="https://www.elspectador.com">https://www.elspectador.com</a>	2025-06-14 00:00:00	yes	
farandula	01/01/2025	06/14/2025	<a href="https://www.diariolibre.com/f">https://www.diariolibre.com/f</a>	2025-06-14 00:00:00	yes	
farandula	01/01/2025	06/14/2025	<a href="https://elcomercio.pe/vmas/f">https://elcomercio.pe/vmas/f</a>	2025-06-14 00:00:00	yes	yes
espectáculos	01/01/2025	06/14/2025	<a href="https://www.timeout.es/madr">https://www.timeout.es/madr</a>	2025-06-14 00:00:00	yes	
espectáculos	01/01/2025	06/14/2025	<a href="https://www.cocemfe.es/infor">https://www.cocemfe.es/infor</a>	2025-06-14 00:00:00	yes	
celebridades	01/01/2025	06/14/2025	<a href="https://es-us.vida-estilo.yaho">https://es-us.vida-estilo.yaho</a>	2025-06-14 00:00:00	yes	
política	01/01/2025	06/14/2025	<a href="https://www.paho.org/es/noti">https://www.paho.org/es/noti</a>	2025-06-14 00:00:00	yes	yes

Figura 5. Ejemplo excel de URLs recolectadas

Fuente: elaboración propia



En el caso de este estudio, el anotador debía:

1. **Leer el contenido completo** de cada noticia para comprender el contexto.
2. **Decidir si el texto correspondía o no a un caso de homicidio.**
3. **Identificar y resaltar manualmente en el texto** las entidades nombradas relevantes. Las categorías de entidades fueron definidas previamente (ver Anexo 2. Palabras NER) e incluyeron variables como: nombres de víctimas y perpetradores, ubicación del crimen, fecha, motivo del homicidio, entre otras (Figura 7).

Este proceso es fundamental para construir un corpus de alta calidad que sirva como base para el entrenamiento de modelos automáticos de extracción de información.

Minuto30.com - En un nuevo caso de **violencia de género** que conmociona a la ciudad, las autoridades reportaron el **feminicidio** de una **mujer** de **75 años** en el exclusivo **barrio El Poblado**. El hecho ocurrió en la tarde del (día) en la **urbanización Vegas de Zúñiga, carrera 48 No. 17A-51 Sur**. La víctima, identificada como **Miriam del Carmen J.A.**, fue encontrada sin vida **al interior de su residencia**, presentando **signos de violencia causados por un objeto contundente**. Según las primeras versiones, el presunto responsable del crimen sería la **pareja sentimental de la mujer**, un **hombre** de **aproximadamente 80 años**, quien se entregó de manera voluntaria en el CAI Caribe, confesando el homicidio. Las autoridades investigan los móviles que llevaron al **hombre** a cometer este atroz acto. Unidades de la Sijin y la Sipol se hicieron presentes en el lugar de los hechos para realizar la inspección técnica del cadáver y adelantar las investigaciones correspondientes. Los vecinos de la zona se mostraron consternados por este nuevo caso de **feminicidio** que enluta a la ciudad. Las autoridades hacen un llamado a la comunidad para denunciar cualquier caso de violencia de género. Es fundamental romper el silencio y actuar de manera conjunta para erradicar este flagelo que sigue cobrando vidas en nuestro país. Aquí más Noticias de Medellín

Legend:

- Victim First Name
- Victim Last Name
- Victim Age
- Victim Profesion
- Victim Gender
- Victim Nationality
- Victim Criminal background
- Perpetrator First Name
- Perpetrator Last Name
- Perpetrator Alias
- Perpetrator Gender
- Perpetrator Age
- Perpetrator Profesion
- Banda Criminal
- Sensacionalismo
- Weapon
- Relationship
- Motive
- Number of Victims
- Number of Perpetrators
- Date
- Municipality
- Department
- Place

Figura 7. Ejemplo de un texto etiquetado en Label Studio

Fuente: elaboración propia

Una vez se etiquetó un grupo de noticias, se precedió a exportar de *Label Studio*, un archivo CSV con diversas columnas (Figura 8), entre las cuales están:

- **ner\_labels**: la cual presenta para cada una de las NER presente en el artículo, la posición de inicio y final, el texto etiquetado y la etiqueta correspondiente.
- **tipo\_noticia**: hace referencia a si la noticia es sobre un homicidio o no
- **texto**: el texto de la noticia.





Con las noticias clasificadas como homicidios, se procedió al entrenamiento del modelo BETO de Hugging Face<sup>2</sup>, específicamente su versión adaptada para clasificación de tokens mediante la clase *AutoModelForTokenClassification*. Para la selección de los hiperparámetros del modelo, se utilizó la técnica de búsqueda aleatoria (*random search*). Esta consiste en la evaluación de un subconjunto aleatorio de combinaciones de hiperparámetros dentro de un espacio de búsqueda definido. A diferencia de la búsqueda en cuadrícula (*grid search*), que prueba todas las combinaciones posibles, la búsqueda aleatoria es más eficiente y menos exigente computacionalmente ya que selecciona combinaciones de forma aleatoria. El espacio de búsqueda fue el siguiente:

- **learning\_rate**: 5e-5, 3e-5, 2e-5
- **batch\_size (train/eval)**: 16 y 32
- **epochs**: 2, 3, 4
- **weight\_decay**: 0.0 y 0.01

Se seleccionó la configuración que minimizó la pérdida en validación (*eval\_loss*). El modelo resultante se guardó y se utilizó en etapas posteriores para la inferencia sobre noticias no anotadas.

### 5.5.3. Aprendizaje Activo

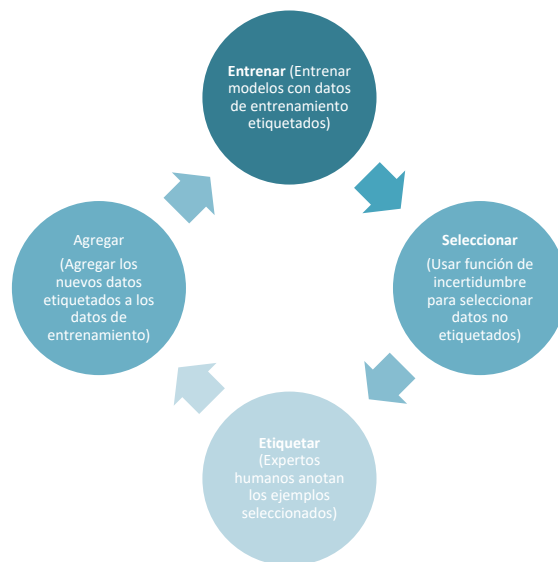
Después del entrenamiento inicial del modelo, y considerando que el etiquetado manual implica un alto costo en tiempo y recursos humanos, se implementó una estrategia de aprendizaje activo (*Active Learning*) para optimizar el proceso de anotación. Esta metodología permite al modelo identificar y priorizar aquellas noticias cuya clasificación resulta más incierta o compleja, de modo que el esfuerzo de etiquetado humano se concentre en los casos con mayor valor informativo para mejorar el desempeño del modelo [16].

---

<sup>2</sup> El modelo utilizado fue: dccuchile/bert-base-spanish-wwm-cased

En particular, se implementaron dos criterios de incertidumbre para priorizar las noticias a etiquetar manualmente: **entropía** y **mínima confianza**. La **entropía** permitió identificar textos en los que el modelo distribuía de manera más uniforme las probabilidades entre varias clases, reflejando **mayor ambigüedad** en sus predicciones. Por su parte, la **mínima confianza** se centró en la probabilidad más baja asignada a la clase predicha, resaltando los casos en los que el modelo mostraba **menor seguridad** respecto a una etiqueta específica.

Este enfoque resultó especialmente útil en el corpus periodístico sobre homicidios en Colombia, donde es frecuente encontrar textos con múltiples víctimas y victimarios o narrativas complejas que dificultan la extracción automática de información. Al priorizar estas instancias para su anotación manual, el modelo pudo refinar progresivamente su capacidad de clasificación con un número reducido de ejemplos adicionales. La Figura 10 presenta el funcionamiento del bucle de aprendizaje activo.



*Figura 10. Bucle de Aprendizaje Activo*

*Fuente: elaboración propia*

## 5.6. Evaluación del modelo

Ambos modelos se evaluaron utilizando el módulo de métricas de *scikit-learn*. Particularmente, se calculó la precisión, el *recall*, el puntaje F1 y la matriz de confusión. Estas métricas se seleccionaron porque permiten una evaluación integral del desempeño del modelo en tareas de clasificación, especialmente cuando pueden existir clases desbalanceadas o cuando los errores tienen diferentes implicaciones:

Estas métricas fueron preferidas sobre otras como la exactitud (*accuracy*) porque, en contextos con clases desbalanceadas, la exactitud puede ser engañosa: un modelo que siempre predice la clase mayoritaria puede tener una alta exactitud pero bajo desempeño real en la clase minoritaria.

## 5.7. Despliegue y monitoreo

Se desarrolló una aplicación web utilizando el framework *Flask* en Python. Esta aplicación consiste en una interfaz donde los usuarios deben ingresar un rango de fechas para posteriormente obtener una tabla de datos estructurada sobre noticias de homicidios.

El backend de la aplicación carga el modelo de clasificación y el modelo NER previamente ajustados. Una vez se recibe el rango de fechas, el sistema ejecuta los siguientes pasos:

1. **Extracción de contenido:** Se utiliza la biblioteca *newspaper3k* para descargar y extraer el texto de las noticias.
2. **Preprocesamiento:** Los textos extraídos se limpian y tokenizan para poder ser procesados por el modelo de clasificación.
3. **Clasificación:** El modelo de clasificación selecciona únicamente las noticias que reportan casos de homicidio.
4. **Reconocimiento de Entidades (NER):** El modelo NER identifica las variables de interés en cada noticia.
5. **Descarga:** Se genera un archivo descargable con el texto de cada noticia y las variables identificadas.

## Capítulo 6

### RESULTADOS Y DISCUSIÓN

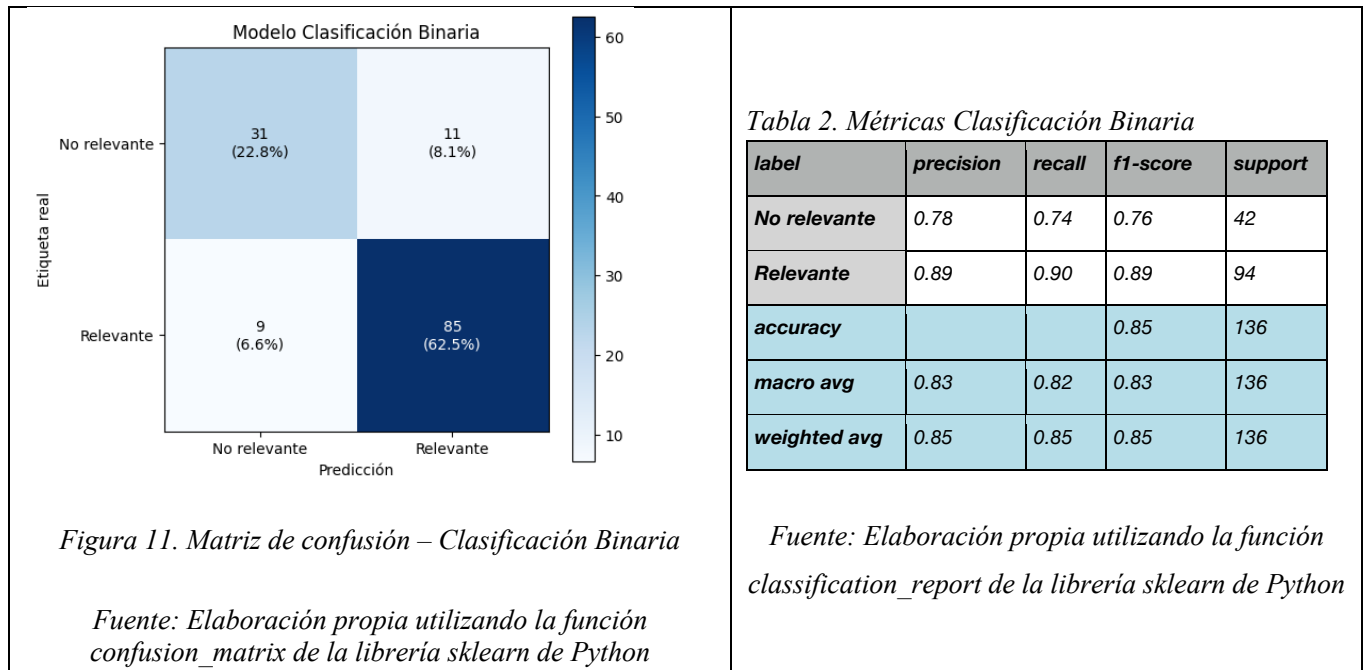
Este capítulo presenta los principales resultados obtenidos del desarrollo del proyecto, destacando el desempeño del modelo de clasificación binaria y NER, así como la funcionalidad de la aplicación web desarrollada. Se analiza la coherencia de los resultados con los objetivos planteados y se discute su relevancia técnica y social.

#### 6.1 Modelo de Clasificación Binaria

Como parte de la solución planteada, se desarrolló un modelo de clasificación binaria para filtrar las noticias que contienen información sobre casos de homicidios. Este modelo constituye un primer paso esencial en el pipeline de procesamiento, pues permite enfocar los recursos de cómputo en textos que efectivamente contienen variables clave para el análisis.

La evaluación del modelo de clasificación binaria se realizó sobre un conjunto de prueba compuesto por 136 noticias, alcanzando un *accuracy* del 85 %, es decir, en más de 8 de cada 10 casos, el modelo predijó correctamente si una noticia es relevante o no para el análisis de homicidios. El modelo mostró un mejor desempeño en la clase relevante, alcanzando una F1-score de 0.89, con un *recall* de 90 %, lo cual indica que casi todas las noticias relevantes fueron correctamente identificadas. Por el contrario, el rendimiento fue menor en la clase no relevante, con una F1-score de 0.76. Este resultado muestra que el modelo priorizó la identificación de noticias relevantes, lo cual es adecuado para el propósito del sistema, ya que el costo de omitir una noticia verdaderamente relevante (falso negativo) es mayor que el de incluir una noticia irrelevante (falso positivo) en el pipeline de procesamiento.

En general, el modelo cumple adecuadamente su función como filtro previo, maximizando la retención de noticias útiles para la tarea de extracción de entidades. La Figura 11 muestra la matriz de confusión, la Tabla 2 presenta las métricas estándar de desempeño del modelo.



*Fuente: elaboración propia*

## 6.2. Modelo NER

El modelo NER fue entrenado sobre un corpus de noticias utilizando el esquema BIO, el cual permite identificar de manera estructurada el inicio y la continuación de las entidades. Se definieron 24 etiquetas que representan variables clave para el análisis de noticias sobre homicidios. Tras el proceso de *fine-tuning* sobre el modelo base, se obtuvieron mejoras significativas en todas las métricas de evaluación (ver Tabla 3).

- **F1-score global:** 0.53
- **Precisión:** 0.48
- **Recall:** 0.60
- **Accuracy:** 94.8 %

Parámetro	Sin entrenar	Entrenado
eval_loss	3.744142770767212	0.4505883753299713
eval_model_preparation_time	0.0016	0.0021
eval_accuracy	0.02449295956594755	0.9481720707918874
eval_f1_	0.00019043387183800424	0.5346534653465347
eval_precision	9.773578758755497e-05	0.4831013916500994
eval_recall	0.003694581280788177	0.5985221674876847
eval_runtime	12.6064	8.0047
eval_samples_per_second	85.988	135.421
eval_steps_per_second	10.788	16.99

Tabla 3. Métricas del modelo NER antes y después del entrenamiento

Fuente: elaboración propia

Estas métricas reflejan una mejora notable en la capacidad del modelo para identificar entidades relevantes dentro del texto, en comparación con el rendimiento inicial del modelo base sin entrenamiento específico, cuyo F1-score, precisión y recall eran prácticamente nulos. El *accuracy* del 94,8 % debe ser tomado con cautela, puesto que gran parte de esos aciertos se deben a clasificar tokens no pertenecientes a ninguna entidad. El F1-score global de 0.53, junto con una precisión de 0.48 y un recall de 0.60, indica que el modelo tiene una capacidad razonable para identificar entidades relevantes, aunque aún presenta errores tanto por omisión como por clasificación incorrecta.

La Tabla 4 presenta las métricas de evaluación para cada una de las variables. A nivel de NER individualizados, se destaca el buen desempeño en entidades como nombres y apellidos de la víctima, fecha del hecho, municipio, departamento y edad, con F1-scores superiores al 0.70. Estas entidades son particularmente relevantes para construir bases de datos estructuradas que permitan análisis espaciales, temporales y sociodemográficos de los homicidios.

No obstante, se observaron resultados limitados en entidades menos frecuentes (como "banda criminal" o "antecedentes penales"), lo cual se atribuye a su escasa representación en los datos de entrenamiento. Este hallazgo valida la metodología empleada, basada en

aprendizaje supervisado, pero también señala oportunidades de mejora mediante la ampliación y diversificación del corpus de noticias. Adicionalmente, estos resultados reflejan una limitación inherente a las propias fuentes noticiosas, que en muchos casos no reportan información detallada sobre estos aspectos, lo que evidencia tanto la dificultad de acceso a ciertos datos como los vacíos informativos presentes en los relatos mediáticos sobre los homicidios.

Destaca el caso de la categoría “motivo”, la cual aparece con frecuencia en las noticias, pero cuya identificación resulta especialmente compleja para el modelo. Una posible explicación es la amplia variabilidad lingüística con la que se describen los motivos en los textos periodísticos. Estos pueden expresarse mediante frases explícitas, por ejemplo, “El homicidio ocurrió por motivos pasionales” o “La causa del crimen fue un ajuste de cuentas”, o a través de formas indirectas como “La víctima había denunciado amenazas de grupos armados semanas antes”. En algunos casos, incluso se emplea lenguaje narrativo, como “Los celos se convirtieron en la chispa que desencadenó la tragedia”. Esta diversidad de expresiones incrementa la dificultad para que el modelo generalice y reconozca correctamente todas las variantes presentes en el corpus.

label	precision	recall	f1-score	support
<b>Banda Criminal</b>	0.08	0.17	0.11	6
<b>Fecha</b>	0.53	0.62	0.57	55
<b>Departamento</b>	0.65	0.77	0.71	31
<b>Motivo</b>	0.13	0.18	0.15	39
<b>Municipio</b>	0.65	0.78	0.71	64
<b>Número de víctimas</b>	0.18	0.33	0.24	6
<b>Edad del victimario</b>	0.33	0.5	0.4	8
<b>Alias del victimario</b>	0.23	0.36	0.28	14
<b>Nombre del victimario</b>	0.6	0.7	0.65	30
<b>Género del victimario</b>	0.25	0.09	0.13	11
<b>Apellido del victimario</b>	0.68	0.7	0.69	43
<b>Profesión del victimario</b>	0.11	0.07	0.08	15
<b>Lugar</b>	0.35	0.47	0.4	92

<b>Relación</b>	0.05	0.1	0.07	10
<b>Sensacionalismo</b>	0.0	0.0	0.0	5
<b>Edad de la víctima</b>	0.47	0.63	0.54	46
<b>Antecedentes criminales de la víctima</b>	0.0	0.0	0.0	1
<b>Nombre de la víctima</b>	0.72	0.82	0.76	109
<b>Sexo de la víctima</b>	0.39	0.44	0.41	27
<b>Apellido de la víctima</b>	0.71	0.85	0.77	99
<b>Nacionalidad de la víctima</b>	0.0	0.0	0.0	2
<b>Profesión de la víctima</b>	0.41	0.61	0.49	46
<b>Arma</b>	0.27	0.38	0.31	53
<b>micro avg</b>	0.48	0.6	0.53	812
<b>macro avg</b>	0.34	0.42	0.37	812
<b>weighted avg</b>	0.49	0.6	0.54	812

*Tabla 4. Métricas del modelo NER por variables después del entrenamiento*

*Fuente: elaboración propia*

La matriz de confusión (Figura 12) evidencia confusiones en variables compartidas por víctima y victimario, como edad, nombre, apellido y profesión. En otras palabras, aunque el modelo logra identificar estos atributos en el texto, presenta dificultades para diferenciar si corresponden a la víctima o al victimario, especialmente cuando ambos se mencionan en la misma noticia con estructuras lingüísticas similares.



### 6.3 Aplicación Web: Herramienta de Sistematización Automatizada

Como producto final, se implementó una **aplicación web funcional**, que representa un **producto viable mínimo (MVP)**, es decir, una versión inicial, operativa y simplificada del sistema, capaz de cumplir con los requerimientos básicos del proyecto. Actualmente, esta herramienta se ejecuta correctamente en entorno local, permitiendo evaluar su funcionamiento completo de forma reproducible. La aplicación está diseñada para que cualquier usuario pueda ingresar un rango de fecha y obtener como salida una tabla con las entidades extraídas de las noticias encontradas. Esta herramienta opera en cuatro etapas principales:

1. **Extracción automatizada del contenido:** utilizando la librería *newspaper3k* junto con funciones personalizadas, se descarga y limpia el texto completo de los artículos.
2. **Clasificación binaria del texto:** se evalúa si cada noticia contiene información relevante sobre un caso de homicidio, evitando aplicar el modelo NER innecesariamente.
3. **Reconocimiento de entidades:** se ejecuta el modelo NER entrenado sobre los textos para extraer variables clave como nombres, fechas, lugares, género, edad, motivo, entre otros.
4. **Visualización estructurada:** los resultados se presentan en una tabla organizada, que puede ser exportada y utilizada directamente en procesos de análisis o sistematización (Figura 13).

Esta aplicación representa una solución reproducible y escalable, pensada para facilitar la labor de organizaciones sociales, investigadoras o periodistas de derechos humanos. Si bien el sistema aún no está desplegado en línea, su funcionamiento en entorno local demuestra que es tecnológicamente viable y que puede ser trasladado a producción con adaptaciones,

como la incorporación de una interfaz gráfica más amigable o su publicación en un servidor institucional.

id	data	Date	Perpetrator Age	Victim Age	Department	Perpetrator First Name	Perpetrator Last Name	Place	Perpetrator Alias
121	['Título: 'Bogotá: capturado en Usme uno de los delincuentes más bus'	['22 de abril del año 2023']	['24 años', '24 años']	['31 años']	['Tolima']	['Edgar Andrés']	['Prieto Santos']	['bar']	
122	['Título: 'Capturan a uno de los más buscados por homicidio en Barran'	['10 de junio']	['28 años']			['Luis Fernando']	['Lamadín Ochoa']	['barrio Ciudadela 20 de']	['alias "Torta"', 'Alias "Torta
123	['Título: 'Lo que dijo la Policía de accidente en el que murió niño Juan I'								
124	['Título: 'Las alarmantes cifras de Medicina Legal sobre la violencia inl'			['12 años']	['vallecaucano']				['corregimiento de Villa
125	['Título: 'Cae a alias Coronel, señalado cabeclla del 'clan del Golfo: h'	['21 de abril del presente año']			['Bolívar']	['José Alejandro']	['Castro Cadavid']		['Coronel o Chiquí', 'Chiquí o
126	['Título: 'Un exparamilitar condenado fue asesinado a tiros en Medell'	['este sábado 12 de octubre']		['51 años']					['sector de Pedregal']
127	['Título: 'Penito de subteniente de la Policía asesinado por un grupo lí'	['lunes, 21 de octubre']							
128	['Título: 'Bogotá: patrullero de la Policía fue asesinado por joven de 17'		['17 años', '17 años', '17 a']	['36 años']					['sector de Las Delicias']
129	['Título: 'El último y triste adiós a Sofía Delgado, la niña asesinada por'	['29 de septiembre', '29 de septie']	['32 años']	['12 años', '12 años']	['Valle del Cauca']	['Brayan']	['Campo']		['barrio La Victoria', 'Vil
130	['Título: 'Herida de metralla y disparo en la cabeza: lo que reveló la au'								
131	['Título: 'Tanques israelíes rodean y disparan contra el hospital Indone'								
132	['Título: 'Dos tragedias y el mismo dolor en hogares de Cali y Candelar'		['32 años']	['15 años', '15 años', '12 años', '1']			['Hárold', 'Hárold', 'Hárold']	['Echeverry', 'Echeverry', 'Ec']	['barrio San Judas', 'boc
133	['Título: 'Multitud, desmayados y caos en el sepelio de Sofía Delgado,'				['Valle del Cauca']	['Brayan']	['Campo Píllimue']		
134	['Título: 'Asesinan a una mujer trans dentro de su apartamento en Cal'			['61 años']	['Valle de Aburrá']				['barrio La Mansión']
135	['Título: 'Capturan en hospital a pareja de mujer asesinada en Repelló'	['15 de octubre']		['33 años']	['Atlántico']	['Alfonso']	['Ramírez Gómez', 'Ramíre']		['barrio El Porvenir']
136	['Título: 'Una joven asesinó a su novio en plena calle porque salió a c'								
137	['Título: 'Comoción en Brasil: adolescente de 14 años asesinó a tíos'								
138	['Título: 'Condenan al policía que asesinó a un joven por colarse en uni'	['6 de enero de 2017']							
139	['Título: 'Brayan Campo, confeso asesino de Sofía Delgado, ya enfrenta'	['domingo 29 de septiembre']		['12 años']	['Valle del Cauca']	['Brayan', 'Brayan']	['Campo', 'Ocampo']		['corregimiento Villago
140	['Título: 'Fotos: Él es Brayan Campo, señalado asesino de Sofía Delga'	['29 de septiembre']	['32', '32 años']	['12 años', '32 años']	['vallecaucano']	['Brayan', 'Brayan']	['Campo', 'Campo']		['corregimiento de Villa
141	['Título: 'El macabro plan de una mujer para robarse a un bebé: cíclic'								
142	['Título: 'Medellín: Un escolta asesinó a un presunto felero en medio'	['lunes 21 de octubre']		['20 y 25 años']					['barrios Miravalle y el i
143	['Título: 'Antioquia: hombre que asesinó a una perrita en El Carmen de'								
144	['Título: 'Atención: Brayan Campo aceptó ante un juez que asesinó a la'	['29 de septiembre']			['Valle del Cauca']	['Brayan', 'Brayan']	['Campo', 'Campo']		
145	['Título: 'Propietaria de la casa donde vivía Brayan Campo, el confeso'	['29 de septiembre']	['31 años']	['12 años']			['Brayan Snaither']	['Campo Píllimue']	
146	['Título: '¿Quién mató al papa de Michael Jordan? Justicia de un giro lí'								
147	['Título: 'Israel confirma que mató a Yahya Sinwar, máximo líder de H'								
148	['Título: 'En Tolima: tres personas fueron capturadas por presuntamen'	['31 de marzo de 2023']	['25 años']	['32 años', '81 años', '50']			['Edgar', 'Juan']	['Prieto', 'Villanueva']	['finca', 'corregimiento ('Yumi']
149	['Título: 'Por los delitos de homicidio y tortura en el sur de Bogotá, do'	['23 de septiembre']		['25 años']			['Germán Lisandro', 'Kevin A']	['Muñoz Silva', 'Romero Jari']	['Patio Bonito', 'barberí']
150	['Título: 'Tragedia familiar en La Guajira: ginecólogo habría sido asesi'	['domingo 13 de octubre']	['37 años']	['42 años']	['La Guajira']	['Nelson José']	['Solano Brochero']		['frente de una residenc
151	['Título: 'Eran como hermanos': habla esposa del ginecólogo que fue'	['domingo 13 de octubre']	['37 años']	['42 años']	['La Guajira']	['Nelson José']	['Solano Brochero']		['frente de una residenc
152	['Título: '¿Qué se sabe del caso del profesor asesinado frente a sus es'								
153	['Título: 'Profesor fue asesinado en la puerta del colegio por un hombr'								
154	['Título: 'Los reclamos del Gobierno Nacional y el Concejo de Bogotá a'								
155	['Título: 'Radiografía del impacto de la intolerancia en los homicidios'								
156	['Título: 'Arrestan a una mujer por presuntamente haber asesinado a s'								

Figura 13. Ejemplo tabla de datos resultado

Fuente: elaboración propia

### 6.4 Discusión General

Los resultados demuestran que es posible automatizar, de forma efectiva, una parte significativa del proceso de monitoreo de prensa de homicidios en Colombia. La herramienta desarrollada:

- Responde directamente a la problemática social identificada, al permitir visibilizar violencias que han sido históricamente invisibilizadas.
- Materializa los objetivos específicos del proyecto, desde la identificación de variables, la recolección de datos, el entrenamiento de modelos, hasta el despliegue de una solución funcional.
- Se construyó con base en una metodología sólida, utilizando el ciclo de vida del aprendizaje automático y arquitectura de inteligencia artificial demostrada ser exitosa en procesamiento de datos secuenciales como *transformers*.

- Genera un producto útil para actores sociales, como ONG y activistas, permitiéndoles reducir drásticamente el tiempo invertido en la sistematización manual.

En conjunto, se cumple no solo con el propósito técnico de extraer variables estructuradas, sino también con una finalidad ética y política: nombrar, clasificar y visibilizar las violencias que afectan a las personas LGBTIQ+ en Colombia.

## Capítulo 7

### CONCLUSIONES Y RECOMENDACIONES

Este trabajo demuestra que es posible aplicar técnicas avanzadas de NLP para abordar problemas sociales complejos, como la sistematización de información sobre homicidios de personas LGBTIQ+ en Colombia. La herramienta desarrollada marca un hito en el análisis automatizado de noticias periodísticas, al permitir extraer información clave de manera estructurada y reducir significativamente el tiempo requerido para estas tareas, tradicionalmente realizadas de forma manual por organizaciones sociales.

Entre los principales hallazgos y aportes del proyecto se destacan:

1. **Automatización del análisis de noticias:** La aplicación desarrollada permite identificar automáticamente variables relevantes como el nombre de la víctima, género, edad, lugar, fecha, arma utilizada y posible motivo del crimen. Esto representa un avance significativo en comparación con los métodos manuales usados actualmente.
2. **Impacto social y ético:** La herramienta contribuye a visibilizar las violencias que afectan a poblaciones históricamente invisibilizadas, como las personas LGBTIQ+. Al facilitar la creación de bases de datos estructuradas, se fortalece la capacidad de análisis sistemático y, por tanto, la incidencia y exigibilidad de derechos de estas comunidades.
3. **Desempeño técnico sólido:** El modelo de clasificación binaria logró un *accuracy* del 85 %, lo cual garantiza que la mayoría de las noticias pertinentes sean correctamente identificadas. El modelo NER, tras el *fine-tuning*, mostró mejoras sustanciales en múltiples entidades clave, como nombres y apellidos de la víctima, fecha del hecho, municipio, departamento y edad, con F1-scores superiores al 0.70

4. **Flexibilidad y escalabilidad:** La solución es modular, lo que permite adaptar el pipeline a otros contextos sociales o geográficos. Además, el enfoque de aprendizaje activo facilita su mejora continua a medida que se dispone de nuevos datos anotados.

La herramienta pretende ser una fuente de información complementaria a la oficial, de manera que organizaciones sociales puedan:

- **Monitorear en tiempo real** la ocurrencia de homicidios, incluyendo variables clave como lugar, fecha, víctimas, victimarios y posibles motivos.
- **Detectar patrones y tendencias** que permitan identificar zonas con mayor incidencia, tipos de violencia predominantes o factores asociados a cada caso.
- **Optimizar recursos y focalizar intervenciones**, dirigiendo programas de prevención y seguridad hacia poblaciones y territorios más vulnerables.
- **Mejorar la transparencia y la rendición de cuentas**, al disponer de datos sistematizados y verificables para la formulación de políticas públicas.
- **Alimentar sistemas de alertas tempranas** que faciliten la toma de decisiones rápida en contextos de violencia o crisis de seguridad.

En conjunto, el impacto social radica en que la información generada no solo fortalece la capacidad de análisis de las instituciones, sino que también contribuye a diseñar políticas públicas basadas en evidencia, con un enfoque preventivo y de protección a las víctimas.

### **Recomendaciones**

Si se desea escalar el modelo y ponerlo a funcionar, se recomienda:

- **Despliegue colaborativo:** Se recomienda explorar alianzas con distintos tipos de organizaciones, para adaptar y utilizar el sistema en sus procesos de monitoreo de derechos humanos y de acuerdo con sus variables de interés.

- **Mejoras en la cobertura del modelo NER:** Aunque el modelo mostró buen desempeño en varias entidades, algunas categorías con baja frecuencia requieren más ejemplos anotados para mejorar su precisión. Se sugiere continuar con el proceso de etiquetado iterativo para fortalecer la cobertura.
- **Ampliación del universo temático:** Una de las ventajas de la arquitectura utilizada es su flexibilidad para ser reentrenada con nuevos conjuntos de datos y etiquetas. Se recomienda evaluar la incorporación de nuevas variables de análisis que permitan adaptar el modelo a otros contextos de interés social, tales como amenazas, violencia policial, afectación por desastres naturales, desplazamiento forzado o ataques contra líderes sociales. Esto aumentaría el potencial del sistema como herramienta transversal de monitoreo de derechos humanos y acción humanitaria.
- **Interfaz más accesible:** Finalmente, para facilitar el uso por parte de personas no técnicas, se recomienda el desarrollo de una interfaz gráfica más amigable e inclusión de modelo de análisis preliminar de noticias.

Este proyecto demuestra que los modelos de NLP pueden ser herramientas éticas y potentes al servicio de la justicia social, al transformar grandes volúmenes de texto en evidencia estructurada y útil para la defensa de derechos humanos. Dicha herramienta

### **Consideraciones éticas**

Durante el desarrollo de este trabajo, se tuvieron en cuenta consideraciones éticas relacionadas con el manejo de información sensible y el respeto a las víctimas. Aunque se basa en noticias de acceso público, se reconoce que los textos analizados hacen referencia a hechos homicidas, muchos motivados por prejuicio, lo cual exige un tratamiento cuidadoso que evite la revictimización y el sensacionalismo. El análisis se enfocó en los patrones lingüísticos y narrativos de los medios, sin emitir juicios de valor sobre las personas mencionadas ni sobre las instituciones que difunden la información.

Una de las principales preocupaciones éticas radica en los sesgos que pueden reproducir los modelos de inteligencia artificial entrenados con noticias [17]. La cobertura mediática no es neutral: suele incluir estereotipos de género, estigmatizaciones territoriales, diferencias en la forma de narrar los hechos según el estatus socioeconómico de las víctimas o la gravedad atribuida a distintos tipos de violencia. Estos sesgos, al trasladarse al modelo, pueden generar representaciones distorsionadas que perpetúen desigualdades y narrativas discriminatorias. Por ejemplo, el sistema podría aprender a asociar ciertos motivos del crimen con características de género o región, no por un patrón real de la violencia, sino por la forma en que los medios lo reportan.

Para mitigar estos riesgos, se buscó garantizar transparencia en las decisiones metodológicas y documentar de manera clara la construcción del corpus, el proceso de anotación y las limitaciones del modelo. Asimismo, se enfatiza que los resultados no deben interpretarse como verdades objetivas ni usarse de forma automática en contextos judiciales o de política pública sin un análisis crítico. En cambio, se propone que el modelo sirva como una herramienta exploratoria que complemente el análisis humano e información oficial, siempre reconociendo sus limitaciones y la posibilidad de sesgos en la información subyacente.

## REFERENCIAS

- [1] Colombia Diversa, «La realidad de la discriminación,» Colombia Diversa, Bogotá, 2024.
- [2] Instituto Nacional de Medicina Legal y Ciencias Forenses, «Forensis 2023 Datos para la vida,» 2024.
- [3] Coalición contra la vinculación de niños, niñas y jóvenes al conflicto armado en Colombia (COALICO), «Boletín de Monitoreo No. 31 del Observatorio de Niñez y Conflicto Armado de la COALICO - ONCA,» 2024.
- [4] Colombia Diversa, «La violencia no nos impide ser y amar Informe Situación de Derechos Humanos Personas LGBT 2021,» 2022.
- [5] Caribe Afirmativo, «Con permiso para despreciar. Informe sobre la situación de derechos humanos de personas LGBTIQ+ en Colombia en 2024.,» 2025.
- [6] Y. Hou y J. Huang, «Natural language processing for social science research: A comprehensive review,» *Chinese Journal of Sociology*, vol. 11, n° 1, pp. 121-157, 2025.
- [7] K. R. Chowdhary, *Fundamentals of Artificial Intelligence*, Jodhpur, Rajasthan, India: Springer, 2020.
- [8] D. Jurafsky y J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 2025.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser y I. Polosukhin, «Attention Is All You Need,» *Advances in neural information processing systems*, vol. 30, 2017.
- [10] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,» *arXiv*, 2018.
- [11] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang y J. Pérez, «Spanish pre-trained BERT model and evaluation data,» *arXiv preprint arXiv*, 2023.
- [12] J. Bergstra y Y. Bengio, «Random Search for Hyper-Parameter Optimization,» *Journal of Machine Learning Research*, vol. 13, pp. 281-305, 2012.
- [13] J. C. Upegui y V. Saavedra, «PretorIA y la automatización del procesamiento de causas de derechos humanos,» *Dejusticia*, 2021.
- [14] Comisión de la Verdad, «Metodologías de análisis de entrevistas utilizando Procesamiento de Lenguaje Natural (PLN),» Comisión de la Verdad, Bogotá, 2022.
- [15] K. R. Varshney, *Trustworthy Machine Learning*, Chappaqua, NY: Independently Published, 2022.
- [16] F. Stollenwerk, J. Öhman, D. Petrelli, E. Wallerö, F. Olsson, C. Bengtsson, A. Horndahl y G. Zarzar Gandler, *Text Annotation Handbook A Practical Guide for Machine Learning Projects*.

- [17] X. Fang, S. Che, M. Mao, H. Zhang, M. Zhao y X. Zhao, «Bias of AI-generated content: an examination of news produced by large language models,» *Scientific Reports*, vol. 14, n° 5224, 2024.
- [18] Defensoría del Pueblo Colombia, «Una Radiografía del Prejuicio Boletín Anual 2023 de Derechos Humanos de Personas OSIGD-LGBTI,» Defensoría del Pueblo, Bogotá, 2024.

## ANEXOS

### Anexo 1. Palabras claves para el proceso de webscrapping

Se utilizaron tres categorías de palabras claves.

#### 1. Sin relación a homicidios o población LGBTIQ+

“farándula”, “espectáculos”, “entretenimiento”, “celebridades”, “política”, “deportes”, “fútbol”, “cultura”, “arte”, “música”, “cine”, “teatro”, “televisión”, “noticias”, “economía”, “tecnología”, “ciencia”, “salud”, “medio ambiente”, “educación”, “historia”, “filosofía”, “literatura”, “viajes”, “gastronomía”, “moda”, “belleza”, “estilo de vida”, “bienestar”, “psicología”, “religión”, “espiritualidad”, “sociedad”, “comunidad”, “derechos humanos”, “justicia”, “seguridad”, “crimen”, “policía”, “investigación”, “sucesos”, “accidentes”, “desastres naturales”, “clima”, “cambio climático”, “salud pública”, “epidemias”, “pandemias”, “vacunas”

#### 2. Relacionadas a homicidios y violencia

“homicidio”, “asesinado”, “asesinada”, “asesinó”, “mató”, “ultimaron”, “asesinato”, “asesino”, “asesina”, “mataron”, “muerte”, “muerto”, “muerta”, “fallecido”, “fallecida”, “falleció”, “fallecieron”, “desaparecido”, “desaparecida”, “desapareció”, “desaparecieron”, “desaparecer”, “delito”, “delitos”, “delincuente”, “delincuentes”, “criminales”, “violencia”, “violento”, “violenta”, “agresión”, “agresiones”, “crimen organizado”, “narcotráfico”, “tráfico de drogas”

#### 3. Relacionadas a población LGBTIQ+

"lgbti", "lgbti+", "lgbt", "lgbtq", "lgbtiq", "lgbtqia", "diversidad sexual", "orientación sexual", "identidad de género", "género no binario", "transgénero", "transexual", "bisexual", "pansexual", "homosexual"

## Anexo 2. Palabras NER

La selección de variables a etiquetar y posteriormente a identificar son las siguientes:

<b>Categoría</b>	<b>Variable</b>	<b>Descripción</b>
Identificación de la víctima	Victim First Names	Primeros nombres de la víctima
	Victim Last Name	Apellidos de la víctima
	Victim Age	Edad de la víctima
	Victim Profession	Profesión de la víctima
	Victim Gender	Género de la víctima
	Victim Nationality	Nacionalidad de la víctima
	Victim Criminal background	Antecedentes penales de la víctima
	Number of Victims	Número de víctimas
Identificación del perpetrador	Perpetrator First Name	Primeros nombres del perpetrador
	Perpetrator Last Name	Apellidos del perpetrador
	Perpetrator Alias	Alias del perpetrador
	Perpetrator Age	Edad del perpetrador
	Perpetrator Profession	Profesión del perpetrador
	Perpetrator Gender	Género del perpetrador
	Banda Criminal	Banda criminal al que pertenece el perpetrador
	Number of perpetrators	Número de perpetradores
Identificación del hecho	Sensacionalismo	Utilización de lenguaje sensacionalista en la noticia
	Weapon	Arma utilizada para cometer el homicidio
	Relationship	Relación entre el perpetrador y la víctima
	Motive	Motivo del hecho
	Date	Fecha de hecho del hecho
	Municipality	Municipio de ocurrencia del hecho
	Department	Departamento de ocurrencia del hecho
	Place	Lugar de ocurrencia del hecho