Universidad de SanAndrés

Herramientas de Inteligencia Artificial en la lucha contra la violencia de género digital

Un Estudio con Enfoque en el Español Rioplatense

Lic. Paula Luvini

Director: Dr. Juan Manuel Pérez

Maestría en Ciencia de Datos

29 de enero de 2024

AGRADECIMIENTOS

Quisiera expresar mi profundo agradecimiento a MSc Laila Sprejer, quien lideró el proyecto original que sirvió como punto de partida para esta investigación. Su visión y liderazgo fueron fundamentales para la concepción y desarrollo de este trabajo. Agradezco sinceramente la oportunidad que me brindó para participar en este proyecto, que fue una experiencia enriquecedora y formativa para mi.

También quiero agradecer a mi tutor y director de tesis, Juan Manuel Pérez, por su contribución a lo largo de este proceso de investigación. Su profundo conocimiento en el campo del Procesamiento de Lenguaje Natural (PLN) y generosidad al compartir material y experiencia fueron fundamentales para el desarrollo y la culminación de esta tesis.

Asimismo, quiero extender mi reconocimiento al Programa de las Naciones Unidas para el Desarrollo (PNUD) Uruguay, al Instituto Nacional de las Mujeres de Uruguay y a la Embajada Británica de Montevideo. Su valioso financiamiento y apoyo han hecho posible la realización de esta investigación, contribuyendo de manera significativa a la comprensión y abordaje de la violencia digital de género.

Este trabajo no habría sido posible sin la colaboración y el respaldo de estas instituciones. Agradezco sinceramente su compromiso con la investigación y la promoción de soluciones innovadoras para abordar problemáticas de violencia de género en el ámbito digital.

RESUMEN EJECUTIVO

El abuso y la violencia digital puede describirse como aquel discurso que ejerce violencia y acoso a través de redes sociales como WhatsApp, Facebook, Twitter, entre otras plataformas. Desde la proliferación de estas plataformas digitales, este tipo de discurso ha tomado gran relevancia en redes sociales, dando lugar a formas específicas de acoso digital y haciendo la definición de abuso en línea compleja. En este sentido, diversos estados y organizaciones de la sociedad civil han estudiado el tema y tomado acciones para moderar el contenido abusivo en línea, además de las acciones propias que las empresas pueden tomar. También se implementaron protocolos en caso de ser víctima de violencia digital.

Varios estudios internacionales y locales revelan que las mujeres y otras identidades feminizadas sufren altos índices de acoso virtual. Esto tiene varias consecuencias, como un impacto negativo en la salud mental de las víctimas o en la participación política de las mujeres, evidenciado por testimonios que muestran cómo recibir amenazas e insultos en línea puede condicionar su vida política.

Los estudios que analizan la violencia digital hacia las mujeres suelen emplear encuestas o entrevistas para comprender la profundidad del impacto en las víctimas. Sin embargo, estas metodologías pueden limitar la visión general. Por este motivo, en esta tesis queremos abordar esta limitación introduciendo un modelo que puede detectar diversos tipos de sentimiento en los tweets que mujeres que están inmersas en la política y el ámbito público pueden sufrir. Para ello, utilizamos un dataset que fue construido para un proyecto solicitado por PNUD Uruguay, el Instituto Nacional de las Mujeres del gobierno de Uruguay y la Embajada Británica de Montevideo. El dataset fue construido en un proceso de anotación con dos grupos de anotadores uruguayos que debían clasificar varias cosas de cada tweet recibido, entre ellas si los tweets eran dirigidos o no y si se trataban de tweets abusivos, críticos, neutrales, positivos o de contra-abuso. Se anotaron 9.000 tweets en total, seleccionados aleatoriamente y mediante active learning para optimizar la eficiencia del modelo.

A partir de la construcción de este dataset se evaluó el agreement entre anotadores y la calidad del mismo encontrando resultados alentadores. Luego se realizó un benchmarking con los principales transformers pre entrenados de lenguaje (Bert, Roberta, Robertuito, Electra y Bertin) y se comparó el desempeño de cada uno. El modelo que mostraba mejores resultados en el dataset de validación fue Robertuito, con el que se evaluó el modelo final y se realizó un análisis de resultados y de error detallado para evaluar próximos pasos a seguir en futuros trabajos.

Índice general

1. INTRODUCCIÓN

El abuso y la violencia digital pueden definirse como aquellos mensajes despreciativos, de burla y que intentan descalificar o impugnar el valor de las personas. Estos mensajes suelen dirigirse particularmente hacia grupos y colectivos de individuos discriminados y amenazados por características como puede ser su género, creencia religiosa u origen. Si bien hay varias definiciones y conceptualizaciones respecto a lo que constituye un mensaje abusivo, no hay disidencias respecto a que estos mensajes son hirientes y afectan la vida personal de las víctimas así como también pueden fomentar ataques de violencia física.

Muchos de los estudios y la literatura existente sobre el tema utilizan metodologías cualitativas como entrevistas a personas damnificadas o encuestas a los grupos de interés. Estas metodologías son muy útiles para orientar las líneas de investigación, dado que a través de testimonios concretos se pueden identificar problemas o contrastar hipótesis. Por ejemplo, Posetti et al. [33] pudo confirmar que una mayoría de periodistas mujeres había recibido acoso virtual debido a su labor, un hallazgo contundente sobre una encuesta de 901 casos.

El mencionado estudio también indaga sobre las consecuencias que tienen estos ataques virtuales a las mujeres, encontrando que un 26 % de ellas sufre un impacto en su salud mental tras los mensajes recibidos. Las consecuencias que los ataques tienen sobre la vida de las víctimas son alarmantes. Organizaciones como la End Violence Against Women [44] reconocieron tempranamente al abuso online en la misma escala de importancia que la violencia *offline*. Esto es relevante tanto por como se dimensiona y conceptualiza el estudio del tema como por la legislación y jurisprudencia que se desarrolla en torno al mismo.

El caso que concierne a esta tesis es la violencia dirigida hacia mujeres y disidencias que ejercen un lugar de influencia en la vida pública. Al respecto, el Equipo Latinoamericano de Justicia y Género (ELA) realizó en 2021 una campaña llamada "Impacto de la violencia política por cuestiones de género" donde entrevistaron a varias mujeres e identidades feminizadas que participan en política en Argentina. Lo notorio de estas entrevistas fue que, a pesar de que pertenecían a fuerzas políticas distintas y opuestas, todas tenían testimonios similares respecto al impacto que la violencia digital tiene en sus vidas. En este sentido, las entrevistadas mencionan que muchas veces los mensajes recibidos por redes sociales no son críticas a sus agendas políticas ni buscan una discusión con argumentos: se busca simplemente la descalificación.

Silvia Lospennato[1], entonces Diputada Nacional, menciona que estos discursos muchas veces habilitan a que la discusión democrática vire hacia un discurso de odio que no es tolerable. Es decir, los ataques virtuales buscan inhabilitar a estas mujeres de la vida política. Tanto es así que la entonces legisladora de la Ciudad de Buenos Aires Ofelia Fernández[2] decía respecto a los ataquentes: "Tienen que hacer que te vayas sola, por la puerta chica, que tomes la decisión de decir: ¡Yo con esto no puedo!". Estos testimonios son una clara muestra de que el abuso digital

[1] `https://www.youtube.com/watch?v=WLjpindydpk`
[2] `https://www.youtube.com/watch?v=RCpVgNgnEg4&t=164s`

funciona como una forma de silenciamiento y de coerción sobre las mujeres, como puntualiza Megarry [25].

A pesar de lo valioso que es contar con testimonios en primera persona, tanto las encuestas como las entrevistas fallan a veces en mostrar un panorama completo de la situación y pueden ser criticados o reducidos a anécdotas. Esto es así tanto por la baja escalabilidad de estas metodologías como por el componente declarativo que tienen: en el fondo, se trata de respuestas y afirmaciones subjetivas de personas con entendimientos que pueden ser diversos sobre el mismo problema. En este sentido, sumar herramientas que evolucionen con el tiempo y que puedan recoger la mayor cantidad de ejemplos posibles ayuda a la fortaleza de estos estudios. Para esto lo que se debe solucionar es el tema de la escalabilidad, es decir cómo logramos tener un diagnóstico amplio en el tiempo y que tenga la mayor cantidad de ejemplos posibles.

Para esto, es necesario el desarrollo de herramientas de inteligencia artificial que permitan procesar gran cantidad de texto y escalar la tarea de la clasificación del mismo como violento. Dentro de las competencias del Procesamiento de Lenguaje Natural (PLN) encontramos a la clasificación de textos según su sentimiento o intención como una de sus tareas más utilizadas. De esta manera, se puede entrenar un algoritmo para que clasifique diversos mensajes según si su contenido es abusivo o no. Estas herramientas, además de ser necesarias para el estudio de este fenómeno también lo son para la moderación del mismo. Es decir, para evitar la proliferación de estos discursos con miles de mensajes siendo enviados en una gran variedad de redes sociales y en un diversas lenguas, es imposible que las plataformas incorporen moderadores humanos que se encarguen de esta tarea.

Es importante que estas herramientas se desarrollen pensando en las sociedades y poblaciones que pretenden ayudar. Contar con modelos entrenados con datos propios, de cuya recolección conocemos su metodología y en idiomas distintos al inglés es muy valioso. Es por eso que esta tesis busca sumar una nueva herramienta pensada para el Sur Global y el español rioplatense. Particularmente, utilizaremos un dataset cosntruido en el año 2022 en un proyecto del Instituto Nacional de las Mujeres, el Programa de las Naciones Unidas para el Desarrollo (PNUD) y la Embajada Británica de Montevideo. El objetivo del proyecto era la confección de un Monitor de Violencia Digital de Género para Uruguay[3], con el interés de seguir los ataques que sufrían algunas personalidades públicas de dicho país. A raíz de ese trabajo se construyó un dataset innovador e inédito en la región, anotado por personas locales y diseñado específicamente para servir a construir una herramienta cuantitativa que provea de evidencia en tiempo real sobre el nivel de agresiones e insultos que reciben figuras públicas como mujeres políticas, periodistas, comunicadoras, activistas y artistas en Twitter.

En esta tesis, hemos recopilado el proceso de anotación y construcción del corpus que se hizo durante este proyecto e incorporado nuevas métricas para evaluar la calidad de este dataset. Además, hemos utilizado el corpus creado para realizar algunos experimentos y entrenamientos distintos a los que se hicieron en el proyecto original, incorporando la clasificación de mensajes a categorías inéditas. Para lograr esto se realizaron experimentos de clasificación buscando

[3] https://www.violenciadigitalmujeres.uy/

obtener un modelo que pueda identificar si un mensaje estaba dirigido hacia una usuaria y su categoría de intención: mensajes abusivos, críticos, positivos y neutrales. Como resultado, nos encontramos con métricas alentadoras y más altas de lo esperado debido a la naturaleza subjetiva de la tarea, tanto a la hora de anotar como al momento de entrenar un transformer para clasificar. Tras un benchmarking con varios modelos, el de mejor performance y elegido finalmente fue RoBERTuito Pérez et al. [31].

Nuestra propuesta tiene la intención de aportar a desarrollar mejores mecanismos automáticos de detección de abuso hacia mujeres y de fomentar la proliferación de estos estudios pensados desde el Sur Global, atendiendo a posibles sesgos e implicancias éticas que tengan estos desarrollos. Por eso destacamos que todos los recursos utilizados para el entrenamiento se encuentran en español, lo cual no es común en la literatura aún cuando se trata del segundo idioma en hablantes nativos del mundo. En este aporte cabe destacar la iniciativa y el interés del gobierno de Uruguay, de PNUD y de la Embajada Británica de Montevideo al financiar la construcción del corpus.

La tesis se encuentra organizada de la siguiente manera: en primer lugar, realizamos una revisión de la literatura existente en Inteligencia Artificial y en deep learning, con el objetivo de entender el recorrido de la misma hasta los modelos de lenguaje preentrenados basados en transformers, que son los que finalmente usamos para entrenar nuestros modelos de clasificación. La siguiente sección revisa literatura previa en estudios particulares de detección de abuso o de discursos de odio con herramientas de aprendizaje automático, y otros estudios con diferentes metodologías orientados a analizar violencia de género. En el siguiente capítulo explicamos y desarrollamos cómo se construyó el corpus utilizado y realizamos una evaluación de calidad del mismo. Acto seguido presentamos los resultados del benchmarking del modelo, con una nota acerca de la función de pérdida particular que utilizamos en este caso para respetar la estructura de nuestros datos. Por último, analizamos la performance del modelo y realizamos un análisis de error que nos permitió dejar algunas discusiones para proseguir en futuros estudios.

2. TRABAJO PREVIO

Esta sección de literatura previa tiene como primer objetivo proporcionar una visión general de los desarrollos más significativos en el campo de Machine Learning y Procesamiento de Lenguaje Natural (PLN). Se repasaran algunas teorías fundamentales, técnicas avanzadas y aplicaciones prácticas que han impulsado su crecimiento y evolución a lo largo del tiempo. Además, en la segunda parte revisaremos algunos trabajos previos sobre violencia digital y cómo se estudió la misma en la literatura, repasando también la aplicación de estos trabajos a cuestiones relacionadas con el género.

2.1. Inteligencia Artificial

Los orígenes de la inteligencia artificial se remontan a la Conferencia de Inteligencia Artificial de Dartmouth en 1955, su evento seminal. Allí se presentó la propuesta "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence"McCarthy et al. [21] donde se utilizó por primera vez el término **inteligencia artificial**. Este evento marcó el punto de partida de un campo que se centraría en el estudio y desarrollo de algoritmos capaces de aprender de datos y realizar predicciones. Esta reunión inicial fue organizada por John McCarthy, en ese momento profesor de matemáticas en Darmouth. En la propuesta, McCarthy afirmó que la conferencia tenía como objetivo "proceder sobre la base de la conjetura de que cada aspecto del aprendizaje u cualquier otro rasgo de la inteligencia puede ser descrito tan precisamente que una máquina puede ser diseñada para simularlo."

El concepto de **aprendizaje automático** por otro lado, fue registrado por primera vez en 1959 por Arthur Samuel mientras trabajaba en IBM. El aprendizaje automático es un campo fundamental en la inteligencia artificial que se enfoca en el estudio y desarrollo de algoritmos capaces de aprender de datos y realizar predicciones precisas. A medida que la disponibilidad de datos continúa creciendo exponencialmente en la era digital, el aprendizaje automático se ha convertido en un campo esencial para desbloquear conocimiento y patrones ocultos en grandes conjuntos de datos, transformando la forma en que interactuamos con la tecnología y tomamos decisiones en diversas industrias.

El PLN es un área interdisciplinaria que se sitúa en la intersección entre la lingüística, la inteligencia artificial y las ciencias cognitivas. Su importancia radica en la capacidad de las computadoras para comprender, analizar y generar texto humano de manera automatizada. A medida que la cantidad de datos textuales disponibles en línea ha aumentado exponencialmente en los últimos años, el PLN se ha convertido en una disciplina crucial para abordar desafíos que van desde la búsqueda de información y la traducción automática hasta la corrección de textos, la generación de resúmenes y la detección de sentimientos.

En el contexto de problemas de PLN como el que revisaremos en esta tesis, nos enfrentamos a desafíos que pueden abordarse mediante la aproximación de una función $f : D \to O$, donde D

es el *dominio* y O el codominio o posibles salidas. Ahora bien, en contextos en los que estamos trabajando con tareas específicas de texto, estas funciones suelen ser desconocidas, altamente no lineales y difíciles de expresar analíticamente. En el caso en que las salidas, o sea el conjunto O sea finito, diremos que estamos ante un problema de **clasificación** y llamaremos a nuestro aproximador $\widehat{f}$ **clasificador** y a cada uno de las posibles salidas **clases**. Los problemas que trataremos en esta tesis son, en efecto, problemas de clasificación.

En estos problemas el objetivo es predecir, a partir de un texto que puede ser un comentario en una red social, alguna característica discreta del mismo, como su polaridad (positiva, neutra o negativa) o la presencia de discriminación (no discriminatorio o discriminatorio). En esta tesis, nuestro objetivo va a ser detectar comentarios abusivos dirigidos hacia mujeres respecto a los no abusivos, que además clasificaremos de acuerdo a su intención. Para abordar este problema, se pueden utilizar técnicas de aprendizaje supervisado que involucran instancias de entrenamiento junto con sus etiquetas correspondientes para entrenar al aproximador $\widehat{f}$. El conjunto de entrenamiento, denotado como $\{(x_1, y_1), \ldots (x_n, y_n)\}$, es esencial para desarrollar modelos efectivos en problemas de clasificación y otros relacionados en el ámbito del procesamiento de lenguaje natural.

El eje de la tesis estará puesto en los modelos basados en redes neuronales. Estos modelos han tomado por completo el estado del arte en PLN para casi cualquier tarea conocida. Para esto, a continuación haremos un repaso de sus distintas variantes.

2.2. Deep Learning

El deep learning es un tipo de enfoque del aprendizaje automático basado en redes neuronales profundas, que aprende representaciones de los datos en múltiples niveles de abstracción o capas. Hoy en día deep learning es sinónimo de redes neuronales, porque estas redes modernas a menudo son profundas y constan de muchas capas.

Las redes neuronales son una herramienta con muchos años de bagaje académico. Su origen se remonta a la neurona McCulloch-Pitts de McCulloch y Pitts [22], un modelo simplificado de neurona humana pensado como un elemento de cómputo. El uso moderno de las redes neuronales "artificiales" en el procesamiento del lenguaje ya no se basa en estas primeras inspiraciones biológicas. Una red neuronal moderna es un conjunto de pequeñas unidades de cómputo que toman uno o más valores de entrada y produce uno o más valores de salida. Las redes neuronales son una poderosa herramienta cuando se aplican a problemas de clasificación. Estas aplicaciones son conocidas como redes feedforward, ya que el proceso avanza de manera iterativa desde una capa de unidades a la siguiente.

2.2.1. RNN, LSTM y GRU

Dentro de los modelos de deep learning aplicados al lenguaje son muy importantes las Redes Neuronales Recurrentes (RNN por las siglas en inglés). La RNN está diseñada para procesar

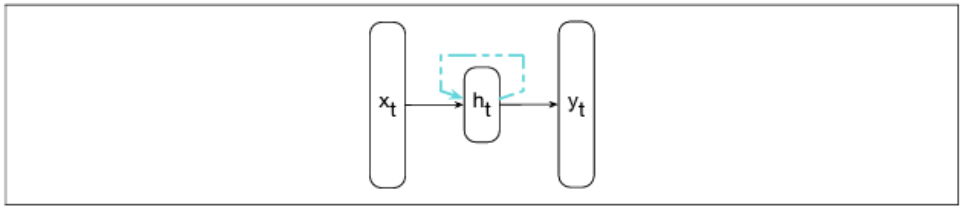


Fig. 2.1: Red Neuronal Recurrente tras Elman (1990). Fuente: Jurafsky y Martin [13]

datos secuenciales, es una red que contiene un ciclo en sus conexiones de red, es decir que el valor de alguna unidad depende directa o indirectamente de sus propias salidas anteriores como entrada. Aunque estas redes son poderosas en términos de su capacidad para modelar secuencias, son difíciles de analizar y entrenar. De todas maneras, dentro de la amplia categoría de redes recurrentes, existen arquitecturas con restricciones que han demostrado ser extremadamente efectivas cuando se aplican al procesamiento del lenguaje. Un tipo de RNN es la red de Elman, conocida como red simple recurrente Elman [9], de donde surgen también las Redes de Memoria a Corto y Largo Plazo (LSTM por las siglas en inglés).

Las RNN tienen un vector de entrada que representa la entrada actual, x_t, que se multiplica por una matriz de pesos y luego se pasa a través de una función de activación no lineal para calcular los valores de la capa de unidades ocultas. Esta capa oculta se utiliza para calcular una salida correspondiente, y_t. La diferencia respecto a las feedforward networks radica en el enlace recurrente mostrado en la línea punteada de la Figura 2.1. Esta iteración recurrente genera en la red neuronal una suerte de "memoria" que va a influir en la siguiente codificación.

Dicho todo esto, las RNN de Elman sufren de varios problemas. Uno de ellos es la dificultad para manejar información que no se encuentra cerca dentro del texto. Porque si bien tienen toda la secuencia en las capas ocultas, las RNN se enfocan más en la información cercana, dificultando algunas tareas de procesamiento de lenguaje. Esto es así porque las capas y sus valores tienen que llevar la información útil para decidir la palabra inmediata y además actualizar y llevar hacia adelante otra información necesaria para futuras decisiones, es decir una doble tarea.

Otro inconveniente que presentan es el de vanishing gradient o los gradientes que desaparecen. Esto sucede cuando durante el entrenamiento las capas ocultas atraviesan varias multiplicaciones que llevan a valores cada vez más pequeños y que hacer que los gradientes disminuyan en valor y desaparezcan. Una manera de solucionar este problema es utilizando las de Memoria a Corto y Largo Plazo (LSTM) (Hochreiter y Schmidhuber [11]) o las Gated Recurrent Units (GRU) (Cho et al. [6]).

Las LSTM lidian con el vanishing gradient al dividir los problemas, eliminando información innecesaria para el contexto y agregando otra relevante para futuras decisiones. Esto se logra mediante la incorporación de capas de contexto explícitas y el uso de unidades especiales con compuertas (gates) que controlan el flujo de información. Las GRU son una variación más simplificada de las LSTM y que también buscan solucionar al vanishing gradient. Se las conoce como una variación de las LSTM porque también utilizan a las compuertas o gates: tienen una compuerta de actualización y una de reseteo. La primera es la que determina cuánta informa-

ción de los pasos anteriores necesita continuar a la siguiente etapa. La segunda compuerta, la de reseteo, es la que decide cuánta información de las etapas anteriores olvidar.

2.2.2. Transformers

Los transformers, al igual que las LSTM también tienen mecanismos para lidiar con el gradiente que desaparece (autoatención y codificaciones posicionales). A diferencia de las LSTM, los transformers no se basan en conexiones recurrentes como las RNN y los LSTM, lo que los hace más eficientes de utilizar. En las redes anteriores, dado que los cómputos eran secuenciales, con h_t calculándose en base a h_{t-1} no podía darse un proceso paralelo con varias representaciones generadas al mismo tiempo. Los transformers solucionan este problema, haciéndolos una herramienta eficiente y útil, pudiendo ser entrenados significativamente más rápido que una red recurrente o convolucional.

Ahora bien, ¿cómo están compuestos los transformers? Los transformers son bloques de redes multicapas que combina capas lineales simples, redes feedforward y capas de self-attention (auto-atención). La utilización del auto-atención es su diferencial, porque hace que la red tenga información de contexto y la utilice pero sin hacer las conexiones recurrentes de las RNN. En Vaswani et al. [38] se introdujo por primera vez la auto-atención a los transformers para la traducción del lenguaje y lo publicaron en el mencionado paper cuyo título era contundente "Attention Is All You Need".

La auto-atención fue introducida por Parikh et al. [27] y se trata de la posibilidad de comparar un elemento con una colección de elementos. En el caso de las redes neuronales, la auto-atención genera representaciones de un vector de entrada al observar la totalidad de la secuencia y no sólo el paso anterior como ocurría en las RNN. De esta manera, los transformers pueden considerar todo el contexto de una secuencia al procesar cada elemento, capturando eficazmente las dependencias de largo alcance. Esto se logra dado que la auto-atención es el mecanismo que pondera la importancia de diferentes palabras de la secuencia respecto a una palabra específica.

En la Figura 2.2 podemos observar la arquitectura del transformer tal como la expresaron Vaswani et al. [38]. De esta manera, el modelo recibe como input una secuencia de texto que va a ser representada por un embedding que tome cada palabra o token y lo transforme en un vector numérico. Una vez que se tienen estos vectores se le agrega un positional encoding que le va a dar información al modelo de dónde estaba ubicado cada token en la secuencia. Los positional encodings deben agregarse porque estos modelos, a diferencia de las redes recurrentes que revisamos previamente, no tienen de otra manera una referencia de la ubicación de la palabra en el texto.

De esta manera, la arquitectura está organizada como encoder-decoder con 6 capas donde cada capa del encoder utiliza un mecanismo de auto-atención, seguido de una capa feedforward. El transformer entonces, está formado por un encoder que toma la secuencia de entrada y la codifica y un decoder que después da como output las probabilidades de cada palabra. El primer componente del encoder es el de auto-atención que como se ha descrito anteriormente permite

al modelo medir la importancia de las diferentes palabras de la secuencia de entrada en relación con una palabra específica. Luego, el output de la auto-atención pasa por una capa de feedforward. El decoder, por otro lado, también tiene una capa de auto-atención pero "enmascarada" para garantizar que cada posición sólo pueda atender a las posiciones anteriores a ella. Esto evita que la información se vea afectada por posiciones futuras durante el entrenamiento. Luego hay otra capa de atención que toma los outputs del bloque del encoder, lo que hace que el decoder pueda centrarse en las partes relevantes de la secuencia de input. Por último hay una capa feedforward al igual que en el encoder que procesa los resultados de las capas de atención.

Como vemos en los gráficos, otro elemento a considerar dentro de los bloques son las conexiones residuales. Las conexiones residuales son conexiones que permiten que la información pase de una capa inferior a una capa superior sin pasar por la capa intermedia. Esto mejora el aprendizaje al permitir que la información avance y el gradiente retroceda sin obstáculos entre las capas, lo que proporciona a las capas de nivel superior un acceso directo a la información de las capas inferiores. En los transformers, las conexiones residuales se implementan sumando el vector de entrada de una capa a su vector de salida antes de avanzar.

Los outputs finales del transformer se obtienen mediante la capa lineal y que está sucedida por una función de activación softmax, como se ve en la Figura 2.2. La capa lineal es una red neuronal que proyecta el vector obtenido del decoder en otro vector más llamado vector logits, que va a tener los score o resultados del entrenamiento. En el caso de un problema de clasificación, tendría la confianza que tiene el modelo sobre cada una de las clases. Por último, la capa softmax convierte normaliza este vector de resultados entre 0 y 1.

Desde la publicación del trabajo de Vaswani et al. [38] los transformers han revolucionado a los modelos de procesamiento de lenguaje natural. Cada modelo de IA desde GPT (*generative pre-training*) introducido en Radford et al. [35], GPT-3 (Brown et al. [2]), GPT-4 (Achiam et al. [1]) y Github Copilot (Peng et al. [29]) están basados en la arquitectura de transformers. En este sentido, sus aplicaciones abarcan tareas como la traducción automática, el resumen de textos, la respuesta a preguntas y el modelado del lenguaje.

2.2.3. Encoders de transformers preentrenados

Tras la revolución de los transformers aparecieron los encoders de transformers bidireccionales. Estos modelos se conocen como bidireccionales porque pueden ver el contexto derecho e izquierdo de texto a la vez. Este cambio ocurre en la capa de auto-atención. En la Figura 2.3 vemos como en todos los transformers que revisamos hasta ahora el conocimiento era incremental token a token. Esto puede hacer que los transformers no sean tan buenos al ser utilizados en clasificaciones secuenciales de texto o en identificación de etiquetas. Por eso es que se introduce la bidireccionalidad, que permite que se use la información de la secuencia total del texto, como se ve en la Figura 2.4.

A su vez, el hecho de que los modelos sean bidireccionales cambió la forma en que estos modelos se entrenan. Antes, los modelos buscaban predecir cuál era la próxima palabra que iba

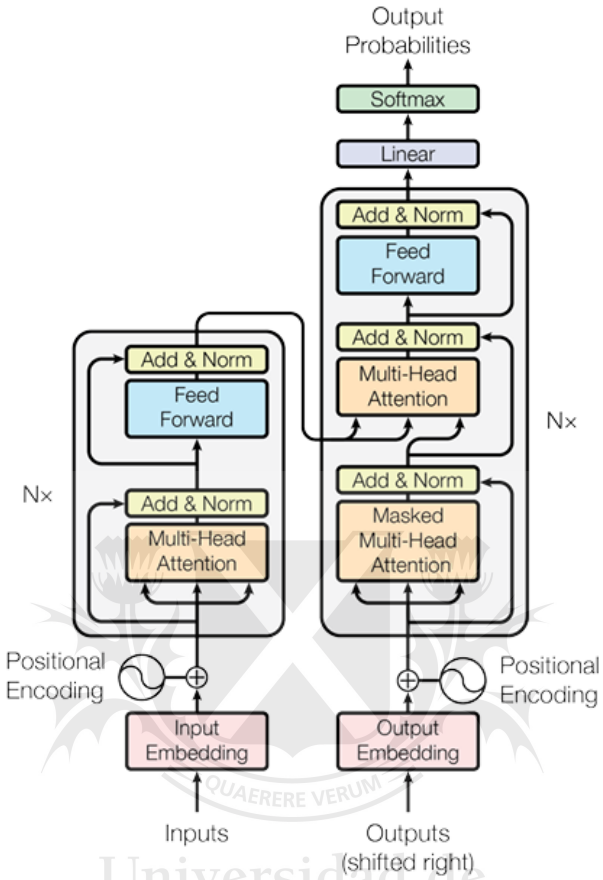


Fig. 2.2: Arquitectura de los modelos transformer. Fuente: Vaswani et al. [38]

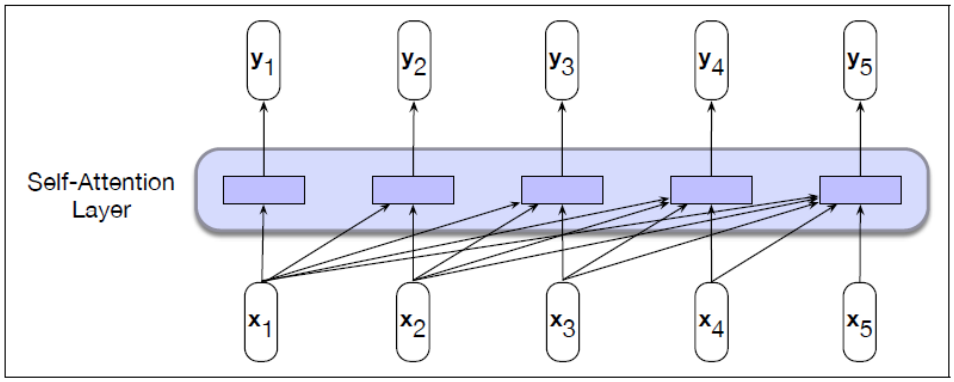


Fig. 2.3: Capa de auto-atención de un transformer unidireccional. Fuente: Jurafsky y Martin [13]

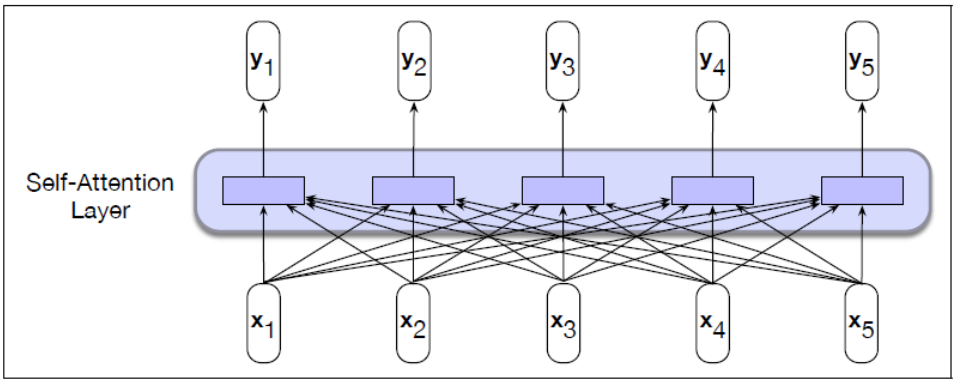


Fig. 2.4: Capa de auto-atención de un transformer bidireccional. Fuente: Jurafsky y Martin [13]

a venir en una oración, ahora al tener el contexto bidireccional se va a entrenar a partir del "masking" o enmascaramiento de partes del texto que no sean necesariamente las últimas. Esta tarea es usualmente conocida como *Cloze task* e introducida en Taylor [37]. Es decir, inicialmente los modelos de texto buscaban predecir de la primera forma y ahora los bidireccionales utilizan la segunda:

Los perros ladran y los pájaros `[MASK]`

Los `[MASK]` ladran y los pájaros cantan.

Otro hito de relevancia que introdujeron estos encoders preentrenados es que pueden ser "finetuneados" para servir a aplicaciones específicas al agregar una capa de clasificación al modelo pre entrenado. El *fine tuning* de un modelo implica, en efecto, tomar un modelo pre entrenado para la predicción de texto contextualizado y re entrenarlo para que pueda hacer otras tareas, como la predicción de etiquetas o clasificación de texto. Para ello se requiere del uso de datos etiquetados de acuerdo a la tarea correspondiente para entrenar al modeo bajo estos parámetros adicionales, manteniendo los parámetros originales del modelo. En machine learning esta tarea es conocida como *transfer learning*: es la reutilización del conocimiento de un modelo pre entrenado para que sirva a otra tarea.

La posibilidad de hacer un *fine tuning* a partir de un modelo preentrenado con inmensas cantidades de texto hace que el proceso de entrenar estos modelos en una nueva tarea no sólo sea sencillo sino también eficiente. Es decir, con estos modelos necesitamos una cantidad de datos mucho menor y con una performance muy buena. El más utilizado de estos modelos es BERT (Bidirectional Encoder Representations from Transformers), introducido por Devlin et al. [8]. Otros modelos pueden ser RoBERTa (Robustly Optimized BERT Pretraining Approach) presentado por Liu et al. [20]. Por último, cabe destacar el desarrollo de RoBERTuito introducido por Pérez et al. [31], un modelo pre entrenado con más de 500 millones de tweets en español y que utilizamos en esta tesis con buenos resultados.

2.3. Violencia digital

El comienzo del siglo XXI fue testigo de la aparición de redes sociales y de mensajería instantánea que trajo aparejadas nuevas formas de comunicarnos. Con ello también aparecieron nuevas formas de acoso y de violencia, que mutaron a expresiones y códigos particulares del ecosistema virtual. Esta violencia digital puede llegar a ser particularmente fuerte contra algunos colectivos y minorías, como mujeres, grupos religiosos o extranjeros.

Estos fenómenos, si bien aparentemente virtuales, terminan traspasando el mundo digital y afectando la vida cotidiana de las personas que los sufren, motivo por el cual es importante estudiarlos y analizarlos. En primer lugar porque las amenazas y denigraciones online pueden afectar a las víctimas que lo sufren como demuestran Williams y Tregidga [43] y McKenna y Bargh [24]. En segundo lugar, es relevante monitorear de estos ataques se porque se ha demostrado que tienen relación con otros actos físicos violentos, como McIlroy-Young y Anderson [23] respecto al tiroteo en una sinagoga de Pittsburgh. En otro ejemplo, Müller y Schwarz [26] encontraron que los tuits islamofóbicos que el presidente Trump emitió durante el ejercicio de su presidencia parecen aumentar los crímenes de odio contra musulmanes en los días siguientes. Lee y Leets [17] también estudiaron las consecuencias de los grupos supremacistas blancos en los adolescentes, al utilizar internet para captar nuevos miembros y transmitir sus mensajes y narrativa. Con la gravedad que esto ha tenido sobre los grupos atacados, también debemos considerar que el entorno virtual retroalimenta asimismo estos ataques violentos. Burnap et al. [3] analiza esto respecto a los ataques terroristas en Woolwich (Londres) en 2013, donde la propagación posterior del tema en Twitter depende fuertemente del contexto social de quien twittea.

En este sentido, varios trabajos se han desarrollado generando corpus de anotaciones y disponibilizando los mismos para el estudio la violencia digital, como Chandra et al. [5], Ibrohim y Budi [12] y Lee, Yoon y Jung [18]. Estos estudios son valiosos tanto por su análisis como por la creación de este tipo de corpus que permiten que otros los utilicen para su análisis y para poder generar herramientas de contra-discurso que puedan generar un impacto positivo en la sociedad. De todas maneras, el idioma y la regionalización de las anotaciones muchas veces hacen que estos datasets tengan limitantes a la hora de aplicarse en otros contextos. Esto aún sin mencionar que al no haber una definición universal de abuso ni de toxicidad dentro del campo del PLN, muchos midan lo mismo nominalmente hablando pero no así en realidad.

Dentro de las herramientas de detección de agresión en texto, una muy utilizada internacionalmente por académicos y empresas para detectar mensajes abusivos o tóxicos es Perspective API[1], un desarrollo de Google y de Jigsaw. Perspective puede ser utilizado por desarrolladores que deseen identificar diversos atributos en texto, como puede ser el nivel de toxicidad, amenazas, insultos que tenga. Lo que se va a hacer a través de la API es obtener para cada texto una probabilidad entre 0 y 1 de cada variante a evaluar, donde un puntaje alto indica grandes probabilidades de que el comentario pueda percibirse como tóxico, amenzante o insultante. En la página web se pueden realizar pruebas fácilmente ingresando texto y obteniendo un porcentaje de toxicidad como se ve en la Figura 2.5 donde probamos utilizando uno de los tweets recolectados para este estudio. La definición de toxicidad que Jigsaw emplea es una de las más

[1] https://perspectiveapi.com/

Fig. 2.5: Captura de interfaz de Perspective de uso libre

utilizadas en este tipo de estudios, tratándose de "un comentario grosero, irrespetuoso o poco razonable que haga que alguien abandone una discusión."

Si bien en la página web sólo se muestra el nivel de toxicidad, como mencionamos utilizando la API se pueden obtener muchos más atributos, como contenido amenazante o insultante. En este sentido, es ampliamente utilizada por empresas de comunicación y por algunas redes sociales para la moderación de contenido tóxico. Muchos trabajos además comparan sus propias modelizaciones de clasificación para detectar discurso abusivo o de odio con los resultados que Perspective puede dar, para tener una comparación baseline.

Definir qué consideramos como abuso o violencia digital es una de las primeras tareas que la literatura intentó acordar, con bastante dificultad como menciona Vidgen, Margetts y Harris [40]. Lo cierto es que las definiciones pueden cambiar respecto al objeto de estudio que tengamos y no existe actualmente una definición universalmente aceptada por la literatura de lo que se encuentra bajo el paraguas de la violencia y agresión escrita. Poletto et al. [32] también se hace eco de esta problemática y de las complejidades de la definición y la distinción entre todas las categorías que puede haber dentro del abuso o toxicidad. En este trabajo, se hace una revisión del estado del arte de las definiciones de estos fenómenos en la literatura existente y además se propone un marco teórico condensado en la Figura 2.6 respecto a varios conceptos y estudios que pueden estudiarse. En este sentido, el abuso o la toxicidad en el discurso va a incluir a los discursos de odio, pero no todo lo que constituye toxicidad va a ser esto. Asimismo, el lenguaje ofensivo no siempre va a ser tóxico o abusivo, siendo que podemos encontrar mensajes que presentan malas palabras pero que no son dirigidas a insultar a una persona sino que pueden ser una muestra de folclore o algarabía.

En esta tesis nuestro concepto de interés va a ser el de abuso hacia mujeres, que definimos como aquellas expresiones despreciativas, de burla y/o que descalifican o impugnan el valor de

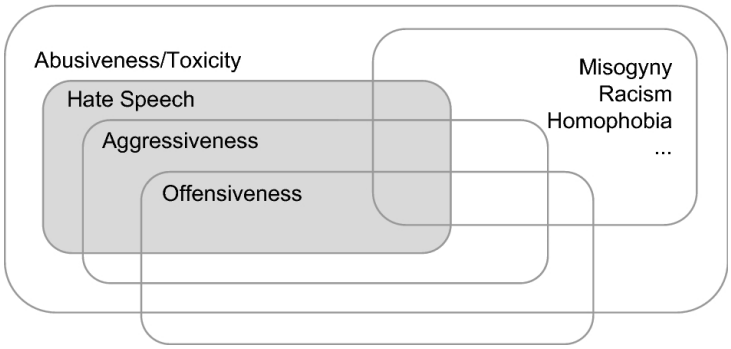


Fig. 2.6: Marco conceptual del abuso, toxicidad y otros relacionados. Fuente: Poletto et al. [32]

las personas, en general relacionadas con su personalidad o cualidades. Dentro de la conceptualización de Poletto et al. [32] estamos incluyendo entonces no sólo discursos de odio por cuestiones de género sino también agresiones y comentarios tóxicos o amenazantes hacia la persona dirigida.

Uno de los trabajos que fue utilizado para esta conceptualización de abuso es el de Vidgen et al. [41]. En ese trabajo, se definen categorías de anotación dentro de las cuales se encuentra la hostilidad contra grupos del este asiático. Se definen dentro de esta hostilidad a los "tweets que expresen abuso o intensa negatividad contra una entidad de Asia Oriental, principalmente mediante la desestimación o el ataque directo". Dentro de esta denominada hostilidad se incluyen también todos los mensajes negativos que incluyan teorías conspirativas de estas poblaciones o mensajes que los identifique como amenzas.

Algunos estudios se enfocan en el concepto de acoso virtual, que es aquel que está dirigido a una persona particular. En el plano internacional, destacan los reportes del Pew Research Center Vogels [42], donde se considera que una persona fue acosada virtualmente si considera que recibió insultos, se los avergonzó de manera intencional, fueron acechados (stalkeados) virtualmente, recibieron amenazas físicas, fueron acosados por un largo período de tiempo o sufrieron acoso sexual. El estudio encuentra que para Septiembre de 2020 un 41 % de los ciudadanos de Estados Unidos sufrieron alguna de las seis formas de acoso virtual nombradas anteriormente.

En el plano local, estos temas se encuentran regulados y con protocolos, en algunas jurisdicciones. En Argentina es relevante la línea 102, que depende de la Secretaría Nacional de Niñez, Adolescencia y Familia (SENAF). La SENAF disponibilizó una guía con instrucciones sobre cómo denunciar o proceder en caso de ser víctima de ciberbullying, grooming, o de la difusión de imágenes privadas sin consentimiento. En este caso, es una línea especial para niños, niñas y adolescentes, uno de los grupos más vulnerables a recibir ataques digitales. El Gobierno de la Ciudad de Buenos Aires (GCBA) también cuenta con un protocolo y recomendaciones en caso de ser víctima de violencia digital[2], a la que define como: "una tipología más en las formas de ejercer violencia y se define como acoso u hostigamiento virtual a través de medios digitales como por ejemplo Facebook, Instagram, Snapchat y Tik Tok, foros en internet, plataformas de juegos o mensajerías y celulares."

[2] `https://buenosaires.gob.ar/violencia-digital`

Una segunda forma de abuso y que distingue Vidgen, Margetts y Harris [40] es cuando el abuso es dirigido hacia un grupo o hacia individuos por pertenecer a dicho grupo, que se suele encontrar y estudiar en la literatura como "discurso de odio". Se advierte que distinguir ambas situaciones en la práctica es difícil y puede haber una superposición entre ambas. La diferencia entre ser víctima de acoso o de discursos de odio muchas veces puede encontrarse en los emisores de los mensajes. Es muy común que las víctimas de abuso en redes tengan como victimarios a miembros de círculos cercanos o conocidos y que quienes reciben mensajes de hostigamiento debido a su género o a pertenecer a algún colectivo reciban esto de usuarios desconocidos.

Dentro de estos discursos discriminadores y segregantes, encontramos a un subgrupo de ellos que tienen como destinatarias a las mujeres y a otras identidades feminizadas. En esta línea de investigación, hay trabajos que se dedican a estudiar violencia hacia grupos de mujeres específicos, como políticas, activistas, periodistas o mujeres en situación de poder. Tal es el caso de Lewis, Rowe y Wiper [19], donde focalizan su trabajo en analizar abuso online dirigido a mujeres que forman parte de debates públicos feministas, particularmente porque la evidencia anecdótica sugería que eran un grupo particularmente atacado. Allí los autores mencionan cómo muchos trabajos sugieren que la violencia online debe ser separada y tratada distinto a un ataque físico. Es por esto que mediante una encuesta online y entrevistas realizadas a actores, se mostró que el abuso online que sufren las activistas feministas comparte muchas características con aquel sufrido *offline*. En este sentido, estemos de acuerdo o no con si ambos ataques deben juzgarse de la misma manera, el estudio resalta que el sistema judicial está fallando al incorporar estas violencias conceptualmente y no dando las respuestas adecuadas a estas situaciones.

Amnistía Internacional (2018)[3] llevó a cabo una investigación en Reino Unido y en los Estados Unidos con varias encuestas y entrevistas, algunas a mujeres en situaciones de poder que sufrieron violencia en Twitter. Dentro de todos estos análisis, hubo uno donde se utilizó aprendizaje automático para medir y analizar el abuso hacia mujeres del Parlamento británico que tenían cuentas de Twitter. El estudio se ubicó temporalmente en los mensajes publicados hasta seis semanas previas a las Elecciones Generales en el Reino Unido y encontró que un 2,85 % de los tweets dirigidos hacia parlamentarias contenían agresiones. Asimismo y más alarmante, se descubrió que Diane Abbott, Secretaria del Interior del Gabinete Sombra y primera parlamentaria negra del Reino Unido, recibió casi la mitad de todos los comentarios abusivos identificados. Esto refuerza la premisa de que estos discursos de odio se dirigen hacia miembros de grupos particulares, en este caso mujeres o minorías.

Posetti et al. [33] es otro antecedente de un trabajo focalizado a grupos donde se condujeron entrevistas online con el objeto de analizar y evaluar la violencia ejercida hacia mujeres periodistas. La motivación surge del interés mismo del Secretario General de las Naciones Unidas (ONU), quien había puntualizado que mujeres periodistas, trabajando en el mismo rubro que otros hombres, recibían mayores agresiones que estos últimos. La encuesta incluyó un total de 901 respuestas de periodistas de 125 países. Algunos hallazgos del trabajo fueron alarmantes, tales como que el 73 % de las mujeres que respondieron a la encuesta declararon haber recibido agresiones virtuales en relación con su trabajo de periodistas. Dentro de las distintas tipologías

[3] https://amnistia.org.ar/wp-content/uploads

de violencia digital, el más frecuente y sufrido por un 49 % de las mujeres fueron los comentarios con lenguaje ofensivo escritos con el objetivo de ofenderlas, lo que indica la importancia del análisis del lenguaje dentro de las redes sociales para sumar evidencia en este sentido.

Este trabajo es también interesante por los resultados que encuentra respecto a las consecuencias e impactos en las mujeres que sufren este tipo de ataques. En este sentido, se destacan las consecuencias de salud mental en las víctimas, reportadas por un 26 % de ellas. Otro hallazgo que también se repite en otros testimonios es la inseguridad que las víctimas sienten aún luego de terminado el ataque y en la vida cotidiana por fuera de las redes sociales. En este reporte, el número de mujeres que reportan esto último alcanza el 17 %. En el trabajo de Amnistía Internacional (2019)[4] para Argentina y en el contexto del debate por la legalización de la interrupción voluntaria del embarazo, se encontró que de todas las mujeres que fueron acosadas verbalmente en redes sociales un 39 % sintió amenazada su integridad física. Esto no se debe solamente al carácter de las agresiones sino también a la amenaza de divulgación de datos personales (como direcciones o números de celular) o de daños en torno a familiares y círculo cercano.

El mencionado artículo de Amnistía Internacional combina metodologías observadas en reportes anteriores: se llevó a cabo una encuesta a 1.200 mujeres, de 18 a 55 años de edad y de todo el país con preguntas referidas a violencia en redes sociales en general y otras específicas al contexto del debate por la ley de interrupción voluntaria del embarazo. Esto se combinó con un análisis cuantitativo de conversaciones e interacciones en Twitter, de donde se seleccionaron tweets y perfiles que exhibieron relevante actividad durante el debate legislativo del 2018. Con los tweets recolectados se hizo un análisis de cuántos de estos contenían agravios y dirigidos a quién, para indagar y analizar estos discursos que propagaban odio. Como resultado se encontró que este discurso violento fue más frecuente hacia algunas de las referentes analizadas que se expresaban y participaban activamente a favor de la sanción del proyecto de ley.

El Equipo Latinoamericano de Justicia y Género (de ahora en adelante ELA) también llevó a cabo un trabajo enfocado en Argentina en el contexto de la campaña electoral de 2019 Justicia y Género (ELA) [14]. En el estudio monitorearon a 22 políticas en la red Twitter y a otras 10 en Facebook e Instagram. Con esto armaron una base de datos y clasificaron los comentarios y palabras basadas en una lista de agresiones que confeccionaron manualmente para medir violencia machista en redes sociales. Como resultado, hallaron que un 5 % de los tweets recolectados contenían algún tipo de insulto. Dentro de estos un 54 % se refería a expresiones discriminatorias, 25 % a acosos, 16 % eran amenazas y un 5 % eran campañas de desprestigio.

ELA es un centro de estudios y de difusión que tiene una trayectoria importante en distintas problemáticas referidas a la inequidad de género. En este sentido, sus trabajos involucran la recolección de muchos testimonios y entrevistas. Estas instancias de recolección son clave porque permiten contar con testimonios certeros acerca del impacto de la violencia que (aunque digital) tiene sobre la vida de las mujeres. Es así que ELA realizó en el marco de su proyecto Mujeres en el poder una serie de entrevistas a mujeres en cargos públicos en Argentina titulado: “Nos quieren sacar de la vida política”. Este material resulta ser muy esclarecedor para

[4] https://amnistia.org.ar/wp-content/uploads

estimar hasta qué punto mensajes maliciosos y agresivos pueden condicionar la participación política de muchas mujeres, independiente de la cantidad de seguidores o la popularidad que tengan. Testimonios como el de la diputada Silvia Lospenatto o de la legisladora Ofelia Fernández ayudan a dimensionar la violencia política por cuestiones de género que sufren mujeres en posiciones de poder y cómo ser víctimas de esto las hace replantearse cosas como el manejo que tienen de sus redes sociales hasta medidas más drásticas como su continuidad en la vida política.

Testimonios similares de figuras públicas tuvieron también mucha repercusión en Reino Unido, al hacerse públicas las denuncias de activistas feministas, políticas miembro del Parlamento y periodistas en el año 2013. El trabajo de la organización End Violence Against Women [44] recoge estos testimonios y otros menos conocidos de ataques y acoso sexual a mujeres utilizando redes sociales.

Megarry [25] puntualiza que una consecuencia de que las mujeres reciban acoso online puede ser el silenciamiento de su vida política. La autora dice que para que sea cierto que las redes sociales son una plataforma segura e igualitaria sus miembros tienen que poder "hablar y expresarse sin miedo de recibir amenazas o acoso". En el trabajo, también analiza la red social Twitter, particularmente lo que sucedió con un hashtag que fue trending topic en 2011: #mencallmethings, donde afirma que el abuso online termina limitando la participación de las mujeres en las redes, restringiendo y silenciando sus voces en debates y foros feministas.

Por todos estos trabajos y testimonios consideramos relevante en esta tesis trabajar y expandir la evidencia disponible sobre el abuso que sufren las mujeres online. Es vasta la literatura que alerta sobre las consecuencias de salud mental y de inhibición a la vida política que sufren las mujeres y los grupos feminizados, resignificando la importancia y la urgencia de que se puedan tomar medidas para morigerar y moderar estos ataques.

3. CONSTRUCCIÓN DE CORPUS

3.1. Recolección de datos

El corpus se construyó utilizando información de Twitter, una de las redes sociales más usada en la región[1]. Amnistía Internacional (2018)[2] agrega: "La naturaleza misma de Twitter fomenta que los usuarios tengan conversaciones públicas y compartan sus opiniones con otros (a menudo, extraños), de modo que podría decirse que los usuarios se benefician más de la plataforma cuando pueden participar abiertamente en los debates." Es por estas razones que esta fuente es particularmente interesante para analizar violencia en redes sociales.

Twitter, a diferencia de otras redes sociales, proveía al momento de realizarse la construcción del corpus[3] el acceso a todos los tweets realizados por usuarios públicos mediante una API, permitiéndonos recolectar tweets, menciones, hashtags y usuarios de una manera muy simple. De esta forma, para construir el monitor accedemos sólo a datos públicos (los perfiles privados no están disponibles) que fueron almacenados de manera segura y anónima en una nube propiedad del Instituto Nacional de las Mujeres, por lo que ninguna información personal fue revelada en la confección del mismo. La información es recolectada y almacenada en tiempo real cada una hora, modificando en la misma frecuencia los resultados del monitor[4].

Los tweets comenzaron a ser recolectados y almacenados desde el 11 de marzo de 2022 de manera continua utilizando la implementación en Python de la librería *tweepy*[5]. Esto se hizo mediante un streamer que permitía buscar tweets en tiempo real bajo los criterios deseados. La ventaja de utilizar un streamer para esta tarea es que permite recolectar información en tiempo real de mensajes que al ser en muchos casos abusivos o insultantes son luego eliminados por los propios usuarios o por los moderadores de la red social. El criterio que se utilizó para recolectar los tweets fue recoger todos aquellos donde se mencionara, etiquetara o respondiera a una de las usuarias de interés del proyecto.

La lista de usuarias seleccionadas fue confeccionada por el Ministerio de Desarrollo Social de Uruguay y PNUD teniendo en cuenta distintos grupos de mujeres de diversas profesiones. Las 180 cuentas seleccionadas fueron a su vez agrupadas según su rol dentro de la escena pública como se muestra en la Tabla 3.1. Es así que todas aquellas mujeres que ejercieran o hayan ejercido alguna profesión en medios de comunicación están englobadas dentro de la categoría de **periodistas** y quienes tuviesen algún rol político o cargo gubernamental quedaron dentro de la categoría de **políticas**. Estas dos categorías fueron las que mayor cantidad de usuarias tenían seguidas por el grupo de **artistas**, donde encontramos actrices, músicas y modelos. Las **comuni-**

[1] `https://gs.statcounter.com/social-media-stats/all/south-america`

[2] `https://amnistia.org.ar/wp-content/uploads/delightful-downloads/2018/05/TOXICTWITTER-report_SP.pdf`

[3] En Marzo de 2023 los términos y condiciones de la plataforma Twitter (actualmente X) cambiaron y restringieron las condiciones de uso y la descarga de tweets.

[4] Esto fue así al menos hasta el cambio de las condiciones de uso.

[5] `https://docs.tweepy.org/en/stable/`

cadoras son un grupo heterogéneo de mujeres que se dedican a difundir contenido y participar activamente de la discusión pública aunque no siempre desde medios de comunicación tradicionales. Por último, las **lideresas** son mujeres activistas o promotoras de causas particulares. Para un mayor detalle de qué cuentas son estas y su clasificación, revisar la Tabla 7.1 del Anexo.

Tipo de usuario	**Cantidad de usuarias observadas**
Periodista	69
Política	41
Artista	27
Comunicadora	27
Lideresa	16
Total usuarias	180

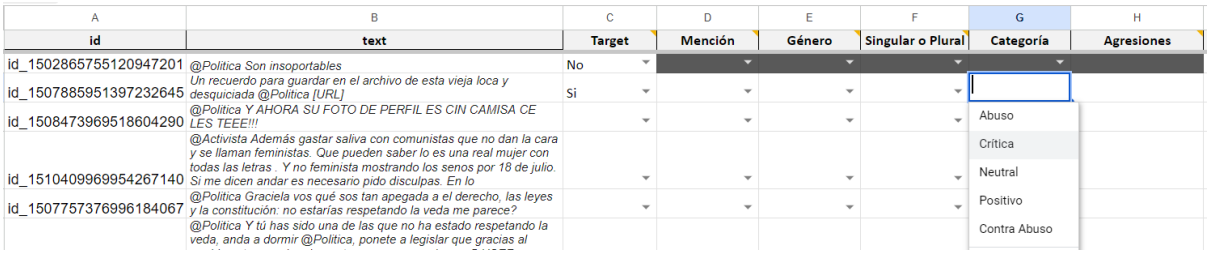
Tab. 3.1: Cantidad de usuarias por tipo

Es necesario aclarar que la lista original de usuarias era en realidad más grande pero por una cuestión de robustez se decidió seleccionar a aquellas que tengan más de 3000 seguidores en sus cuentas. Este número es un piso en general considerado en diversos análisis de este estilo para desestimar aquellas cuentas que no tengan interacción o presencia en redes sociales y que, en definitiva, no tendrán impacto en las métricas finales.

Del dataset construido se filtraron y seleccionaron solamente aquellos mensajes realizados en español, que por tratarse de tweets de usuarios uruguayos se trata específicamente del español rioplatense. Esto hace a la construcción del corpus y posterior entrenamiento del modelo muy valioso, puesto que no hay una gran cantidad de recursos anotados para estas tareas en esta variante del español y por fuera del inglés, haciéndolo una contribución especial de esta tesis.

3.2. Manual de anotaciones

Para que el modelo sea entrenado entonces fue necesario contar con tweets clasificados de manera manual por personas. Si bien el objetivo era identificar tweets abusivos dirigidos a las mujeres seleccionadas, son varias las cosas que se identificaron de cada tweet. Por ello, fue necesario generar una metodología que permita clasificar tweets entre abusivos y no abusivos hacia usuarias monitoreadas. Para esto se planteó una guía de anotaciones donde se listan los aspectos claves que distinguen a los tweets entre estas dos categorías. Esta guía fue compartida y utilizada para entrenar al grupo de anotadores que participó en el proyecto clasificando los tweets que se utilizaron para el entrenamiento. La guía se basó en una utilizada en el marco de un proyecto de la Oficina de Proyectos Especiales (Unidad de Planificación y Control de Gestión) de la Honorable Cámara de Diputados la Nación Argentina. Para la misma, se adaptaron y expandieron previas categorizaciones hechas por Justicia y Género (ELA) [14] y el Alan



	A	B	C	D	E	F	G	H
	id	text	Target	Mención	Género	Singular o Plural	Categoría	Agresiones
1	id_1502865755120947201	@Politica Son insoportables	No					
2	id_1507885951397232645	Un recuerdo para guardar en el archivo de esta vieja loca y desquiciada @Politica [URL]	Si					
3	id_1508473969518604290	@Politica Y AHORA SU FOTO DE PERFIL ES CIN CAMISA CE LES TEEE!!!						
4	id_1510409969954267140	@Activista Además gastar saliva con comunistas que no dan la cara y se llaman feministas. Que pueden saber lo es una real mujer con todas las letras . Y no feminista mostrando los senos por 18 de julio. Si me dicen andar es necesario pido disculpas. En lo						
5	id_1507757376996184067	@Politica Graciela vos qué sos tan apegada a el derecho, las leyes y la constitución: no estarías respetando la veda me parece?						
6		@Politica Y tú has sido una de las que no ha estado respetando la veda, anda a dormir @Politica, ponete a legislar que gracias al						

Abuso
Crítica
Neutral
Positivo
Contra Abuso

Fig. 3.1: Captura del excel recibido por anotadores y anotadoras

Turing Institute en Vidgen et al. [41]. Para que un tweet sea considerado abusivo según las características del proyecto tenía que cumplir con dos condiciones:

1. Estar dirigido hacia una usuaria de la selección.
2. Estar clasificado como abusivo.

Los anotadores recibieron planillas con los tweets a anotar y las columnas que tenían que completar, que esquematizamos en la Figura 3.2. Como se ve en la figura 3.1 las opciones de las respuestas estaban pre-cargadas para evitar que se complete con resultados indeseados. Es decir, dentro de la columna de target y mención sólo podían completar Sí o No y dentro de la columna de categorías sólo tenían las cinco opciones disponibles. Asimismo, dado que si un tweet no tenía target no se debían completar el resto de las columnas, la línea se autocompletaba con un color negro para que el anotador no siga completando el resto de las columnas. Además, los usuarios etiquetados en el tweet fueron reemplazados por tokens: las usuarias observadas tenían un token correspondiente al grupo al que pertenecían (@Politica, @Periodista, etc) y el resto de los usuarios aparecía simplemente como @USER.

De esta manera, los anotadores tuvieron que en primer lugar identificar si el tweet está dirigido hacia una persona o grupo de personas, que de acá en adelante llamaremos que tiene o no un **target**. Específicamente, la instrucción de anotación era la siguiente:

Parte A): Target. De quién o qué está hablando principalmente el tweet.

Dirigido a una/s persona/s: El tweet, ¿está directamente hablando sobre o hacia una persona o grupo chico e identificable de personas?

Aclaración: En caso de estar dirigido hacia un grupo, éste debe ser un grupo chico/concreto/definido, es decir, no una generalización de un colectivo o una abstracción. En algunos casos va a haber más de un target. Utilicen su criterio para decidir cuál es el mensaje principal del tweet y a quién está dirigido.

Ejemplo de tweet	**Explicación**
Sí (personas, grupos de personas)	
FELICITACIONES @politica [URL]	Felicita a una persona

@POLITICA @POLITICA Háganse un tik tok para los infradotados que los siguen	Le habla a dos personas y la acción es hacia ambas ("Háganse un tik tok")
@POLITICA Vos vivís la vida que queremos todo gordo vago…	Le habla directamente a POLITICA ("Vos vivís..")
No (colectivo genérico de personas, leyes, frases, situaciones)	
Así las basuras infrahumanas desechables peronchas utilizan los recursos del Anses @POLITICA @USER @USER	Está dirigido hacia lxs peronistas, un colectivo del cual no podemos identificar individualmente a personas específicas.
@USER @POLITICA Ah pero creditos UVA, me suena a ah pero Macri…	Está hablando de una situación o una frase, no habla de una persona ni está dirigido a una persona.
Es terrible lo que están haciendo. Ahora les pagamos el AVIÓN para que vengan a votar los parásitos extranjeros. Es una LOCURA TODO @POLITICA @USER @USER @USER	Está dirigido al colectivo de extranjeros, no se puede identificar individualmente a una o más personas.

Una vez identificado si el target era sí o no se debía identificar si el mensaje del tweet estaba dirigido hacia alguna de las personas arrobadas que no sea un @USER. Es decir, si el receptor del mensaje era alguna de las mujeres observadas y con los token distintivos. Para ello, se completa la columna de **mención**. Cabe aclarar que de acá en adelante **sólo se anotaron los tweets con target**. Aquellos tweets que no tenían un grupo definido fueron dejados de lado en las siguientes columnas de anotación.

Parte A): Mención. La persona a la que el tweet está principalmente dirigido, está @arrobada y NO es un @user?

Aclaración: En algunos casos no va a ser claro si el tweet está dirigido hacia la persona arrobada o hacia unx tercerx. En esos casos usen el contexto y la intuición para elegir la opción que crean correcta.

Ejemplo de tweet	**Explicación**
Sí tiene mención	
@POLITICX Vos vivís la vida q queremos todo gordo vago…	El tweet está dirigido a @POLITICX
@politica Mentirosa desvergonzada, ustedes solo se cuidan a si mismos.	El tweet está dirigido a @POLITICA
No tiene mención	

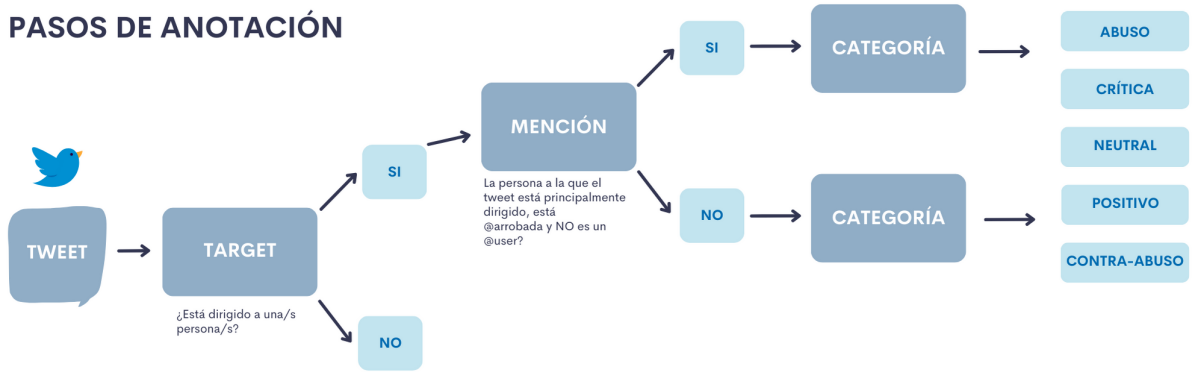


Fig. 3.2: Diagrama de flujo de anotación

La primera vez que me aguanto ver a esa forraaaa de Sarnosa por @POLITICX la quiere callar y no puede, amo	El tweet está principalmente dirigido a "Sarnosa", que es una persona diferente de @POLITICX
@USER @POLITICX Los mismos empresarios nefastos con los que se reunió Alberto 1 semana después de asumir.	El tweet está dirigido a los empresarios, y el único arroba que no es USER es POLITICX

Hasta ahora con estas dos columnas se identificaba si el tweet tenía un target definido y si estaba además dirigido o no hacia alguna usuaria. Como se puede ver en el diagrama de flujo de las anotaciones de la figura 3.2, restaría anotar la intención del tweet, que se clasificó entre cinco categorías posibles de las cuales podían ser: abuso, crítica, neutral, positivo y contra-abuso. El codebook detallado de las categorías de mensajes es el siguiente:

Parte B): Categorías.

Categoría. ¿Cuál de las siguientes opciones describe mejor el tweet? En algunos casos un tweet puede contener más de una categoría. Sin embargo, sólo pueden elegir una. En esos casos, usen su criterio para elegir la categoría que domina el tono del tweet.

Abuso: Expresiones despreciativas, de burla y/o que descalifican o impugnan el valor de las personas, en general relacionados con personalidad y/o cualidades de las personas (incluyendo capacidades mentales, físicas o emocionales). Las amenazas también son un tipo de abuso. Acusaciones de obrar con malas intenciones (por ejemplo jugar sucio) también lo consideramos abuso.

Crítica: Comentarios críticos sobre el accionar, pensamiento o habilidades de las personas. Ser crítico no es una forma "suave" de ser abusivo. Para que un comentario sea crítico no debe contener insultos o lenguaje agresivo.

Neutral: El tweet no incluye comentarios emocionales o sentimentales sobre una persona. En general enuncian hechos o descripciones de eventos.

Positivo: Cuando el tweet apoya, elogia o incluye comentarios positivos hacia a una persona.

Contra abuso: Cuando el tweet denuncia explícitamente el lenguaje abusivo de otras personas para defenderlo. Puede referir de manera directa a un comentario abusivo específico, o hacer una crítica general a la mala forma en la que ciertas personas fueron tratadas.

Abuso
1. @Politica Pero, pedazo de HDP, USTEDES DEJARON DE DAR EL PLAN DE LAS COMPUS, no se puede ser más CÍNICO HDP, si existiera un Dios justo vos y tus compañeros HDP no tendrían que existir en este planeta hace rato. 2. @Politica RIDICULOS!!!!! pongan los pies en la tierra salames, utopicos, no entienden nada , ... no les daa verguenza ?y claaa, si a los que tenemos y JXC tampocoR I D I C U L O S 3. @Politico @USER Bueno, no te escondas abajo de la cama, porque te vamos a encontrar, ex barrabrava. #EllosBardeanElFrenteHace
Crítica
1. @Politica creerá que poniendo puteadas en todos los spots va a ganar votos? 2. @Politica @USER Quien ha perdido la credibilidad? Yo no, trata de hablar por vos misma, tu resentimiento no te deja pensar y unirte a siniestros personajes de la @Politica te están convirtiendo en peor persona. VIVA EL CONICET. 3. @Politica Me aburrieron lo tibios que son con el otro lado de la grieta..., y sí!... @USER tiene razón, y lo mejor que pueden hacer, es escucharlo más.
Neutral

1. En barrio San Nicolás, llevamos adelante la pavimentación a través de fondos financiados por la Secretaría de Obras Públicas de la Nación, a cargo de @Politica
2. 🙌 Este martes 24 de agosto visitará nuestra ciudad la precandidata a Diputada Nacional por el Frente de Tod🌞s, @Politica. #LaVidaQueQueremos [URL]
3. 👉 #JorgeFerrarello Concejal 👉 @USER Senadora Provincial 👉 @Politico Diputado Nacional @USER @USER @USER @USER
Positivo
1. Defender la República NO es para tibios y vos NO lo sos , por eso te voto @Politica 💪🇦🇷
2. Y AÚN ASI PUDO VOLVER DE ESO. FELICITACIONES MARIU @Politica 👏 [URL]
3. Vamos con la lista 501 para reafirmar el rumbo de @Politico y @Politica Vamos con @politica ✌ 🇦🇷 Vamos a #LaVidaQueQueremos [URL] Te banco a muerte @Politico [URL]
Contra Abuso
1. Los que le dicen pelotudo a @Politico no entienden nada
2. 📰 @USER Repudio en redes a la agresión antisemita de @USER a @Politica [URL]
3. Repudiable el ataque de este nazi @Politica. Se ve que están preocupadxs por la izquierda [URL]

Tab. 3.2: Ejemplos de tweets categorizados.

Como también se identifica en la figura 3.1, se anotaron dos variables adicionales y en las que no vamos a entrar en detalle: género, cantidad y agresiones. Las primeras dos se referían a la persona a quien el tweet estaba dirigido y que estaba en la mención. Es decir, si el mensaje estaba dirigido a una de las usuarias monitoreadas, entonces el tweet sería de género femenino y singular y si eran varias usuarias femenino y plural. Asimismo, si estaba dirigido a otros usuarios y el contexto del tweet permitía inferir su género, dado que el español permite en algunos casos conjugaciones que nos dejen inferir géneros, también debían completarlo. Estas columnas no fueron luego utilizadas para el análisis.

Ese no fue el caso de la columna de Agresiones, donde los anotadores debían anotar los insultos y agresiones como se encontraban escritas literalmente en el caso de que la categoría fuera abuso y que fue luego utilizado en el monitor. Con esta última columna se enriqueció un diccionario de keywords de agresiones con el que se contaba de proyectos anteriores y que se utilizó para mostrar en el monitor los insultos presentes en los mensajes. Esos insultos además fueron categorizados de acuerdo a una conceptualización desarrollada en conjunto entre el PNUD e InMujeres, utilizando como base un proyecto realizado en Argentina a cargo de la Oficina de Proyectos Especiales (Unidad de Planificación y Control de Gestión) de la Honorable Cámara de Diputados de la Nación.

3.3. Proceso de anotación

Con los mencionados criterios de anotación se organizaron seis rondas de anotaciones con dos grupos de dos anotadores cada uno que clasificaron un total de 9000 tweets. Los anotadores eran uruguayos y se postularon mediante una búsqueda laboral que difundió PNUD en Uruguay. Se trataba de cuatro jóvenes (tres mujeres y un varón), tres de ellas estudiantes de Sociología y Antropología. Se tuvo una reunión inicial con todos en la que se les presentó el proyecto y se les explicó la tarea a realizar. Se dividieron a los cuatro anotadores en dos grupos que se mantuvieron todas las rondas y que anotaban tweets en paralelo. Es decir, el grupo 1 anotaba los mismos tweets y luego los comparaba entre sí y el grupo 2 anotaba otros tweets y luego los comparaba entre sí. Los horarios de trabajo y tiempos fueron avisados con anticipación en la búsqueda, así como el pago recibido.

El proceso de anotación se realizó entonces individualmente (aunque los anotadores tenían contacto y podían hacerse consultas) y el chequeo fue en parejas, extendiéndose semanalmente de jueves a lunes. Los jueves por la mañana (9 am) cada pareja recibía la muestra de tweets en el excel correspondiente con las columnas preseteadas para anotar y debían realizar las anotaciones de manera individual. Tenían tiempo hasta el domingo a la mañana (9 am) para debe devolver el excel completo e individual. Con eso, se les devolvían las diferencias en las anotaciones a cada pareja en otro archivo. Durante ese día y hasta el día siguiente (lunes) tenían tiempo para reunirse y consensuar las diferencias, entregando el último archivo corregido. Si aún hubieran diferencias debían dejar la columna vacía.

La cantidad de tweets de cada ronda fue escalonada considerando que a medida que se anota una mayor cantidad de tweets el tiempo que lleva revisarlos es cada vez menor debido a la práctica. De esta manera, a los anotadores se les dio un entrenamiento sobre cómo clasificar tweets y una ronda de prueba donde anotaron 50 tweets. Luego se llevó adelante una ronda de 1000 tweets, dos de 1600 tweets y una de 1800 tweets, cantidades que se dividieron de manera equitativa entre los dos grupos de anotadores. Debido a que el beneficio de contar con una mayor cantidad de tweets clasificados era grande se llevaron a cabo dos rondas más de 1500 tweets cada una. Cabe agregar que luego de cada ronda de anotación se tenían reuniones con los anotadores para disipar posibles dudas y que el contacto con ellos era continuo durante la anotación mediante un canal de Slack.

Para seleccionar los tweets a anotar se tomó una parte de los mismos de manera aleatoria de la recolección con que se contaba hasta el momento (el 30 % de la cantidad a anotar). Para llegar a la cantidad restante se utilizó la técnica de **active learning** que permite identificar cuáles son los textos que el algoritmo tiene mayor dificultad en identificar (Settles [36]). Active learning es una técnica aplicada en machine learning según la cual las observaciones a clasificar y anotar no se seleccionan aleatoriamente dentro del pool de datos sin anotar sino que se seleccionan aquellos que al modelo le cuesta más diferenciar. Para esto, luego de cada ronda se agregan las nuevas anotaciones a los datos anteriores y se vuelve a calcular el active learning en base a eso. Esto permite que, eligiendo aquellas observaciones más difíciles de reconocer para el modelo, este vaya mejorando de manera más rápida que seleccionándolas aleatoriamente. Es particularmente útil cuando no se cuenta con la posibilidad de tener muchas anotaciones o su costo es más bien alto.

Para llevar a cabo esta técnica se debe definir en primer lugar qué medida tomar para calcular si un tweet es más fácil de clasificar para el modelo o no. Para este proyecto se evaluaron dos distintas: menor confianza y entropía. Lo que se hizo fue tomar una muestra de 1300 tweets uruguayos recolectados y utilizando el clasificador entrenado calcular la probabilidad de que el tweet sea abusivo o no abusivo. Con esto para cada caso se evaluaron las dos medidas.

En el de menor confianza se seleccionan prioritariamente aquellos casos en que el modelo está menos seguro de cómo clasificar la etiqueta. Es decir que si un tweet tenía una probabilidad de 0.9 de ser considerado abusivo y otro una de 0.6 se quedaría con este último porque es aquel sobre el que menos seguridad a la hora de clasificar tiene. Como en el caso de 0.9 es bastante claro que el modelo se inclinó por la etiqueta de que el tweet sea abusivo entonces no es necesario agregarlo a la ronda de anotación. En el segundo caso se observa la entropía de los valores de probabilidad brindados, siendo p_k las probabilidades dadas.

$$S = -\sum_{k=1}^{n}(p_k - \log(p_k)) \tag{3.1}$$

Luego de contar con este cálculo para todos los tweets de la muestra se seleccionan aquellos cuya entropía sea más alta porque nos indica mayor diferencia entre valores y por ende mayor incertidumbre a la hora de que el modelo clasifique. Habiendo calculado estas dos medidas para los primeros tweets recolectados, se decidió elegir a la entropía porque proveía mayor variabilidad en los valores obtenidos, permitiendo así distinguir más a un tweet de otro y entendiendo que esto daba un ordenamiento más preciso.

En la primera ronda entonces, teniendo datos de otro proyecto anotados previamente para Argentina y su correspondiente modelo, se utilizó esto sobre los tweets uruguayos recolectados hasta el momento y se seleccionaron los 680 primeros cuyo valor de entropía era más alto. Este procedimiento se repitió ronda a ronda luego de agregar los nuevos tweets anotados y reentrenar el modelo con estas nuevas observaciones. En la segunda y tercera ronda se seleccionaron 1100 tweets con active learning y en la cuarta 1240. Las últimas dos rondas consistieron solamente

Medida	Ronda 1	Ronda 2	Ronda 3	Ronda 4
Target	85 %	70 %	75 %	90 %
Menciones	70 %	50 %	60 %	75 %
Categorías	65 %	60 %	60 %	75 %

Tab. 3.3: Porcentaje de aciertos en las gold labels

Ronda	Grupos de anotadores	Tweets totales	Active Learning	Random Sample	Gold Standard
1	2	1000	680	300	20
2	2	1600	1100	480	20
3	2	1600	1100	480	20
4	2	1800	1240	540	20
5	1	1500	1500		
6	1	1500	1500		

Tab. 3.4: Distribución de tweets en rondas de entrenamiento

en tweets con active learning.

Por último, se agregaron 10 tweets por grupo de anotadores que habían sido previamente clasificados por expertos del proyecto y que fueron usados como gold standard para testear el entendimiento de anotadores sobre la tarea. En la siguiente tabla 3.3 se puede ver el agreement que tuvieron los anotadores en los tweets con gold labels a lo largo de las rondas, calculado simplemente como el porcentaje de aciertos. Además de servir como métricas finales para medir qué tan bien se entendió la tarea de anotación, fueron indicadores muy valiosos para identificar malentendidos durante el proceso y corregirlos.

A las cuatro rondas planeadas originalmente se le agregaron dos rondas más que no estaban planificadas porque se consideró que había aún una ganancia relevante en sumar más anotaciones al entrenamiento del modelo. Esto se detalla en la siguiente sección donde se especifican las métricas obtenidas por rondas de anotación. Las rondas entonces se dividieron de la siguiente manera, como muestra la Tabla 3.4.

3.4. Revisión de anotaciones

Es importante para la construcción de un dataset evaluar su calidad a medida que se produce la anotación. En tareas tan difíciles de determinar por la subjetividad que conllevan, que los anotadores mantengan una relativa concordancia al realizar su trabajo individual es rele-

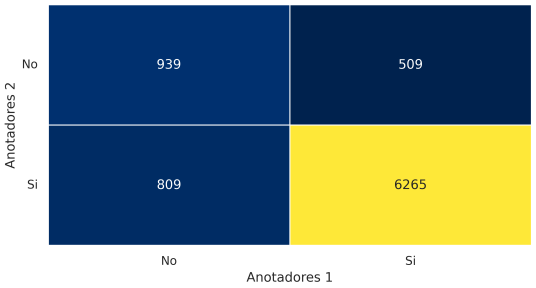


Fig. 3.3: Matriz de acuerdo de target

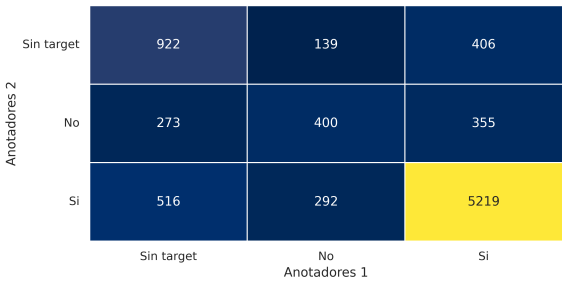


Fig. 3.4: Matriz de acuerdo de mención

vante. Además, también es esperable encontrar una mejora en el acuerdo entre anotadores por el aprendizaje mismo de la tarea a medida que avanzan en la misma. Es decir, si en la primera ronda no estaba tan claro cómo clasificar entre target, mención o de categoría, es esperable que a medida que avancen las rondas de anotación esto se corrija y por ende haya menos discrepancias entre lo que dos anotadores opinan respecto a un texto.

Para esto revisamos en primer lugar las matrices de acuerdo de las anotaciones. Es decir, veremos cuántas veces el anotador 1 coincidió en la etiqueta utilizada con el anotador 2 (de ambos grupos). Para estas matrices consideraremos 8522 tweets de los 9000 anotados en total, puesto que en algunos casos hay algún anotador que no finalizó su tarea y por ende no tiene sentido la comparación[6]. En el primer item a anotar -el target del tweet- vemos en la figura 3.3 que hay una gran mayoría de tweets anotados con concordancia (7204 de 8522). En la segunda matriz del etiquetado de la mención de la figura 3.4 se ve que estas discrepancias iniciales entre el target traen diferencias en lo que se considera con mención o no, dado que la anotación se de de manera encadenada. Inclusive, las diferencias en estas anotaciones provienen más por los tweets que anteriormente se anotaron sin target que por discrepancias en si tienen o no mención (un total de 647 tweets).

Quizás la más interesante y rica sea la matriz de acuerdo de las anotaciones de las categorías. En la figura 3.5 observamos que en la diagonal hay bastante acuerdo entre categorías en anotadores. Hay 5265 de 8522 tweets anotados en los que los anotadores coinciden en la categoría de discurso a la que pertenecen. Considerando la dificultad de la tarea de elegir entre todas las categorías, la coincidencia es razonable. Inclusive, en las discrepancias hallamos sentido en los resultados. Los tweets anotados como crítica y abuso son los que más confusión tienen entre sí: en muchos casos la distinción entre estas dos categorías puede ser confusa o subjetiva de cada persona. De todas maneras, es preponderante el acuerdo frente al desacuerdo.

Una categoría en la que parece haber muy poco acuerdo -o casi nulo- es en contra abuso. Muy pocos mensajes fueron anotados dentro de esta categoría, pero los que lo hicieron no fueron mayoritariamente en acuerdo. En efecto, hay más desacuerdo por tweets que un anotador consideró contra-abuso y el otro no y viceversa que en los 6 únicos tweets que ambos acordaron que era contra-abuso.

[6] Esto no significa que los tweets no fueron anotados. En primer lugar los anotadores etiquetaban por separado (acuerdo que estamos analizando en esta sección) y luego en conjunto decidían la etiqueta final. Este último proceso sí se realizó para los 9000 tweets, asegurando la tarea.

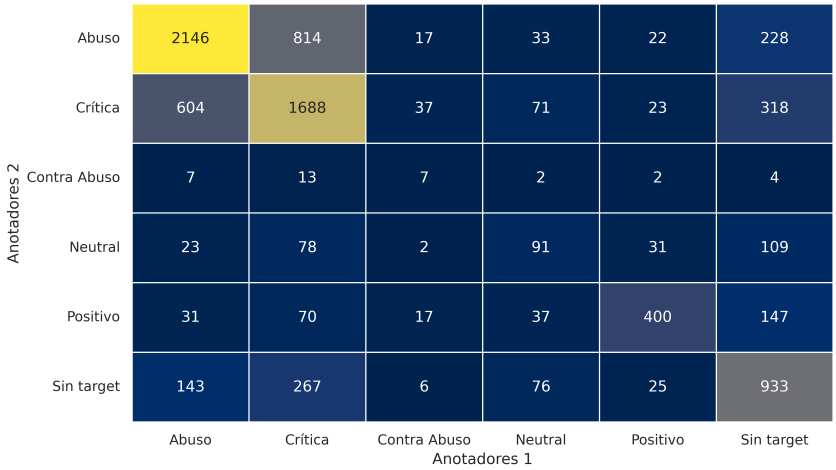


Fig. 3.5: Matriz de acuerdo de las categorías

Variable	Ronda 1	Ronda 2	Ronda 3	Ronda 4	Ronda 5	Ronda 6
Target	35 %	51 %	48 %	35 %	58 %	59 %
Mención	40 %	59 %	52 %	52 %	30 %	42 %
Categorías	27 %	42 %	46 %	48 %	66 %	64 %

Tab. 3.5: Krippendorff alpha por ronda

3.5. Acuerdo entre anotadores

Además de esta herramienta más visual, para analizar la robustez y la concordancia de las anotaciones también calculamos el alpha de Krippendorff (Krippendorff [16].) El alpha de Krippendorff es una medida estadística muy utilizada para evaluar el *agreement* entre anotadores. Es atractiva, especialmente para nuestro caso, porque se puede usar con cualquier número de anotadores, de etiquetas y maneja los datos incompletos. Esto último va a permitir que evaluemos en conjunto las tres variables a anotar aún cuando más de la mitad haya sido anotada por el grupo 1 y el resto por el grupo 2. No importa si hay dos anotadores "faltantes", la medida está implementada para que podamos ingresar la matriz con nuestro dataset entero y hallar los resultados. Para su implementación utilizamos la librería *krippendorff* implementada en Python[7].

Como se muestra en la Tabla 3.5, el acuerdo entre anotadores aumenta a medida que avanzan las rondas. Considerando que para estas métricas un acuerdo entre un 41 % y 60 % es moderado podemos decir que los niveles de acuerdo son razonables. La primera ronda es bastante pobre pero esto es esperable dada la curva de aprendizaje que suelen tener estas tareas. Las siguientes mejoran bastante, con excepciones claro como el target en la Ronda 4 y la mención en la Ronda 5. Pero el aprendizaje en la tarea y por ende el mayor acuerdo es notorio.

[7] https://pypi.org/project/krippendorff/

Variable	Grupo	Alpha
Grupo 1	Target	48 %
	Mención	53 %
	Categorías	46 %
Grupo 2	Target	51 %
	Mención	55 %
	Categorías	42 %

Tab. 3.6: Krippendorff alpha entre anotadores por grupo y ronda

Los alpha anteriores estaban calculados para todas las anotaciones pero también es importante comprobar que todos los grupos de anotación tengan desempeños similares. Es decir, que no haya un grupo mucho más atrasado que el otro. Por esto calculamos el acuerdo por grupos de anotación en la Tabla 3.6. Allí vemos que los alpha son muy similares entre ambos grupos, si bien el Grupo 2 suele tener mayores desacuerdos relativos en las categorías.

3.6. Estructura del dataset construido

Como se mencionó en el apartado anterior, luego de anotar los mensajes individualmente los anotadores debían juntarse, resolver las discrepancias y llegar a las anotaciones finales. Para esto contaban con la posibilidad de hacer consultas al equipo si no podían acordar alguna discrepancia. Las anotaciones finales fueron revisadas por el equipo de expertos, particularmente en aquellos casos donde hubieran discrepancias, y de considerar que estaba mal anotado algunas fueron corregidas.

Para el caso de esta tesis, por los bajos valores de acuerdo encontrados en la primera ronda de anotación y por tratarse también de una ronde de aprendizaje decidimos no utilizar la primera ronda para el entrenamiento final. De esta manera, contamos con un total de 8.000 tweets anotados, cuya gran mayoría tiene un target (el 83 %), Tabla 3.7. De estos tweets se tienen aquellos que tienen una mención y su categoría. Dentro de estas categorías un 44 % de las mismas es abuso, seguido por un 39 % de crítica, lo que habla de que la toxicidad de los tweets recolectados es alta, como se ve en la Tabla 3.8. La categoría de Contra Abuso no va a ser considerada de acá en adelante dado que con sólo 26 observaciones en toda la muestra no hay información suficiente para que el modelo se entrene en esa clasificación.

Variable	Cantidad de tweets	Porcentaje %
Target		
Sí	6645	83,06 %
No	1355	16,94 %
Total	8000	100 %
Mención		
Si	5655	85,10 %
No	990	14,90 %
Total	6645	100 %

Tab. 3.7: Distribución de tweets según Target y Mención

Categoría	Cantidad de tweets	Porcentaje (%)
Abuso	2909	43,77 %
Crítica	2598	39,10 %
Positivo	706	10,62 %
Neutral	406	6,11 %
Contra Abuso	26	0,39 %
Total	6645	100 %

Tab. 3.8: Distribución de tweets según Categorías

De un análisis descriptivo preliminar que se hizo durante el entrenamiento del modelo (tomando tweets recolectados hasta el 11 de junio de 2022) encontramos que el comportamiento y los mensajes que recibían cada tipo de usuaria en Twitter variaba según grupo como se observa en la Tabla 3.9. Por ejemplo, las periodistas eran la mayor cantidad de usuarias que se recolectaron y sin embargo generan menos interacción que las políticas o las lideresas. Esto nos indica un comportamiento distinto en las distintas usuarias y en la forma de relacionarse con sus seguidores.

Esta información también puede ser observada en la página del monitor publicado[8]. Allí, en la tercera pantalla del dashboard están expresados los porcentajes de tweets violentos que recibió cada grupo de usuarias como se ve en la Figura 3.6. El grupo de usuarias con mayor cantidad de mensajes recibidos es el de políticas (en efecto constituyen el 66 % de todos los tweets recolectados) y asimismo el que mayor porcentaje de tweets abusivos recibió, seguido de las Periodistas. Este detalle se puede ver por grupo de usuaria en la segunda slide del dashboard

[8] `https://www.violenciadigitalmujeres.uy/`

Tipo de usuaria	Cantidad de menciones	Ratio entre menciones y usuarias por tipo
Política	209.919	5.120
Periodista	52.819	765
Lideresa	13.044	815
Comunicadora	3.153	117
Artista	1.875	69

Tab. 3.9: Cantidad tweets recibidos por usuarias por tipo y ratio con cantidad de usuarias

3.7, donde además del porcentaje están registrados las cantidades de tweets recibidos además de si en esos había alguno de los insultos listados en la Tabla y en las categorizaciones hechas.

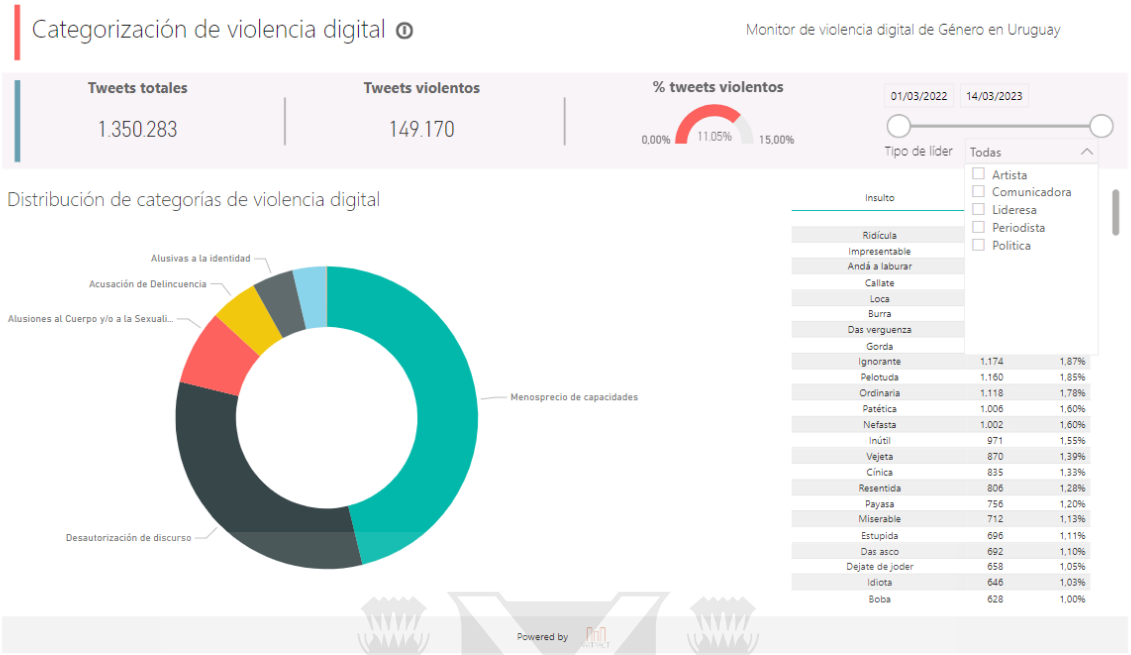


Fig. 3.6: Captura del Monitor de violencia digital de género en Uruguay. Fuente: `https://www.violenciadigitalmujeres.uy/`

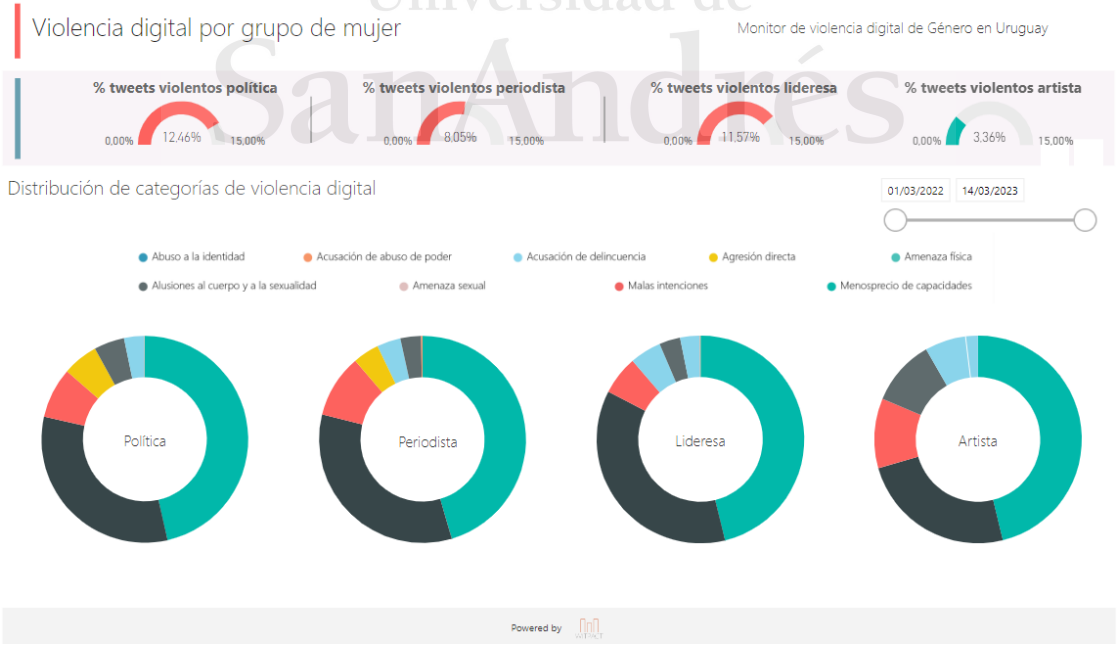


Fig. 3.7: Captura del Monitor de violencia digital de género en Uruguay. Fuente: `https://www.violenciadigitalmujeres.uy/`

4. METODOLOGÍA DE ENTRENAMIENTO

Como se mencionó anteriormente, el objetivo del proyecto de monitor era poder registrar y monitorear la violencia que sufren las mujeres en redes, en este caso figuras relevantes de la política, medios o coyuntura uruguaya. Por esto es que se decidió entrenar un modelo que sea capaz de distinguir si un tweet es abusivo hacia la figura o figuras etiquetadas o no. El modelo utilizado en ese caso está entrenado sobre BETO[1] (Cañete et al. [4]). BETO es un transformer de tipo BERT entrenado por un equipo del Departamento de Ciencias de la Computación de la Universidad de Chile sobre un gran corpus de textos en español.

Para el trabajo realizado en esta tesis se decidió comparar la performance del modelo original utilizado con otros modelos de lenguaje. Es decir, se realizó un benchmarking tomando como referencia a Roberta[2] (Gutiérrez-Fandiño et al. [10]), Bertin[3] (De la Rosa et al. [7]), Electra[4] y Robertuito[5] (Pérez et al. [31]).

4.1. Cambios en la función de pérdida

Dada la naturaleza de la anotación y de las predicciones esperadas, se cambió la función de pérdida por una customizada. Esto se hizo porque a la hora de anotar los tweets, si un tweet era considerado sin target el resto de las columnas quedaba incompleta. Es decir, sólo se anotó si el tweet tenía una mención dirigida a una de las referentas de interés y la categoría del mismo si el tweet tenía un target definido. Es así que calcular la pérdida de las variables de mención y categoría cuando el tweet no tiene target no tiene mucho sentido.

Para sortear esta dificultad se modificó a la función de pérdida para que sólo tenga en cuenta al target cuando este sea igual a 0 (es decir, sin target) y no al resto de las variables, y para que cuando el target sea 1 considere la predicción de todas las variables. Es así que si tenemos una instancia con los valores (t,m,cat) donde t y m sólo pueden tomar valores binarios de 0 o 1 y cat valores del 0 al 4 por cada categoría y hago una predicción ($\hat{t}$, $\hat{m}$, $\hat{cat}$), la pérdida se calculó de la siguiente manera:

Si t = 0:

$$L = L_t \tag{4.1}$$

Si t = 1:

$$L = L_t + L_m + L_{cat} \tag{4.2}$$

[1] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[2] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne
[3] https://huggingface.co/bertin-project/bertin-roberta-base-spanish
[4] https://huggingface.co/mrm8488/electricidad-base-discriminator
[5] https://huggingface.co/pysentimiento/robertuito-base-uncased

Esto también se puede escribir como:

$$L = t \cdot (L_t + L_m + L_{\text{cat}}) + (1 - t) \cdot L_t \tag{4.3}$$

Para el cálculo de la pérdida del target y de la mención se utiliza una función de entropía binaria. Esta función mide la disimilitud entre la distribución de probabilidad prevista y los valores binarios reales del objetivo. Por otro lado, para la evaluación de la pérdida de la variable de categorías se utilizó una función de entropía cruzada. La función de entropía cruzada penaliza al modelo cuando hace predicciones incorrectas e intenta así que aprenda las probabilidades correctas.

5. RESULTADOS

En esta sección vamos a revisar los resultados obtenidos del entrenamiento. En primer lugar repasaremos los resultados obtenidos del benchmarking de modelos sobre el set de validación. Luego analizaremos los resultados del entrenamiento final sobre los datos de testeo y por último analizaremos los errores cometidos por el modelo, esbozando posibles explicaciones e identificando vetas sobre las que seguir trabajando.

5.1. Resultados del Benchmarking

Para comenzar a trabajar con los datos y entrenar los modelos dividimos a los 8000 tweets clasificados en sets de entrenamiento, testeo y de validación. En el grupo de entrenamiento quedaron 5040 tweets (un 63 % del total), en el de testeo 2000 tweets (un 25 %) y en el de validación 960 (el 12 %). Si bien se habían anotado un total de 9000 tweets se decidió dejar la primera ronda de anotaciones de lado dado que el acuerdo entre anotadores no era muy alto.

Una vez divididos los datos se tomaron los modelos del benchmarking, se corrieron cinco veces cada uno y se guardaron sus resultados para evaluarlos en conjunto y disipando posibles outliers propios de la inicialización aleatoria del entrenamiento. Toda la evaluación del benchmarking se realizó exclusivamente sobre el set de validación. Las métricas utilizadas fueron la Precisión, el Recall y el F1. Se calculó cada una para las distintas métricas a predecir: el target, la mención y las categorías.

La Precisión va a medir la proporción de Verdaderos Positivos (VP) -aquellas observaciones donde el modelo predice correctamente la clase positiva - sobre el total de las observaciones positivas que predijo. De esta manera mide la calidad del modelo, evaluando el porcentaje de resultados que identificó como positivos y realmente lo eran.

$$\text{Precision} = \frac{VP}{VP + FP} \tag{5.1}$$

El Recall o exhaustividad va a medir la proporción de VP sobre todos aquellos que son verdaderamente positivos, es decir los VP y los Falsos Negativos (aquellos que eran positivos pero que no identificó como tales). Nos va a indicar la cantidad de casos positivos que es capaz de identificar sobre el total de casos positivos.

$$\text{Recall} = \frac{VP}{VP + FN} \tag{5.2}$$

Modelo	Precisión		Recall		F1-score	
	Promedio	Desvío est.	Promedio	Desvío est.	Promedio	Desvío est.
Beto	88,6 %	<0,01	93,9 %	0,01	91,2 %	<0,01
Electra	88,7 %	0,02	93,8 %	0,02	91,1 %	<0,01
Bertin	87,7 %	0,01	94,1 %	0,02	90,7 %	<0,01
Roberta	89,8 %	<0,01	94,1 %	0,01	91,9 %	<0,01
Robertuito	91,5 %	<0,01	93,8 %	0,01	92,6 %	<0,01

Tab. 5.1: Benchmark de Target

El F1-score combina precisión y recall en un solo indicador. Asumiendo que nos importa de igual forma la precisión y la exhaustividad, calcula la media armónica de la precisión y el recall.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.3}$$

Para observar los resultados, a mostramos las tablas con el promedio de estas tres métricas y los desvíos para los cinco modelos analizados, de las cinco corridas que tuvieron cada uno.

Tanto los resultados del target (Tabla 5.1) y de las menciones (Tabla 5.2) son altos. Como vimos en la distribución de observaciones, un 83 % de los tweets de la muestra anotada eran target y de esos un 85 % tenía una mención. De esta manera, la tarea de clasificación es un poco más sencilla y por ende obtenemos niveles de precisión y de recall muy altos. Es especialmente en este último donde los modelos muestran una mejor performance: son muy buenos acertando en la cantidad de targets y menciones del total. Respecto a los modelos, la performance es muy buena en todos. Electra y Bertin tienen medias similares al resto pero con menos estabilidad en esos resultados al mirar sus desvíos estándar. Robertuiti se destaca como el mejor modelo en términos de precisión, seguido de Roberta. Ambos con poca variabilidad en sus resultados.

La predicción de las categorías demostró ser un poco más difícil que la del target y mención como se puede ver en la Tabla 5.3. En esta tarea se nota más la heterogeneidad de los modelos y la dificultad que encuentran con algunas clases. En la evaluación de las categorías, Robertuito se destaca ampliamente por ser el modelo con mejor rendimiento. Si observamos la Tabla 5.3 es quien suele tener las métricas más altas y además las más estables. Porque si bien en la predicción de la categoría Neutral Electra tiene mayor precisión, el Recall es muy bajo y además con un desvío estándar muy alto, lo que hace a Robertuito una mejor opción. En términos de performance el segundo mejor es Roberta.

Respecto a qué tan buenos son los modelos para cada una de estas categorías, podemos afirmar que tienen mejor desempeño al predecir Abuso, Positivos, Críticas y Neutrales, en ese

Modelo	Precisión		Recall		F1-score	
	Promedio	Desvío est.	Promedio	Desvío est.	Promedio	Desvío est.
Beto	88,6 %	<0,01	93,6 %	0,01	90,9 %	0,01
Electra	88,3 %	0,02	91,7 %	0,02	89,9 %	0,01
Bertin	88,6 %	0,01	94,3 %	0,02	90,2 %	0,01
Roberta	90,3 %	<0,01	92,9 %	0,01	91,5 %	0,01
Robertuito	90,9 %	<0,01	94,1 %	0,01	92,5 %	<0,01

Tab. 5.2: Benchmark de Mención

orden. Robertuito muestra una precisión promedio de 71,9 % y un Recall de 71,4 % al predecir Abuso, con desvíos de 0,02. Considerando la dificultad de la tarea del problema multiclase se trata de valores altos. Es destacable que también muestre una elevada performance en encontrar aquellos tweets positivos, siendo que esta categoría no contaba con muchas observaciones pero que evidentemente son distintas al resto. El peor desempeño se da en la categoría Neutral, aquella con menos observaciones entre estas cuatro (406). No sólo tiene promedios más bajos sino que los desvíos de todas las corridas son más altos también. De todas maneras, con valores de precisión y recall de 54,5 % y 56,1 % respectivamente en Robertuito, se trata de una tarea mejor que el azar.

Dado que el F1 es una métrica que se calcula por clase, para la evaluación de la variable de categorías es necesario agregarlos de alguna manera para evaluar el comportamiento de la variable en conjunto. Además, como la cantidad de observaciones de categorías están desbalanceadas por clase, se calcula el F1 Ponderado que atiende este problema. El F1 Ponderado se calcula tomando la media de todas las puntuaciones F1 por clase ponderadas por el soporte de cada una (el número de instancias verdaderas para cada etiqueta).

$$\text{F1 Ponderado} = \frac{\sum_{i=1}^{N}(\text{F1-score}_i \cdot \text{support}_i)}{\sum_{i=1}^{N}\text{support}_i} \tag{5.4}$$

A continuación calculamos el F1 ponderado para la variable de categorías y sacamos el promedio y desvío para los cinco modelos. Obviamente las tendencias se verifican con los resultados anteriores siendo los mejores modelos Robertuito, Roberta y Beto. El F1 Ponderado nos ayuda a verificar que el modelo está entrenado correctamente para encontrar a las categorías en general. Siendo que la de peor performance era también la de menos observaciones (Neutral), el F1 Ponderado permite dar cuenta de esto y que no deprima el puntaje general a la hora de evaluar el desempeño del modelo. Como podemos comprobar en la Tabla **??**, Robertuito es el modelo con mejores resultados, seguido de Roberta y Beto. Además de tener un alto promedio, Robertuito es también muy estable, con un desvío estándar por debajo de 0,01.

Modelo	Precisión		Recall		F1-score	
	Promedio	Desvío est.	Promedio	Desvío est.	Promedio	Desvío est.
Neutral						
Beto	54,4 %	0,09	48,0 %	0,02	50,6 %	0,04
Electra	65,4 %	0,04	35,5 %	0,15	43,5 %	0,16
Bertin	57,8 %	0,14	52,7 %	0,05	53,7 %	0,03
Roberta	53,7 %	0,08	49,2 %	0,02	51,0 %	0,03
Robertuito	54,5 %	0,11	56,1 %	0,05	54,2 %	0,02
Positivo						
Beto	74,0 %	0,04	50,9 %	0,02	60,2 %	0,01
Electra	69,1 %	0,06	42,2 %	0,04	52,4 %	0,05
Bertin	59,6 %	0,03	48,6 %	0,05	53,3 %	0,03
Roberta	76,3 %	0,04	55,6 %	0,02	64,3 %	0,02
Robertuito	80,3 %	0,01	64,3 %	0,01	71,4 %	0,01
Crítica						
Beto	55,6 %	0,01	63,7 %	0,01	59,4 %	0,01
Electra	54,5 %	0,01	63,0 %	0,07	58,3 %	0,04
Bertin	56,4 %	0,01	59,5 %	0,02	57,9 %	0,01
Roberta	60,0 %	0,01	68,4 %	0,01	63,9 %	0,01
Robertuito	62,9 %	0,01	66,3 %	0,02	64,5 %	0,01
Abuso						
Beto	65,4 %	0,01	62,7 %	0,02	64,0 %	0,01
Electra	64,3 %	0,03	65,1 %	0,05	64,4 %	0,01
Bertin	63,7 %	0,02	63,8 %	0,01	63,8 %	0,01
Roberta	69,3 %	0,01	66,0 %	0,02	67,6 %	0,01
Robertuito	71,9 %	0,02	71,4 %	0,02	71,6 %	0,01

Tab. 5.3: Benchmark de Categorías

Modelo	F1 Ponderado para categorías	
	Promedio	Desvío est.
Beto	61,0 %	0,01
Electra	59,5 %	0,03
Bertin	59,8 %	0,01
Roberta	64,8 %	0,01
Robertuito	67,7 %	<0,01

Tab. 5.4: F1 Ponderado

Variable	Precisión	Recall	F1	F1 Pond.
Target	91,5 %	94,9 %	93,2 %	
Mención	88,2 %	97,9 %	92,8 %	
Abuso	70,6 %	71,9 %	71,3 %	67,4 %
Crítica	61,0 %	67,7 %	64,2 %	
Positivo	79,4 %	62,3 %	69,8 %	
Neutral	78,2 %	43,9 %	56,2 %	

Tab. 5.5: Métricas finales con Robertuito

En conclusión y tras haber analizado los resultados y el rendimiento de los cinco modelos, podemos afirmar que Robertuito es el que mejor desempeño muestra. Además, siendo comparado con Roberta que suele secundarlo en las métricas, Robertuito es más eficiente en los tiempos de entrenamiento, motivo que se suma a la elección del modelo como final. De esta manera, en la siguiente sección evaluaremos al modelo final de Robertuito sobre el set de testeo.

5.2. Resultados finales de entrenamiento

Una vez seleccionado Robertuito como el mejor modelo para entrenar, guardamos el último entrenamiento y lo utilizamos para predecir el dataset de testeo. En esta sección entonces detallaremos los resultados de entrenamiento, en tablas similares a las revisadas anteriormente y en matrices de confusión.

Como se puede observar, confirmamos los altos resultados que se habían visto sobre el dataset de evaluación: tanto en target como en mención las métricas finales son muy altas. Esto también se puede ver en las Figuras 5.1 y 5.2. En este sentido son varias las hipótesis que podemos esbozar que expliquen métricas por encima del 90 %.

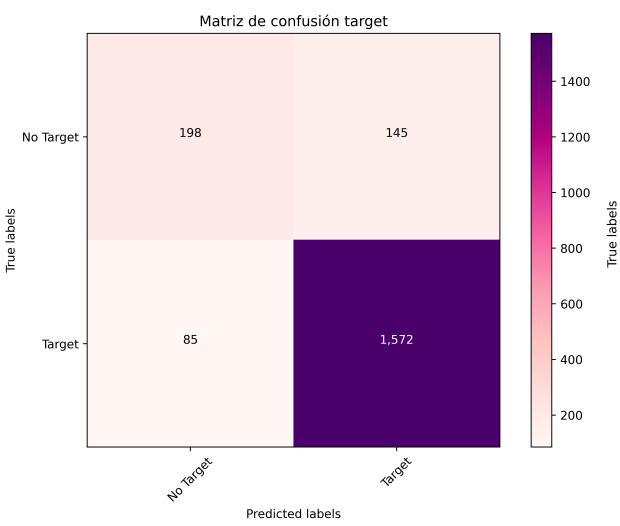


Fig. 5.1: Matriz de confusión de la variable target

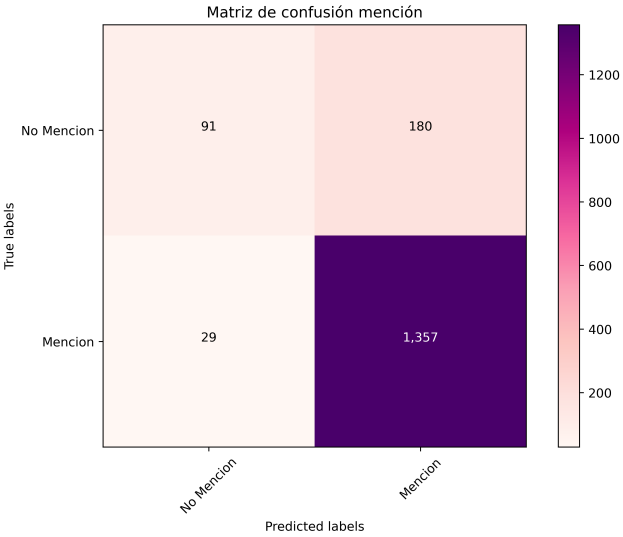


Fig. 5.2: Matriz de confusión de la variable mención

Variable	**Especificidad**
Target	57,8 %
Mención	33,58 %
Abuso	76,52 %
Crítica	71,6 %
Positivo	98,3 %
Neutral	99,1 %

Tab. 5.6: Métricas finales de especificidad

Primero, cabe mencionar que la distribución de los datos puede estar determinando los resultados de la tarea, siendo que un 83,06 % de los tweets son targeteados y de estos un 85,1 % tienen algún tipo de mención. Es decir, probabilísticamente es más sencillo acertar en que un tweet tiene un target o una mención, dado que en general la respuesta va a ser "Sí." Estos datos desbalanceados, con una performance mejor en las respuestas positivas se puede observar en las matrices de confusión. Los falsos negativos son tan sólo el 5,1 % de los tweets de target (85 de 2000) mientras que los falsos positivos son un 42,3 % (145 de 2000). Es decir, al modelo le cuesta mucho más predecir aquellos tweets que no tienen target que aquellos que sí, en los que casi nunca se equivoca. Una métrica para resumir este razonamiento es la especificidad, es decir la proporción de predicciones negativas verdaderas entre todas las instancias negativas reales.

$$\text{Especificidad} = \frac{TN}{TN + FP} \tag{5.5}$$

La especificidad tanto del target como de la mención son menores a la precisión y al recall,

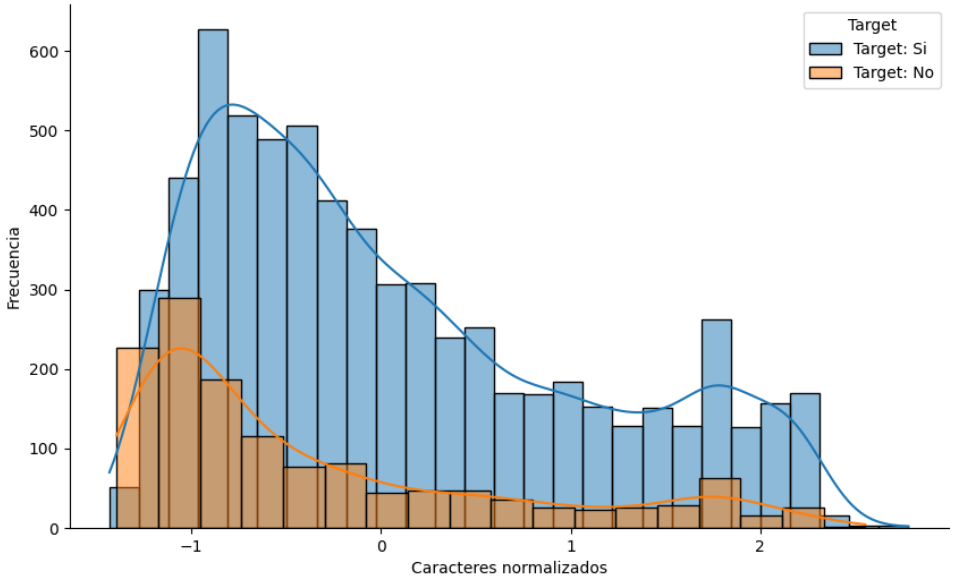


Fig. 5.3: Distribución de caracteres en los tweets según el target

como se puede observar en la Tabla 5.6. La especificidad del target es del 57,8 % y la de la mención es 33,58 %, lo que indica que el modelo es menos eficaz a la hora de identificar correctamente los casos negativos ("No").

En segundo lugar, haciendo una inspección manual de los tweets podemos encontrar algunos patrones que también pueden estar provocando las métricas altas en el target y en la mención. Hay una gran cantidad de tweets de muy corta extensión, que son muchas veces una palabra o simplemente algunos emojis, que son clasificados como sin target y sin mención. En particular, 129 de los 198 tweets sin target que fueron correctamente clasificados tienen menos de 50 caracteres. En efecto, si miramos la distribución de los caracteres totales de los tweets según si son target o no observamos una diferencia como se ve en la Tabla 5.3 donde se normalizaron los caracteres. Si calculamos la mediana de los tweets sin target encontramos que es de 53 caracteres y la de los tweets con target de 99, casi el doble. Además, el 25 % de los tweets sin target tienen 27 caracteres o menos. A continuación, en la Tabla 5.7 se detallan algunos ejemplos de tweets sin target que fueron correctamente clasificados y cuya longitud es corta.

Nro.	Tweet
1	@Politica Y?
2	@Politica 🤮🤮
3	@Politica señales
4	@Politica ¿Qué es esto?
5	@Politica Nooooo
6	@Politica Ponele
7	@Politica No droga?
8	@Politica .
9	@Politica 😂 😂🙏🙏🙏
10	@Periodista @USER Ojo con los polvos

Tab. 5.7: Tweets sin target predichos correctamente

Las dos variables además están muy relacionadas entre sí. Entre las anotaciones de target y de mención se observa una correlación de Pearson significativa del 68 %. Es decir, cuando un tweet es clasificado con target suele también ser clasificado con mención. Por eso también ambas variables tienen resultados similares. En síntesis, los mencionados factores podrían estar contribuyendo a que las métricas de resultados de target y de mención sean tan altas. De todas maneras, admitimos que puede haber algún otro factor que cause este buen desempeño y que no estemos registrando.

Los resultados de las categorías también son buenos, con un F1 Ponderado de las cuatro categorías de 67,4 %, como se ve en la Tabla 5.5. La predicción de tweets abusivos resultó ser muy buena y balanceada en la precisión y en el recall, con un 70,6 % y un 71,9 % en cada una. Crítica resultó tener la menor precisión con 61 % y con un recall más elevado que le permite tener un F1 de 64,2 %. Los positivos tienen la mayor precisión con casi un 80 %, aunque un menor recall de 62,3 %. Por último, la categoría neutral, con 406 observaciones en total y 98 en el dataset de testeo tiene una precisión muy alta encima del 78 % pero un recall bajo de 43,9 %. En general, la variable de categorías tiene buenos resultados, estando el F1 Ponderado en un 67,4 % para todas las categorías.

Las categorías en las que más se confunde el modelo son entre crítica y abuso como se puede ver en la Figura 5.4. En esto consideramos que hay dos factores en juego. El primero es que la distinción entre los tweets críticos y los tweets abusivos puede muchas veces no ser tan clara, como detallamos en la siguiente sección de Análisis de error. Inclusive, estas eran las categorías donde los anotadores más se confundían entre sí. La dificultad de la tarea, por la subjetividad de la misma, hace que sea difícil tanto para el modelo como para los anotadores distinguir entre ambas.

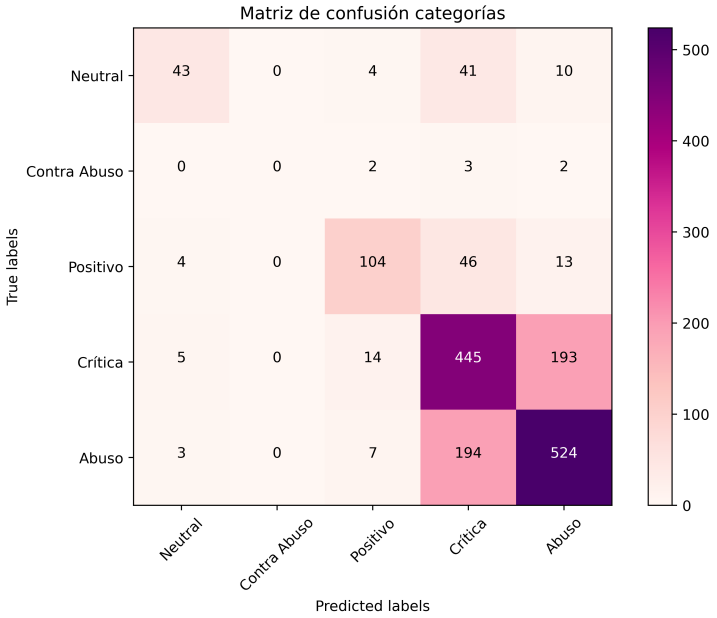


Fig. 5.4: Matriz de confusión según la variable categoría

Además de esto, en la matriz de confusión también identificamos otro factor importante: en general, el modelo se equivoca más en identificar los tweets críticos del resto. En efecto, la precisión de crítica es la menor de todas las categorías: 61 %. Esta confusión del modelo con tweets que fueron clasificados como positivos o neutrales baja -estrepitosamente para el caso neutral- el recall de estas categorías. Hay sólo 43 tweets neutrales que el modelo clasifica bien mientras que hay otros 41 que son neutrales y clasifica como crítica. En resumen, la categoría de crítica parece confundir bastante al modelo y afectar las métricas del resto, a diferencia de los abusivos donde la única confusión aparece con los de crítica. La diferencia entre los tweets de abuso y los positivos y los neutrales es más tajante en comparación, lo cual es de todas maneras un buen chequeo de calidad para el entrenamiento.

5.3. Análisis de error

En esta sección vamos a hacer un análisis cualitativo de los tweets donde el modelo se equivocó. Particularmente, nos interesa analizar la confusión entre crítica y abuso, que identificamos desde las anotaciones, y también crítica frente al resto de las categorías dado que fue bastante problemática.

5.3.1. Falsos Positivos y Falsos Negativos: crítica y abuso

En primer lugar vamos a revisar los Falsos Positivos de la categoría crítica. Es decir, aquellos tweets anotados como críticos pero predichos como abuso. Para mostrar los ejemplos elaboramos la Tabla 5.8 con el mensaje y el score que el modelo puso a que el tweet sea crítico o abusivo. Es decir, si el modelo le puso a la categoría abuso un score muy alto significa que "tenía pocas dudas" sobre que la categoría del tweet era abuso. Por el contrario, si el score es bajo y el de crítica o

algún otro es más alto, indica que no estaba tan seguro de si un tweet era abusivo, crítico u otro.

De estos tweets podemos obtener algunas lecciones. En primer lugar, que la tarea de anotación es difícil y subjetiva y por ende los criterios por los que para una persona un tweet es crítico o abusivo pueden cambiar. En este sentido, algunos tweets que el grupo de anotación consideró críticos parecen más bien abusivos, como el quinto de la Tabla 5.8. No es un tweet con insultos pero sin embargo descalifica a la política arrobada y cuestiona el ejercicio de su profesión, además de que incluye emojis de enojo. En este sentido, puede que el tweet no esté bien clasificado como crítico. Este componente de error que puede venir de errores de anotación es irreducible y se debe en parte a la subjetividad de la misma tarea.

De la misma manera, la falta de contexto también dificulta la tarea. Por ejemplo, el primer tweet de la Tabla 5.8 fue clasificado como crítico y quizás esto es correcto, pero también puede darse que sea un mensaje gordofóbico dirigido a una persona a la que se ataca constantemente por este tema. Si bien que los anotadores hayan sido del país del que se recolectaron los tweets, cosa que mitigaría un poco este problema al estar ellos inmersos en el panorama de la política o actualidad local, hay contexto de la conversación que no se pueden rescatar con un solo mensaje.

En cualquier caso, la mayoría de los casos en los que el modelo se equivoca parecen no ser groseros y tener sentido. Es decir, encontrar un mensaje clasificado como el sexto de la Tabla 5.8 como abuso parece estar bastante bien, a pesar de que no estaba anotado así. Inclusive, en aquellos casos dudosos entre crítica y abuso como el cuarto o el octavo de 5.8 tampoco es un error grave porque también es discutible su clasificación. La tarea de distinción entre crítica y abuso es difícil inclusive para un ser humano.

Esto no significa que las anotaciones hayan sido malas, porque todas tuvieron el chequeo cruzado de dos anotadores. Es decir, deben haber visto los tweets individualmente y luego haber estado de acuerdo en eso o de lo contrario discutir la respuesta. Un posible motivo por el que aparezcan estos tweets aparentemente mal clasificados era porque no eran lo "suficientemente agresivos" como otros que vieron los anotadores. Por ejemplo, mensajes como "@Politica Una pobre idiota infeliz. Ya sabe lo que es la bondiola ?.", "@Politica Un asco esta vieja cucaracha" o "@Lideresa Pedazo de yegua es una persona con problemas mentales" tienen un nivel de agresión tan alto que pueden matizar o confundir la tarea de anotar. Si un anotador se encuentra con estos mensajes de agresión y luego con uno como el cuarto o el sexto de la Tabla 5.8 puede parecerle que ambos mensajes no pertenecen en la misma categoría y que este último sea más crítica que abuso.

También, hay casos bien anotados en los que el modelo se equivoca, como el segundo de la Tabla 5.8. Este tweet en nuestra opinión está correctamente anotado como crítico y sin embargo mal clasificado como abusivo.

Mirando ahora aquellos tweets abusivos que son erróneamente clasificados como crítica (Falsos Negativos) comprobamos que se repiten patrones similares al caso anterior. En primer lugar hay algunos mensajes donde la distinción entre abuso y crítica es discutible. Por ejemplo, el primer tweet de la Tabla 5.9 podría ser un tweet crítico porque tiene comentarios sobre el accionar

Nro	Tweets	Score Crítica	Score Abuso
1	@Politica Jugatela con unos sándwiches	0,36	0,38
2	@Politica Nadie violó más la veda que Ud, señora	0,47	0,48
3	@Politica Tas al dope	0,38	0,48
4	@Politica @USER @USER Espero q ud se ponga al frente de los mugrientos y no de manija sentadita. Ya lo hicieron con los pobres estúpidos e ignorantes q uds mandaron al frente. No profundice la grieta	0,47	0,48
5	@Politica @USER Mira el roto se ríe del descosido q usted lo único q hace es escribir barbaridades lamentable y vergonzoso q sea senadora 😡 😡 😡	0,10	0,89
6	@Politica @USER Atrevida, criticaste al Plan Ceibal porque lo hizo otro gobierno llamaste a los niños idiotas digitales y no resistís un archivo. Da pena tu decadencia intelectual.	0,08	0,90
7	@Politica @USER @USER @Politica @USER Pintarla como quiera. Están dándole nuestro dinero a femibolches parasitas.	0,07	0,91
8	@Politica Sinverguenza	0,06	0,92

Tab. 5.8: Falsos Positivos: muestra de tweets anotados como críticos y clasificados como abusivos

de la política a la que se dirige pero también podría tomarse como una crítica al ejercicio de su profesión. En este sentido, múltiples políticas mujeres han mencionado que el asedio constante de mensajes maliciosos por redes sociales les hace replantearse su profesión y por ende agraviar su carrera política, como los ejemplos de la Diputada Silvia Lospennato y de la legisladora Ofelia Fernández mencionados en la sección anterior. Es decir, un tweet puede ser agraviante y por ende abusivo, lo que hace la tarea de clasificación difícil, como en el caso mencionado. El modelo entonces considera al tweet como crítico y lo cierto es que es discutible este límite entre abuso y crítica. Otro mensaje al que le pasa algo similar a este puede ser el séptimo de la Tabla 5.9.

En algunos mensajes, al modelo le cuesta más clasificar entre abuso y crítica como al cuarto de la Tabla 5.9, donde el score entre abuso y crítica es casi el mismo. A priori, la línea divisoria puede ser difícil por no ser directamente agraviante, pero también parece estar juzgando la salud mental de las personas lo que constituye una expresión descalificativa. El modelo puede haberlo entendido como un mensaje más bien neutral y no con el sarcasmo que podría tener.

Nuevamente en estos casos nos encontramos con errores esperables por parte del modelo y de la anotación. Por ejemplo, el segundo ejemplo de la Tabla 5.9 está clasificado como abuso probablemente porque descalifica utilizando “estupidez” pero el modelo no identifica eso y lo

considera crítica. Del mismo modo, el sexto fue anotado como abuso pero no es tan claro si es una expresión despreciativa o que impugne el valor de una persona. Hacia ambos lados, los errores parecen razonables. De todas maneras, hay ejemplos que están simplemente mal anotados, como el quinto y el octavo de la mencionada tabla. En ambos casos no parecen ser comentarios abusivos que estén menospreciando o denigrando sino que son más bien críticos de la labor y ejecución política.

Nro	Tweets	Score Crítica	Score Abuso
1	@Politica @USER Unas ganas de ser como vos cuando sea grande.. y ganar plata SOLO por publicar bobadas TODO el día en tw 🙏🏽🤘🏽	0,4	0,12
2	@Periodista Nunca escuché tanta estupidez junta jajajan	0,41	0,28
3	Decime @Politica , que sentís cuando ves esto, tremenda felicidad supongo, difícil que esa piedra que tenés por corazón siquiera se mueva. Estás de regalo en esa posición, aprovecha estos dos años, por que por suerte no se te verá más la cara en ese recinto.	0,44	0,31
4	@Politica señora es hora de ir a consulta	0,46	0,45
5	@Politica Lo que hace el FA y las posteriores declaraciones de Sanchez son patoterismo parlamentario	0,81	0,14
6	@Politica Y esto fue presidente....	0,82	0,08
7	@Politica Estabas vos de directora...ya ahí es la demostración cabal de que la educación pública venía mal..	0,84	0,12
8	@Politica @USER Tiene una lista de todos los familiares que colocó su coalición? O precisa que se la proporcionemos?	0,85	0,04

Tab. 5.9: Falsos Negativos: muestra de tweets anotados como abusivos y clasificados como críticos

5.3.2. Neutrales predichos como crítica

Ahora bien, es interesante también ver por qué hay tantos errores entre tweets anotados como neutrales y clasificados como críticos. Esta confusión puede ser más grave que la de abuso y crítica por cómo afectó al recall de la categoría neutral y porque a priori parecen categorías con una distinción más clara.

Entre los ejemplos encontramos algunos casos de tweets que parecen estar bien anotados como neutrales y en los que el modelo se equivoca clasificándolos como crítica. Por ejemplo, el quinto tweet de la Tabla 5.10 tiene un registro de diario o noticia que debería estar siendo clasificados como neutrales y el modelo los asigna en crítica. Como se puede apreciar en el score, es un caso en el que el modelo no tenía tanta certeza sobre qué categoría era la correcta. Simi-

larmente, el séptimo ejemplo tampoco parece crítico, es sólo un comentario sobre una historia. Sin embargo, el modelo lo clasifica como crítico, puede que sea por la presencia de la palabra "presunto" para referirse a un colono: quizás de los tweets de entrenamiento interpreta la presunción de cosas sobre las personas como una manera de poner en duda ciertas capacidades o realidades y por ende tiene una connotación negativa.

Luego hay otros ejemplos cuyo contenido sin contexto puede generar más dudas. Por ejemplo en la Tabla 5.10 el cuarto tweet puede estar escondiendo un mensaje sarcástico o negativo, porque se puede asumir que ese padre no era bueno. O el tercer mensaje nos hace preguntarnos si ese ejemplo era bueno o malo. A priori no lo sabemos y por ende estaría bien que sean tweets neutrales. Lo mismo podría estar sucediendo con el segundo tweet. Este caso es un poco más dudoso porque puede tener un mensaje gordofóbico, como discutíamos en un tweet similar anterior.

Nro	**Tweets**	**Score Neutral**	**Score Crítica**
1	@Periodista Noooo vos solo saltas por el diente lopez... Es el unico jugador que pediste desde que estaba tabarez... Pero ta si no lo citan es oor algo mamita...no debe ser muy buen compañero... Ahh y el piojo es de LIVERPOOL	0,06	0,63
2	@Politica Dona tus sándwichitos	0,17	0,58
3	@Politica @USER Usted es un claro ejemplo	0,07	0,45
4	@Politica Cada dia mas parecida al padre	0,08	0,45
5	La senadora nacionalista, @Politica, fue elegida como Presidenta de la Comisión de Educación del Senado. A quien le enviamos un especial saludo y éxito en su gestión. [URL]	0,32	0,35
6	@Periodista Creo que tendrás que reclamar a la justicia de Nicaragua a Daniel Ortega a ver qué te dicen. Aquí vamos a tener que apretar las clavijas y poner orden a todo este libertinaje de los Nefastos 15 años [URL]	0,03	0,84
7	@Periodista Patricia: hay una investigación sobre un presunto colono que es senador de la república... Estaría bueno para un libro	0,04	0,8
8	@Politica @USER No soy fanático ni resentido, pero luego de escuchar al presidente entre morisqueta y morisqueta, me doy cuenta que sólo gobierna para los empresarios y sus privilegios! VOTO SI!!!!	0,06	0,75

Tab. 5.10: Muestra de tweets anotados como neutrales y clasificados como críticos

En el caso del primer y del último tweet, ambos parecen correctamente clasificados como

críticos. En el primero, el dirigirse hacia la Periodista como “mamita” tiene un tono peyorativo y que intenta disminuir a la destinataria de alguna manera. Por ende, tratarlo de neutral puede ser incorrecto. El último, por otro lado, también parece expresar una posición crítica y para nada neutral. El mensaje tiene un descontento hacia el presidente a quien acusa de hacer “morisquetas” y además expresa una opinión a partir de eso. En estos casos, podríamos tomar al score del modelo como correcto y pensar en que fue un error de anotación.

5.3.3. Positivos predichos como crítica

En los tweets positivos clasificados como crítica encontramos varios mal clasificados por el modelo, que son positivos. Particularmente, en todos los tweets analizados, parece haber una mayoría de errores y tratarse de ejemplos positivos mal clasificados. En la Tabla 5.11 pusimos algunos ejemplos de esto junto con otros que podrían ser errores de anotación.

Es así que del primero al tercer tweet de la Tabla 5.11 son casos de tweets positivos mal clasificados. Esto puede que se deba a algunas palabras dentro de los mensajes que traigan una connotación negativa, como “politólogos sesgados” en el primer tweet, o la mención a la injusticia y la ilegalidad del segundo tweet. De la misma manera, en el tercero la mención a “los comunistas” con terror también podría confundir al modelo.

Algunos tweets pueden ser un poco más confusos y quizás no es tan claro si son positivos. Un ejemplo es el cuarto tweet, donde no queda claro si el mensaje es sarcástico o un halago, en parte por la falta de contexto. El séptimo mensaje definitivamente no parece ser un mensaje positivo que elogie a alguien sino más bien está expresando una posición crítica. El quinto tampoco parece positivo porque más bien parece ser un ejemplo de contra abuso, al dar un halágo a Tabaré Vasquez mediante la crítica a Lacalle. Similarmente, el octavo tweet habla a favor de la política arrobada pero mediante la crítica a otras personas que interactúan con ella.

En resumen, en el caso de los tweets positivos hay una gran mayoría de errores del modelo que no clasifica bien tweets efectivamente positivos. Una posibilidad es que los seleccionados son mensajes con pocos emojis y los tweets positivos suelen tener varios de ellos, en aplausos y caritas felices, confundiendo de esta manera al modelo. Hay otros pocos casos que mostramos acá donde los mensajes son confundidos por críticos porque intentan decir algo bueno o halagar a alguien mediante la crítica hacia otra persona.

Nro	Tweets	Score Positivo	Score Crítica
1	@Politica Excelente, bien clara. Seguro los politólogos sesgados no la leyeron o no la seguirán.	0,38	0,41
2	Gracias Sen. Graciela Bianchi @Politica por recibir a FAMILIARES de PRISIONEROS POLÍTICOS, escuchar lo que teníamos que decir y los detalles de la injusticia, además de la ilegalidad de que se siga sin respetar la ley Caducidad 2 veces votada por la mayoría del pueblo [URL]	0,19	0,44
3	@Politica @USER SOS mi ídola ,te tienen terror los comunistas viven pendiente de tus comentarios y más aún cuando se empiezan a destapar estafas y fraudes que involucran a muchos ídolos de ustedes!!	0,19	0,45
4	Soy yo o @Periodista y @USER son separadas al nacer??? 🧐 🧐 Igualitas [URL]	0,04	0,47
5	@Politica @USER Mientras más gobiernan Lacalle Arbeleche, más grande tiene que ser el monumento a Tabaré Vasquez Danilo Astori.	0,03	0,87
6	@Politica @USER No les des explicaciones	0,04	0,87
7	@Politica @USER Los presos políticos hoy, son militares. Vergüenza nacional de tenerlos presos mediante juicios truchos.	0,03	0,83
8	En @USER está @Politica, hay momentos que 2 de los periodistas cuando Graciela habla como corresponde del MLN, pareciera que hasta se ofenden y les duele, es demasiado obvio, disimulen che. Se nota tanto en las interrupciones y las re preguntas dependiendo el invitado.	0,02	0,79

Tab. 5.11: Muestra de tweets anotados como positivos y clasificados como críticos

5.4. Discusiones y trabajo futuro

Tras haber realizado el análisis del corpus construido, de las anotaciones y de los resultados del modelo identificamos algunos aprendizajes y puntos pendientes que quedarán a desarrollar en trabajos futuros. En cuanto a las anotaciones, algo a mejorar puede ser la recolección de datos no completamente necesarios y que luego no fueron utilizados, como si el tweet estaba dirigido a un género o a una persona o un grupo. Estos datos serían útiles en caso de hacer análisis comparativos entre abuso y críticas hacia miembros de diversos géneros, o colectivos. A la hora de generar un corpus es ambicioso intentar recolectar todos estos datos pensando en su utilidad a futuro pero también se debe considerar que puede traer consecuencias negativas

como hacer la anotación más dificultosa, cuando la tarea por sí sola ya lo es.

Otro tema que apareció en la revisión de las anotaciones y en el análisis de error es el efecto negativo que tiene la falta de contexto a la hora de anotar un tweet. Hay mensajes cuyo contenido puede ser interpretado de maneras distintas. Por ejemplo, en la Tabla 5.8 el primer tweet y en la Tabla 5.10 el segundo tweet, pueden verse como más bien neutrales o no agresivos o estar escondiendo una crítica o mensaje gordofóbido. El modelo no puede incorporar esto y los anotadores podrán hacerlo en determinados contextos, si se trata de una figura pública muy conocida y atacada por esto podrían intuir que la están atacando a ella, o si el suceso ocurrió recientemente. Pero al haber enmascarado los usuarios y no tener el hilo completo de tweets en muchos casos puede faltar contexto para determinar la verdadera naturaleza del mensaje.

En este sentido, la discusión sobre el uso del contexto está abierta. Pérez et al. [30] evidencia que agregando información de contexto mejora el rendimiento de la detección del discurso de odio para predicción binaria y multietiqueta, incrementando su Macro F1 en 4,2 y 5,5 puntos, respectivamente. Esto pone en relevancia que la falta de contexto puede disminuir la performance de estas tareas y también abre futuras investigaciones acerca de cuál es la mejor manera para introducir esto. Pavlopoulos et al. [28] también evalúa la inclusión del contexto en las anotaciones y comprueban que el mismo tiene un efecto estadísticamente significativo en las anotaciones. El contexto a la hora de anotar entonces en puede amplificar o atenuar la toxicidad percibida de los mensajes, aunque el efecto en la anotación solo lo observan en una parte pequeña de su dataset. De todas maneras, no encuentran efectos sobre el entrenamiento de los clasificadores de toxicidad respecto a aquellos que usan anotaciones sin contexto.

6. CONCLUSIONES

En esta tesis, hemos abordado la tarea de detección de abuso hacia mujeres en posiciones de poder y exposición, aportando una nueva herramienta cuantitativa para la medición del fenómeno desarrollada especialmente con español rioplatense. Para hacer esto, se construyó un corpus específico utilizando tweets recolectados donde se arrobaba a este grupo de mujeres uruguayas. Se realizó un trabajo conceptual importante para definir el objetivo y los límites del estudio que se iba a llevar a cabo, teniendo en cuenta la literatura existente en términos de violencia digital y trabajos similares realizados en otras latitudes.

Una vez superada esta etapa, construimos el corpus con anotadores locales de Uruguay, teniendo en cuenta el componente cultural y coyuntural que pueden tener discursos abusivos o críticos, esto le dio un gran sello de calidad a la anotación. A partir de este dataset se realizaron experimentos de clasificación buscando obtener un modelo que pueda identificar si un mensaje estaba dirigido hacia una usuaria y su contenido abusivo. Como resultado, nos encontramos con métricas alentadoras y más altas de lo esperada debido a la naturaleza subjetiva de la tarea, tanto a la hora de anotar como al momento de entrenar un transformer para clasificar.

En el camino identificamos algunos puntos pendientes a seguir profundizando en futuros trabajos en la materia. Particularmente, en relación con las anotaciones identificamos que hay varias subjetividades a la hora de anotar y corregir las mismas por la falta de contexto. Este tema además se replica a la hora del entrenamiento del problema con estos datos dado que el modelo tampoco cuenta con contexto. Si bien gran parte de los mensajes más tajantemente abusivos y denigrantes son fácilmente clasificables por anotadores y detectables por los algoritmos actuales, hay un subconjunto de estos mensajes que necesitan del contexto completo en el que se inserta para entender su sentimiento. Para mejorar o pensar en posibles soluciones a futuro también destacamos la dificultad de separar el lenguaje crítico o sarcástico de uno tóxico o denigrante. Si bien esto último es difícil y subjetivo inclusive para dos seres humanos, la tarea es aún difícil para una máquina.

Si bien los resultados y las métricas halladas para el entrenamiento de RoBERTuito con nuestros datos son muy altos y más considerando la subjetividad de la tarea, aún así hay algunos mensajes donde el clasificador falla en detectar comentarios abiertamente insultantes. Inclusive teniendo algunos de ellos insultos explícitos. Lo cierto es que la detección de abuso es una tarea difícil y no siempre perfecta pero que contribuye ampliamente a visibilizar la violencia digital hacia las mujeres. Porque al utilizar estas herramientas podemos contrastar las historias y testimonios personales con números y ejemplos concretos. A partir de la evidencia, creemos que se pueden tomar mejores decisiones para morigerar estos fenómenos, tanto desde el lado de la moderación en las redes sociales, como por posibles decisiones de política pública que se puedan tomar para lidiar con esta situación.

De todas maneras, todo este trabajo y la lectura de la literatura previa nos recuerda lo importante que es que sigamos trabajando y perfeccionando estos métodos de detección de

abuso. Aún con falencias como las identificadas en el análisis de error, estamos frente a herramientas poderosas para captar lenguaje insultante, tóxico y discriminatorio en redes sociales. Es importante seguir trabajando en el mejoramiento de estos algoritmos para que puedan ser potencialmente usados en la moderación de agresiones. En este sentido, lo comentado respecto a la dificultad de que el modelo entienda ciertas metáforas o lenguaje más críptico es una tarea en la que se puede continuar afinando los resultados.

Fomentar y perfeccionar el uso de estas herramientas es muy relevante. En un contexto en que por minuto los mensajes que se proliferan en redes sociales son miles y en todos los idiomas, la necesidad de contar con herramientas de Inteligencia Artificial que puedan alivianar la tarea es fundamental. Es imposible contar con humanos que se dediquen a hacer esta tarea manualmente, además de que también estaríamos frente a sesgos propios e incontrolables para nosotros. Por esto es muy importante continuar con estas investigaciones y con su difusión, explicando conceptos que para un público no especializado puede ser de difícil comprensión. Es crucial que estas herramientas sean vistas como algo positivo y con un impacto social constructivo para que sean aceptadas por la sociedad. En este sentido, trabajar en disminuir posibles sesgos en la recolección de datos y en aumentar la cantidad de modelos creados especialmente para una localización es importante. A lo largo de esta tesis se ha detallado cómo se recolectaron datos en español rioplatense y atendiendo al contexto particular de esta lengua y su contexto cultural, así como se han utilizado conceptualizaciones y clasificaciones elaboradas localmente. Participar en la elaboración de estos modelos y poder fortalecerlos desde el Sur Global es un gran paso que debemos dar para adoptar correctamente estas tecnologías.

7. ANEXO

Nro.	Nombre	Twitter	Tipo
1	Carolina Cosse	CosseCarolina	Politica
2	Beatriz Argimón	beatrizargimon	Politica
3	Laura Raffo	lauraraffo	Politica
4	Graciela Bianchi	gbianchi404	Politica
5	Veronica Maria Alonso Montaño	veronica_alonso	Politica
6	Constanza Beatriz Moreira Viñas	Constanza_FA	Politica
7	Graciela Villar	GVillar_uy	Politica
8	Liliam Kechichian	likechichian	Politica
9	Cristina Lustemberg	LustembergC	Politica
10	Adela Dubra	adeladubra	Politica
11	Micaela Melgar	MicaMelgar	Politica
12	Lucia Topolansky Saavedra	TopolanskyLucia	Politica
13	Gloria Rodriguez	gloriasaravista	Politica
14	Carolina Ache Batle	CarolinaAche	Politica
15	Mónica Bottero Tovagliare	MnicaBottero	Politica
16	Carmen Sanguinetti Masjuan	Carmensangui	Politica
17	Dra. Irene Renée Moreira Fernández	IreneMoreiraUy	Politica
18	Bettiana Díaz	bettianadiazrey	Politica
19	Verónica Leticia Mato Correa	Veronica_Mato	Politica
20	Carol Aviaga	CarolAviaga	Politica
21	Matilde Antía	matiantia	Politica
22	Cecilia Bottino Fiuri	ceciliabottino	Politica
23	Patricia Soria Palacios	PatriciaSoria_	Politica
24	Susana Pecoy Santoro	susypecoy	Politica
25	Ana María Olivera Pessano	AnaO1001	Politica
26	Carmen Asiaín	casiain1	Politica
27	Patricia Natalia Kramer Belo	patakramer	Politica
28	Silvia Nane Vincon	SilviaNaneFA	Politica
29	María Eugenia Rosello	marurosello	Politica
30	Lilián Galán	liliangalan1	Politica
31	Silvana María Pissano	silvanapissano	Politica
32	Fernanda Araújo	faraujo404	Politica
33	Carmen Niria Tort González	CarmenTort3000	Politica
34	Susana Pereyra	SPereyra609	Politica
35	Macarena Rubio Fernández	MacarenaRubioF	Politica
36	Cecilia Sena Bargas	chechisena	Politica
37	Valentina Dos Santos	vds2525	Politica
38	Laura Andrea Tabarez Martinez	lautabarez	Politica
39	Lucía Etcheverry Lima	LuciaEtchever11	Politica
40	Sandra Lazo	sandralazo2	Politica
41	Elsa Capillera	ECapillera	Politica
42	Emilia Diaz	emiliadays	Lideresa
43	Susana Andrade	MaeSusanaAndrad	Lideresa
44	Valeria Ripol	Valeriaripoll3	Lideresa

Nro.	Nombre	Twitter	Tipo
45	Patricia Rodríguez	PatriciaRSifpom	Lideresa
46	Selva Anderoli	SelvaAndreoli	Lideresa
47	Patricia González	PataGonzaV	Lideresa
48	Andrea Tuana	TuanaAndrea	Lideresa
49	Lilián Abracinskas	labracinskas	Lideresa
50	Karina Nuñez Rodriguez	KarinaNEZ16	Lideresa
51	Soledad gonzalez	solsticia_uy	Lideresa
52	Ana Ines Martines	AnaInesMartinez	Periodista
53	Blanca Rodríguez	blancarodgon	Periodista
54	Patricida Madrid	PatriciaJMadrid	Periodista
55	Cecilia Bonino	ceciliabonino	Periodista
56	Iliana Da Silva	dasilvailiana	Periodista
57	Carolina García	caro_garcia10	Periodista
58	Lucía Brocal	luciabrocal	Periodista
59	Sofía Rodríguez	Sofiarayl	Periodista
60	Viviana Ruggiero	ViviRuggiero	Periodista
61	Ximena Barbé	ximenabarbe	Periodista
62	Camila Ciblis	Camila_Cibils	Periodista
63	Andrea Tabárez	AndreaTabarez	Periodista
64	Carina Novarese	carinanovarese	Periodista
65	Carolina Domínguez	carodominguezuy	Periodista
66	María Noel Marrone	NoleMarroneok	Periodista
67	Soledad Ortega	SoleOrtega_Uy	Periodista
68	Ana Laura Pérez	perezanalaura	Periodista
69	Tania Tabárez	taniauy	Periodista
70	Silvia Pérez	sperezprensa	Periodista
71	Valeria Superchi	ValeriaSuperchi	Periodista
72	Georgina Mayo	georgina_mayo_	Periodista
73	Natalia Uval	nataliauval	Periodista
74	Ana Laura Román	AnaLauraRoman	Periodista
75	Valeria Tanco	valeriatanco	Periodista
76	Ana Matyszczyk	analiamaty	Periodista
77	Nadia Fumeiro	nadiafumeiro	Periodista
78	Diana Piñeyro	dianapineyro1	Periodista
79	Paula Scorza	pscorza	Periodista
80	Malena Castaldi	mcastaldi_uy	Periodista
81	Mónica Willengton	MonicaBW	Periodista
82	Silvia Techera	stecheraoficial	Periodista
83	Romina Andrioli	andrioliromina	Periodista
84	Marcela Dobal	marcedobal	Periodista
85	Paula Barquet	PaulaBarquet	Periodista
86	Magui Prado	maguipradop	Periodista
87	Pilar Teijeiro	pilar_teijeiro	Periodista
88	Silvana Goicoechea	silgoicoechea	Periodista
89	Paota Botti	bottipaola	Periodista
90	Rosina Mallarini	Rmallarini	Periodista
91	Lorena Bomio	LorenaBomio	Periodista
92	Maria Eugenia Garcia	tuquegarcia	Periodista
93	Claudia García	GarciaMaClaudia	Periodista

Nro.	Nombre	Twitter	Tipo
94	Verónica Chevalier	VeroChevalier	Periodista
95	Ana Laura Gonzales	analaurafaral	Periodista
96	Patricia Vicente	patriciavicente	Periodista
97	Valentina Giménez	vale_gimenez	Periodista
98	Sofía Berardi	sofiberardi	Periodista
99	Camila Pirez	camilapirez22	Periodista
100	Stephanie Demirdjian	desde_abajo	Periodista
101	Belen Fourment	belenfourment	Periodista
102	Déborah Friedmann	debyfri	Periodista
103	Camila bello	CamilaBelloR	Periodista
104	Rosario Rodríguez	RosarioRodri22	Periodista
105	Soledad Gago	soledadgago	Periodista
106	Florencia Traibel	flotraibel	Periodista
107	Verónica Amorelli	veroamorelli	Periodista
108	Carolina Delisa	carodelisa	Periodista
109	Macarena Saavedra	macasaav_	Periodista
110	Mayte de León	maytedeleonfa	Periodista
111	Alejandra Casablanca	NegraCasablanca	Periodista
112	Lorena Nachajon	lorenanachajon	Periodista
113	Majo Borges	BorgesMajo	Periodista
114	Lucía Camila	Lucia4_Camila	Periodista
115	Vivana Aguerre	Vaguerre	Periodista
116	Alexandra Morgan	amorganvilaro	Periodista
117	Silvana Nicola	silvana_nicola	Periodista
118	Gabriela Casullo	GabrielaCasullo	Periodista
119	Rosana Tropiano	tropianorosana	Periodista
120	Adriana Laca	AdrianaLaca1	Periodista
121	Fabiana Goyeneche	Fabigoyen	Lideresa
122	Virginia Cardozo	Virginia_567	Lideresa
123	María Rosa Oña	MarosaOnia	Lideresa
124	DaisyTourné	MujerMira1	Lideresa
125	Mariana Percovich	PercoMariana	Lideresa
126	Laura Falero	MochilaAsesina	Lideresa
127	Manuela da Silveira	manudasil	Artista
128	Camila Rajchman	camilaraj	Artista
129	Patricia Wolf	PatriWolf	Artista
130	Eunice Castro	EuniceCastroC	Artista
131	Andy Vila	Andyyvila	Artista
132	Paola Bianco	paobianco	Comunicadora
133	Karina Vignola	karinavignola	Comunicadora
134	Denisse Legrand	ouicoucou	Comunicadora
135	Ximena Torres	XimeTorres8	Comunicadora
136	María Gomensoro	MariaGomensoro	Comunicadora
137	Sara Perrone	saritaperrone	Comunicadora
138	Yisela Moreira	yiselamoreira	Comunicadora
139	Natalie Yoffe	NatalieYoffe	Artista
140	Varina de Césare	VarinaDeCesare	Artista
141	Martina Graf	MartinaGrafG	Comunicadora
142	Annasofia Facello	annasofiaok	Artista

Nro.	Nombre	Twitter	Tipo
143	Victoria Zangaro	vzangaro	Artista
144	María Noel Riccetto	RiccettoMaria	Artista
145	Giannina Silva	Gianninasilvaok	Artista
146	Maria Ines Obaldía	minesobaldia	Comunicadora
147	Eleonora Navatta	EleonoraNavatta	Comunicadora
148	Verónica Piñeyrúa	veropinie	Comunicadora
149	Eliana Dide	ElianaDide	Comunicadora
150	Noelia Echeverry	noe_etcheverry	Comunicadora
151	Leticia Cicero	LeticiaCicero	Comunicadora
152	Lourdes Ferro	lourdes_ferro	Comunicadora
153	Adriana Da Silva	dasilvaactriz	Artista
154	Leonor Svarcas	Leonorsvarcas	Artista
155	Patricia Fierro	patricia_fierro	Artista
156	Valeria Alonso	valealonso09	Comunicadora
157	Andrea Menache	andymenache	Comunicadora
158	Natalia Roba	natiroba	Comunicadora
159	Sandra Rodríguez	SandraCelinaR	Comunicadora
160	Luciana Gonzalez	lucianitagonzal	Artista
161	Jimena Sabaris	jimesaba	Artista
162	Alejandra Rey	mariaalerey	Comunicadora
163	Catalina de Palleja	catadepalleja	Comunicadora
164	Virginia Ithurbide	Vithurbide	Comunicadora
165	Cecilia Olivera	CeciliaOlivera6	Comunicadora
166	Isabelle Chaquiriand	ichaquiriand	Comunicadora
167	Verónica Lavalle	asiestudia	Comunicadora
168	Mariana Garcia Seijas	MGARCIASEIJAS	Comunicadora
169	Natalia Gemelli Molfino	NatuGemelli	Comunicadora
170	Claudia Fernandez	cfernandezok	Artista
171	Florencia Infante	florenciainfant	Artista
172	Cata Ferrand	CataFerrand	Artista
173	Daiana Abracinskas	DaiAbracinskas	Artista
174	Sofía Romano	alcafecafe	Artista
175	Luciana Acuña	Lucianaacuna	Artista
176	Natalia Oreiro	VenenososdS	Artista
177	Mónica Navarro	navarromonica	Artista
178	Ceci Caputi	cecicaputi	Artista
179	Laura Martínez	Lauramartinezof	Artista
180	Carla Lorenzo Selave	carlalorenzok	Artista

Tab. 7.1: Listado de cuentas estudiadas.

BIBLIOGRAFÍA

[1] Josh Achiam et al. "Gpt-4 technical report". En: *arXiv preprint arXiv:2303.08774* (2023).

[2] Tom Brown et al. "Language models are few-shot learners". En: *Advances in neural information processing systems* 33 (2020), págs. 1877-1901.

[3] Pete Burnap et al. "Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack". En: *Social Network Analysis and Mining* 4 (2014), págs. 1-14.

[4] José Cañete et al. "Spanish pre-trained bert model and evaluation data". En: *arXiv preprint arXiv:2308.02976* (2023).

[5] Mohit Chandra et al. "AbuseAnalyzer: Abuse detection, severity and target prediction for gab posts". En: *arXiv preprint arXiv:2010.00038* (2020).

[6] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". En: *arXiv preprint arXiv:1406.1078* (2014).

[7] Javier De la Rosa et al. "Bertin: Efficient pre-training of a spanish language model using perplexity sampling". En: *arXiv preprint arXiv:2207.06814* (2022).

[8] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". En: *arXiv preprint arXiv:1810.04805* (2018).

[9] Jeffrey L Elman. "Finding structure in time". En: *Cognitive science* 14.2 (1990), págs. 179-211.

[10] Asier Gutiérrez-Fandiño et al. "Maria: Spanish language models". En: *arXiv preprint arXiv:2107.07253* (2021).

[11] Sepp Hochreiter y Jürgen Schmidhuber. "Long short-term memory". En: *Neural computation* 9.8 (1997), págs. 1735-1780.

[12] Muhammad Okky Ibrohim e Indra Budi. "Multi-label hate speech and abusive language detection in Indonesian Twitter". En: *Proceedings of the third workshop on abusive language online*. 2019, págs. 46-57.

[13] Daniel Jurafsky y James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

[14] Equipo Latinoamericano de Justicia y Género (ELA). "Violencia contra las mujeres y disidencias en política a través de redes sociales. Una aproximación a partir del análisis de la campaña electoral en Twitter, Facebook e Instagram durante 2019". En: (2020).

[15] Emma Kavanagh y Lorraine Brown. "Towards a research agenda for examining online gender-based violence against women academics". En: *Journal of Further and Higher Education* 44.10 (2020), págs. 1379-1387.

[16] Klaus Krippendorff. "Computing Krippendorff's alpha-reliability". En: (2011).

[17] Elissa Lee y Laura Leets. "Persuasive storytelling by hate groups online: Examining its effects on adolescents". En: *American behavioral scientist* 45.6 (2002), págs. 927-957.

[18] Younghun Lee, Seunghyun Yoon y Kyomin Jung. "Comparative studies of detecting abusive language on twitter". En: *arXiv preprint arXiv:1808.10245* (2018).

[19] Ruth Lewis, Michael Rowe y Clare Wiper. "Online abuse of feminists as an emerging form of violence against women and girls". En: *British journal of criminology* 57.6 (2017), págs. 1462-1481.

[20] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". En: *arXiv preprint arXiv:1907.11692* (2019).

[21] John McCarthy et al. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955". En: *AI Magazine* 27.4 (dic. de 2006), pág. 12. DOI: 10.1609/aimag.v27i4.1904. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904.

[22] Warren S McCulloch y Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". En: *The bulletin of mathematical biophysics* 5 (1943), págs. 115-133.

[23] Reid McIlroy-Young y Ashton Anderson. "From "Welcome New Gabbers" to the Pittsburgh Synagogue Shooting: The Evolution of Gab". En: *Proceedings of the International AAAI Conference on Web and Social Media* 13.01 (jul. de 2019), págs. 651-654. DOI: 10.1609/icwsm.v13i01.3264. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/3264.

[24] Katelyn YA McKenna y John A Bargh. "Coming out in the age of the Internet: Identity" demarginalization" through virtual group participation." En: *Journal of personality and social psychology* 75.3 (1998), pág. 681.

[25] Jessica Megarry. "Online incivility or sexual harassment? Conceptualising women's experiences in the digital age". En: *Women's Studies International Forum*. Vol. 47. Elsevier. 2014, págs. 46-55.

[26] Karsten Müller y Carlo Schwarz. "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment". En: *American Economic Journal: Applied Economics* 15.3 (jul. de 2023), págs. 270-312. DOI: 10.1257/app.20210211. URL: https://www.aeaweb.org/articles?id=10.1257/app.20210211.

[27] Ankur P Parikh et al. "A decomposable attention model for natural language inference". En: *arXiv preprint arXiv:1606.01933* (2016).

[28] John Pavlopoulos et al. "Toxicity detection: Does context really matter?" En: *arXiv preprint arXiv:2006.00998* (2020).

[29] Sida Peng et al. "The impact of ai on developer productivity: Evidence from github copilot". En: *arXiv preprint arXiv:2302.06590* (2023).

[30] Juan Manuel Pérez et al. "Assessing the impact of contextual information in hate speech detection". En: *IEEE Access* 11 (2023), págs. 30575-30590.

[31] Juan Manuel Pérez et al. "Robertuito: a pre-trained language model for social media text in spanish". En: *arXiv preprint arXiv:2111.09453* (2021).

[32] Fabio Poletto et al. "Resources and benchmark corpora for hate speech detection: a systematic review". En: *Language Resources and Evaluation* 55 (2021), págs. 477-523.

[33] Julie Posetti et al. *Online violence against women journalists*. 2020.

[34] James Pustejovsky y Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* " O'Reilly Media, Inc.", 2012.

[35] Alec Radford et al. “Improving language understanding by generative pre-training”. En: (2018).

[36] Burr Settles. “Active learning literature survey”. En: (2009).

[37] Wilson L Taylor. ““Cloze procedure”: A new tool for measuring readability”. En: *Journalism quarterly* 30.4 (1953), págs. 415-433.

[38] Ashish Vaswani et al. “Attention is all you need”. En: *Advances in neural information processing systems* 30 (2017).

[39] Bertie Vidgen, Emily Burden y Helen Margetts. “Understanding online hate, VSP Regulation and the broader context”. En: *Alan Turing Institute* 11 (2021).

[40] Bertie Vidgen, Helen Margetts y Alex Harris. “How much online abuse is there”. En: *Alan Turing Institute* 11 (2019).

[41] Bertie Vidgen et al. “Detecting East Asian prejudice on social media”. En: *arXiv preprint arXiv:2005.03909* (2020).

[42] Emily A Vogels. “The state of online harassment”. En: *Pew Research Center* 13 (2021), pág. 625.

[43] Matthew L Williams y Jasmin Tregidga. “Hate crime victimization in Wales: Psychological and physical impacts across seven hate crime victim types”. En: *British Journal of Criminology* 54.5 (2014), págs. 946-967.

[44] End Violence Against Women. “New Technology: Same Old Problems”. En: *Report of a Roundtable on Social Media and Violence Against Women and Girls, available online at http://www. endviolenceagainstwomen. org. uk/resources/61/new-technology-same-old-problems-dec-2013* (2013).