Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 187-198 recibido 16-12-2022 revisado 01-02-2023 aceptado 12-02-2023

Violencia Identificada en el Lenguaje (VIL). Creación de recurso para mensajes violentos

Violence Identified in Language (VIL). Creation of a resource for the detection of violent messages

Beatriz Botella, Robiert Sepúlveda-Torres, Patricio Martínez Barco, Estela Saquete
Department of Software and Computing Systems, University of Alicante, Spain
{beatriz.botella, rsepulveda, patricio, stela}@dlsi.ua.es

Resumen: La sociedad avanza cargada de conocimientos nuevos y muy accesibles, que se publican en el mundo virtual. Es una realidad que las Tecnologías de la Información y la Comunicación (TIC) han traído muchos beneficios a nuestras vidas pero también vemos como año tras año aumenta el uso de violencia en plataformas digitales. Nuestro trabajo se enfoca en la creación de recursos que permitan la detección de mensajes violentos en la red social Twitter. Se parte de la creación de una guía de anotación de grano fino para anotar un corpus de mensajes violentos (VIL) con el fin de utilizar herramientas de aprendizaje automático que nos ayuden a detectar automáticamente el problema. Con este corpus se entrenan dos modelos de lenguaje (BETO y RoBERTa_base) con los que se alcanza un valor en la métrica F_1m de 97.03 % y 96.51 % clasificando si un tuit es o no violento.
Palabras clave: Procesamiento Lenguaje Natural, Guía Anotación, Anotación Corpus, Detección Mensajes Violentos.

Abstract: Society is moving forward full of new and very accessible knowledge, which is published in the virtual world. It is a reality that ICTs have brought many benefits to our lives but we also see how year after year the use of violence on digital platforms increases. Our work focuses on the detection of violent messages in the social network Twitter. Starting from the creation of a fine-grained annotation guide to obtain a corpus of violent messages (VIL) in order to use Machine Learning tools that help us to automatically detect the problem Two language models are trained with this corpus (BETO and RoBERTa_base) with which a value of 97.03 % and 96.51 % is reached in the F_1m metric, classifying whether or not a tweet is violent.
Keywords: Natural Language Processing, Annotation Guideline, Dataset Annotation, Detection of Violent Messages.

1 Introducción

Internet se ha convertido en parte imprescindible de nuestras vidas, siendo utilizado prácticamente en todas las actividades cotidianas de la sociedad. Actualmente es posible tener contacto con cualquier persona del mundo a través de un dispositivo electrónico de manera inmediata. La sociedad avanza cargada de conocimientos nuevos y muy accesibles que se publican en el mundo virtual. Las relaciones personales también se han visto afectadas, no solo en el ámbito privado, sino también en el laboral.

Según WeAreSocial y Hootsuite (2022), casi 44 millones de personas en España son usuarias de Internet pasando más de 6 horas al día en la Red y alrededor de 41 millones de españoles son usuarios de redes sociales. Es una realidad que las TIC han traído muchos beneficios a nuestras vidas, pero también, gracias a la posibilidad de ser un usuario anónimo y la ausencia de observar cara a cara el daño que pueden generar nuestras palabras, se crean problemas aún por solucionar (Flores y Casal, 2008). En especial, muchos los investigadores denominan a este tipo de acción violenta como discurso del odio, una conducta ofensiva a través del lenguaje hacia personas o colectivos y cuya detección está siendo un problema para los investigadores, ya que, cabe la posibilidad de que la violencia no esté empleada de una forma explicita en un discurso, si no, ser una única pala-

ISSN 1135-5948 DOI 10.26342/2023-70-15
©2023 Sociedad Española para el Procesamiento del Lenguaje Natural

bra o incluso mediante una forma implícita con el uso de emoticonos (Alonso y Vázquez, 2017), o usando el humor, la ironía, el sarcasmo (Frenda, Patti, y Rosso, 2022; Frenda et al., 2022) o esteriotipos (Sánchez-Junquera et al., 2021).

Dada la cantidad de usuarios presentes en las redes sociales se hace imposible un control manual de los comentarios que se registran y su intención, creando una impunidad a las personas que utilizan estas redes con el fin de hacer daño. La identificación de mensajes violentos y controlar el discurso del odio en Internet se ha abordado desde diferentes puntos de vista, siendo imprescindible la utilización de Procesamiento del Lenguaje Natural (PLN) para desarrollar sistemas computacionales que ayuden a interpretar y procesar el lenguaje humano de forma rápida y efectiva.

Una barrera que encontramos nada más empezar el estudio es la recopilación de mensajes en las redes sociales, ya que como apunta Bruns (2019), la restricción al acceso de datos de las redes sociales dificulta el análisis de cuestiones de gran importancia como el lenguaje abusivo, el acoso, el discurso de odio o las campañas de desinformación.

Es por ello por lo que en la presente investigación se usará la red social Twitter donde cómo define Ott (2017): "El discurso de Twitter es irrespetuoso porque su registro es informal, y porque despersonaliza las interacciones sociales". Esta investigación persigue el objetivo de aportar soluciones a los problemas existentes en la detección de mensajes violentos en redes sociales de una forma rápida, automática y eficaz. La principal contribución del trabajo es un esquema de anotación de grano fino que vaya más allá de marcar un mensaje como violento o no, sino que permite una anotación semántica mucho más compleja del mismo, permitiendo un nivel de detalle mucho más exhaustivo que la simple detección binaria.

El artículo está estructurado de la siguiente manera: Sección 2, se muestran los principales trabajos realizados en la materia y las formas de detección; en la sección 3 describimos cual han sido los pasos de anotación para etiquetar mensajes violentos. La sección 4 explica el proceso de compilación, anotación, así como una prueba piloto para verificar la anotación de nuestra guía. La validación de nuestro corpus y experimentos se encuentran en la sección 5; la sección 6, muestra los resultados de la experimentación realizada y por último en la sección 7, conclusiones y trabajo futuro.

2 Estado de la cuestión

Son muchos los estudios que se han llevado a cabo sobre el análisis de mensajes violentos en redes sociales y medios de comunicación. En concreto, se puede encontrar mucha investigación centrada en descubrir las características del comportamiento humano que promueven la emisión de dichos mensajes, así como los que se centran en descubrir las características de los propios mensajes a través de técnicas de PLN. Si bien nuestro estudio está enfocado a este último grupo, revisaremos algunos de los trabajos más importantes para ambos casos.

2.1 Estudio del lenguaje y comportamiento

Hay una gran cantidad de estudios acerca del comportamiento humano ante los mensajes violentos y el lenguaje empleado. Como dijo McMenamin (2017), "el discurso del odio se estudia según cómo se define, cómo se interpreta, y cuáles son las mejores prácticas para enfrentarlo". Es por ello que encontramos trabajos como Salado (2022), que basaron su investigación en un análisis sintáctico del lenguaje, y descubrieron que hay distintos elementos lingüísticos a tener en cuenta que están presentes en las formas del habla violentas como, la categoría lingüística, el léxico empleado o cómo están colocadas las palabras. Del Arco et al. (2022) realiza un estudio de los fenómenos lingüísticos implícitos y explícitos del lenguaje ofensivo. Otros como Gitari (2015) se centraron en algo tan específico como la creación de un listado de verbos que pueden ser indicadores de mensajes violentos. Por otra parte existen trabajos que se centran en los roles presentes en estos actos como por ejemplo, Nielsen (2002) que a través de unas entrevistas y estudio de los participantes, observó las consecuencias para la víctima, su daño y la posibilidad de delito en los mensajes.

2.2 PLN aplicado a la detección de mensajes violentos

La aplicación del PLN es fundamental en este tipo de investigaciones dado el gran volumen de datos existentes, lo que facilita un gran avance en la investigación de la detección de

este tipo de mensajes, gracias a las siguientes técnicas:

- **Clasificadores basados en palabras claves**

  Una parte de las investigaciones en este campo se han centrado en la elaboración de lista de insultos que ayuden a una detección automática. En este sentido, se han desarrollado lexicones y diccionarios con el fin de observar si la presencia de estos términos determina la violencia en el mensaje (Sood, Churchill, y Antin, 2012).

  Aunque este tipo de listas han ayudado a la detección, se ha quedado escaso a la hora de ser la única herramienta para determinar la violencia. El lenguaje violento evoluciona constantemente,varía según el lugar donde ocurra y es posible que existan términos que en algunas zonas geográficas sean insultos y en otras no (Nobata et al., 2016).

- **Aprendizaje automático**

  La mayoría de los trabajos relacionados con la detección de mensajes violentos abordan esta problemática con la utilización de algoritmos clásicos de aprendizaje automático (ML). Trabajos como Xu et al. (2012) y Dadvar et al. (2013) han utilizado máquinas de soporte vectorial (SVM) en sus investigaciones obteniendo resultados satisfactorios, demostrando ser muy eficaz con muestras de entrenamiento de grandes dimensiones. SVM no es el único algoritmo clásico utilizado en las investigaciones de este campo, trabajos como Arcila-Calderon et al. (2021), utilizaron otros algoritmos, mostrando en sus resultados que el que ofrecía mejor rendimiento es la regresión logística, seguida de Naive Bayes y las SVM.

  La mayoría de los clasificadores basados en ML utilizan representaciones de textos tradicionales como bolsa de palabras (BOW), n-grams, frecuencia de términos (TF), entre otras. En Burnap y Williams (2014) se utilizan todas las técnicas citadas anteriormente. Esta investigación compara los resultados obtenidos de forma individual por los clasificadores con la utilización de un conjunto de clasificadores (ensemble) que los integra a todos, demostrando mayor precisión en este el último. El análisis de sentimientos es otra de las herramientas más utilizadas en este campo. Con ella podemos extraer la polaridad del mensaje y utilizar este indicador junto a otras tareas para determinar con mayor exactitud si estamos ante un mensaje violento o no (Martins et al., 2018).

  El desarrollo de corpus, tienen un papel importante en las investigaciones del lenguaje ofensivo cuando se aplican técnicas de ML. En los últimos años hemos observado un gran volumen de trabajo por parte de investigadores en PLN para generar estos recursos (Wiegand et al., 2018; Qian et al., 2019; Olteanu et al., 2018; Fortuna y Nunes, 2018; Poletto et al., 2021; Rosenthal et al., 2020). Estos autores crearon recursos en inglés, siendo SOLID (Rosenthal et al., 2020) el recurso que contiene más de nueve millones de tuits en inglés etiquetados de forma semisupervisada.

  Por otra parte, HurtLex (Bassignana, Basile, y Patti, 2018) es un léxico multilingüe de palabras de odio que abarcan varios idiomas y Hatebase3[1] es un repositorio colaborativo de discurso de odio también multilingüe. El principal inconveniente de estos recursos es su escasez de términos en español, y los que están presentes se han recopilado utilizando una traducción semiautomática de otro idioma, dejando de lado la importancia de los factores culturales y lingüísticos de cada país. Sin embargo, a pesar de que el español es una de las lenguas más habladas del mundo, encontramos escasez de recursos en este idioma para llevar a cabo la tarea de detección.

  Existen recursos en español de palabras ofensivas como Plaza-Del-Arco et al. (2020) para términos misóginos y xenófobo; y Share (Plaza-del Arco et al., 2022) que los etiquetan como ofensivo y no. Tras el estudio realizado sobre la literatura se considera necesaria la elaboración de otro corpus donde recoger más características presentes en los mensajes violentos, que puedan ayudar en la explicabilidad y el detalle de la detección.

- **Aprendizaje profundo**

[1] `https://hatebase.org/`

Dentro de la Inteligencia Artificial existen otras técnicas más complejas que también se han utilizado en esta tarea. Nos referimos al aprendizaje profundo (DL), como es el caso de la investigación de Arcila-Calderón et al. (2021) que tras utilizar las herramientas de aprendizaje automático y redes neuronales, estas últimas mejoraron las métricas de evaluación frente a los modelos generados con algoritmos de ML tradicionales. Con el mismo fin Badjatiya et al. (2017) usa modelos de DL para entrenar diferentes incrustaciones de palabras validando que, utilizar estas representaciones, obtiene mejor resultados que representaciones tradicionales como frecuencia de término – frecuencia inversa de documento (TF-IDF) o BoW.

Los modelos basados en arquitectura *transformer* como es el caso de BERT, RoBERTa y ALBERT, ostentan los mejores resultados del estado del arte en la detección de mensajes violentos en tareas reconocidas como OffensEval o HatEval (Sarkar et al., 2021). En Sarkar et al. (2021) se realiza un ajuste fino (*fine-tuning*) a BERT utilizando SOLID, el mayor corpus de identificación de lenguaje ofensivo en inglés, mejorando los resultados obtenidos con BERT en las tareas mencionadas anteriormente.

En Song et al. (2021) se utiliza un conjunto de clasificadores (*ensemble*) basados en RoBERTa y BERT que obtiene los mejores resultados en la tarea compartida "SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense"[2] que incluye una subtarea de detección de mensajes ofensivos. Este trabajo consiste en hacer un ajuste fino de estos modelos para crear un clasificador y agruparlos en un conjunto de clasificadores basados en *stacking*.

Una vez estudiada la literatura al respecto de la tarea y la importancia de la aplicación de PLN y técnicas de ML y DL, y debido a que estas técnicas se nutren de datos de entrenamiento, se concluye la necesidad de crear un recurso en español que pueda ser utilizado en la detección efectiva de mensajes violentos, con un nivel de detalle que va más allá de la simple detección binaria, marcando características, que detallamos en nuestra guía de anotación, como el grado de violencia, el rol o el tipo de violencia, puesto que consideramos que si detección en los mensajes mensajes que puedan ayudar a la futura explicabilidad en las decisiones tomadas.

3 Guía de anotación

Persiguiendo el objetivo de crear un recurso que ayude a la detección de los mensajes violentos, decidimos generar una guía de anotación de grano fino para los mensajes, con un cierto grado de complejidad semántica, donde no solamente se marca si un mensaje es violento o no, si no también determinados elementos importantes al respecto del contenido del mensaje.

3.1 Violento vs NoViolento

Para empezar nuestra anotación elegimos si el mensaje es violento o no, definiendo que entendemos por mensaje violento:

- **Violento**: Cuando el contenido del mensaje contiene la acción de hacer daño o emplea violencia en sus palabras, ejemplos: "Deberían estar en la cárcel", "eres idiota","no soporto la mierda de este puto país".
- **NoViolento**: El mensaje no contiene acción de hacer daño, aunque puede existir "palabras violentas", como por ejemplo: "Que idiota soy, por casi me lo creo", "Hoy es un gran día", "me he levantado insoportable hoy".

  Si el contenido del mensaje es *NoViolento*, solo anotaremos si contiene insultos o palabras negativas. Las demás categorías no se anotarán.

3.2 Contenido del mensaje

En este apartado nos centraremos en analizar el contenido del mensaje más en detalle para determinar tres elementos fundamentales en el mismo: i) el grado de violencia, ii) el rol del mensaje y iii) el tipo de violencia. Dentro del contenido del mensaje observamos por ejemplo si los mensajes contienen insultos o expresiones negativas por si puede ser un indicador de violencia, ejemplo: "matón de patio", "me cago en tus muertos", así como la identificación de la estructura de un

[2] http://bit.ly/3J8uHOX

mensaje violento, que no contenga un insulto o palabra ofensiva directa, ejemplo: "Te vas a enterar", "Espero encontrarte a solas en la calle". Para todo ello se seleccionan las palabras que contienen violencia.

3.2.1 Grado de Violencia

Según el contenido del mensaje podemos catalogarlo en 2 niveles de violencia:

- **Moderado**: Se recogen aquí todos los mensajes que lleven contenido violento pero no atenten contra la vida o integridad física de las personas. Groserías, desaprobación con personas o cosas, ridiculizar, insultos por ideología o política. Ejemplos: "Ese es Gilipollas", "Que asco de hotel", "Maria la tetona", "me cago en tu puta madre".
- **Grave**: En este nivel los mensajes atentan contra la vida o integridad de las personas. Amenazas, agresión física, desear el mal. Encontramos verbos como ojalá o deseo con acciones negativas y acusar a personas de delitos graves, como asesino, violador, proxeneta. Ejemplos: "Ojalá te murieses", "Ten cuidado a ver si te pasa algo", "te voy a dar una ostia cuando te vea", "eres una asesina".

3.2.2 Rol del mensaje

Definir de que forma actúan los usuarios en mensajes violentos.

- **Rol 1 - Incitador**: Su mensaje incita a los demás lectores a que escriban mensajes violentos o propicia el odio en la red. Ejemplos: "Deberíamos decirle los españoles lo idiota que es esta tía","Tendríamos que ir todos a tu casa a darte tu merecido".
- **Rol 2 - Ejecutor**: Mensaje de un usuario individual con acción directa de violencia. Ejemplos: "Deberías morirte","Eres retrasado".
- **Rol 3 - Pasivo**: Emplea violencia sin estar dirigida a nadie en concreto. Ejemplos: "La política de este país es una mierda", "Siempre nos cuentan lo mismo, se creen que somos idiotas".
- **Rol 4 - Informativo**: Es mero transmisor de la violencia pero no participa en ella. Ejemplo: "No soporto la violencia".

3.2.3 Tipo de violencia

Es importante definir que tipo de violencia se está empleando en el mensaje. Según el Ministerio del Interior del Gobierno de España, en su informe emitido en 2021 [3] los delitos de odio cometido en internet y redes sociales son: racismo/xenofobia en primer lugar, seguido de orientación sexual e identificación de género, ideología, discriminación por razón de sexo, creencias o prácticas religiosas, discriminación generacional, delitos de odio contra personas con discapacidad, descriminación por razón de enfermedad, antigitanismo, antisemitismo y aporofobia. Dado el gran volumen existente de datos, se decidió etiquetar los mensajes en 5 tipos de violencias. Añadiendo la violencia machista, no presente en el informe anterior, por su alerta en la sociedad y con el fin de estudiar este problema en futuras investigaciones.

- **Machista**: el contenido del mensaje conlleva una actitud despectiva contra las mujeres. Ejemplos: "Lo has conseguido por ser una guarra y ponerte de rodillas", "Las mujeres no sabéis hacer otra cosa", "Estás con él por tu dinero".
- **Homófoba**: mensajes violentos hacia la homosexualidad o las personas homosexuales. Ejemplos: "Los gays están enfermos y tienen que curarse", "Bolleras de mierda".
- **Ideología_religiosa**: ataques contra las ideologías religiosas. Ejemplos: "Me río de tu dios Alá", "Los católicos son asquerosos".
- **Política**: violencia hacia cualquier ideología política o persona/s política que los representa. Ridiculizar nombres políticos. Ejemplos: "Peperos de mierda", "Los de podemos son asquerosos".
- **Xenofobia/Racismo**: rechazo a cualquier persona por el mero hecho de no compartir la misma nacionalidad o actitud o ideología donde una raza o grupo étnico se considera superior a otra. Ejemplos: "Moro de mierda", "Panchitos", "No podemos ser iguales que los negros, ellos son una escala inferior".
- **Otro**: otro tipo de violencia que no corresponda a los anteriores. Ejemplos:

[3] https://bit.ly/3FXZyxt

"Madrilistas de mierda"(Violencia Deportiva), "Ojala maten al perro"(Violencia animal), "Asco de los toreros"(Profesión). Este ultimo apartado se les pedía a los que en el apartado de Observaciones escribieran qué tipo de violencia era la que habían catalogado como Otro.

4 Corpus VIL

Una vez definida la guía de anotación, se procede a la construcción de un corpus basándonos en esa guía. Las fases seguidas en la construcción del recursos serán presentados en la siguientes subsecciones.

4.1 Proceso de compilación

Para poder tener un corpus de mensajes violentos etiquetados, primero se pensó en que red social era la más apropiada para descargar mensajes de los usuarios. Basado en la justificación mostrada en la sección 1, se escogió Twitter debido a la manera informal de expresarse que tienen los usuarios en esta red social. Además, Twitter permite descargar tuits con gran facilidad. A continuación, se pensaron 3 escenarios en los que se presencian opiniones que afectan a la sociedad actual, con lo que obtendríamos tuits reales con alta probabilidad de violencia. Los tuits seleccionados están relacionados con 3 acontecimientos ocurridos en España:

- Entrevista de La Sexta realizada a la política Cayetana Álvarez de Toledo.
- La campaña de Irene Montero, ministra de igualdad sobre la campaña "Sola y borracha quiero llegar a casa".
- Isabel Díaz Ayuso, en la manifestación que ocurrió el 13 de noviembre 2022 por la Sanidad en Madrid.

Esta elección se hizo en base a la actualidad en la sociedad española y el odio existente a los políticos de nuestro país. Los tuits se descargaron mediante la herramienta "Social Analytics" (Fernández et al., 2017), en total unos 12500 tuits. Con los tuits descargados se realizó un proceso de limpieza para eliminar tuits repetidos y retuits, generándose un total de 90 paquetes de tuits, donde cada paquete contiene 100 tuits.

4.2 Prueba piloto anotación

Para asegurarnos de que nuestra guía de anotación era correcta, contamos con la ayuda de 6 anotadores entrenados por un experto en la guía de anotación, cada uno de ellos anotó la misma cantidad de tuits (40 tuits). En el proceso de anotación de esta prueba piloto se observó que la guía era demasiado compleja, derivando en confusión para los anotadores, con anotaciones incorrectas. Con esta prueba se decidió hacer una serie de modificaciones a la guía de anotación, simplificando los pasos a seguir por los anotadores. En la guía inicial contábamos con 3 niveles de violencia, (leve, moderado y grave), pero la línea de decisión entre las opciones leve y moderado era muy difícil de definir por los anotadores, debido a la subjetividad de la violencia dependiendo de la persona que etiquetaba. Por ese motivo se modificó dejando solos dos niveles de violencia (presentes en la guía actual) y se añadieron más ejemplos que permitiesen reducir todas las dudas que suscitaban. También se añadieron ejemplos en el resto de opciones para asegurar un etiquetado correcto.

4.3 Anotación del corpus VIL

Como resultado del proceso de anotación después de las dos primeras fases explicadas anteriormente se procedió a la anotación masiva de los tuits recopilados. Con esta anotación se obtiene el corpus Violencia Identificada en el Lenguaje (VIL), el cual contiene un total de 2874 tuits anotados con 1491 *Violento* y 1383 *NoViolento*. La cantidad de tuits *Violento* y *NoViolento* que contiene el corpus es similar, evitando así que los modelos que lo utilicen se vean afectados por un posible desbalance entre las clases que contiene.

Con el fin de evaluar el rendimiento de futuros modelos entrenados utilizando este corpus, se realiza un particionamiento del mismo en (entrenamiento, validación y prueba). La partición de prueba fue extraída aleatoriamente utilizando el 20 % de los tuits anotados, de los tuits restantes el 20 % se reserva para evaluar experimentos (partición de validación) y el resto para la partición de entrenamiento. La tabla 1 muestra la distribución de etiquetas por cada partición. El conjunto de datos VIL esta disponible para su descarga y utilización en `http://bit.ly/3ZVwUnL`.

Más concretamente, dato este conjunto de datos seleccionados sobre los tres eventos mencionados actualmente, la distribución por tipo de violencia es la siguiente: 13 mensajes machistas, 5 mensajes homófobos, mensajes de 0 ideología religiosa, 174 mensajes políti-

	Violento	NoViolento	Total
Entrenamiento	957	882	1839
Validación	236	224	460
Prueba	298	277	575
Total	1491	1383	2874

Tabla 1: Distribución de etiquetas en las particiones de entrenamiento, validación y prueba.

cos, 2 de Xenofobia y racismo y 1381 mensajes que no corresponden a ninguno de los anteriores. Este grupo sería el más amplio dada la complejidad de la clasificación, siendo para estos eventos recopilados concretamente los mensajes racistas los más escasos y ninguno de ideología religiosa. El tipo de violencia va muy ligado al tipo de situación o evento que se esté analizando, y debido a este balanceo en trabajos futuros será necesario la ampliación de los tipos de violencia más escasos.

Para la anotación de este corpus se ha utilizado la herramienta de anotación Brat (Stenetorp et al., 2012). Esta permite la anotación de mensajes de una forma intuitiva mostrando una ventana para seleccionar la anotación deseada. Previamente se configura los campos específicos de la anotación así como la jerarquía en las anotaciones. Los insultos etiquetados mediante Brat se encuentran disponibles en el siguiente repositorio GitHub: `https://bit.ly/3ZVwUnL`. La figura 1 muestra algunos ejemplos de tuits anotados en el conjunto de datos VIL.

5 Validación del esquema de anotación y del corpus VIL

Esta sección presenta una validación realizada al esquema de anotación para corroborar que la guía de anotación es clara y precisa, derivando en una anotación del corpus acorde con la definición de la misma. Para cumplir este objetivo se realiza una validación de acuerdo entre anotadores. Por otra parte, se valida la pertinencia del corpus VIL para crear un sistema de detección de mensajes violentos en Twitter.

5.1 Validación entre anotadores

Con el objetivo de medir la calidad de la tarea de anotación se realizó un acuerdo entre dos anotadores. Los dos anotadores elegidos son criminólogos y está selección se hizo por su conocimiento en el comportamiento violento entre humanos. Estos anotaron independientemente 200 tuits entre *Violento* y *NoViolento*, calculando un índice de acuerdo en la anotación. Se utilizó el índice *kappa* (Cohen, 1960) para calcular el acuerdo en las anotaciones (índice común en procesos de validación de anotaciones entre dos anotadores). Se obtuvo un *kappa* de 0,868, lo que representa un valor alto de acuerdo entre dos anotadores, validando así el proceso de anotación.

Adicionalmente se calculó el acuerdo teniendo en cuenta el contenido del mensaje marcado, primeramente evaluando si el tuit contiene insultos, alcanzándose un *kappa* de 0.896. Por último, teniendo en cuenta la anotación del grado de violencia, el índice de acuerdo es de 0.753, sustancialmente menor que el resto de anotaciones, lo que evidencia que esta etiqueta es la más compleja de anotar.

En cualquier caso, se considera que los valores *kappa* obtenidos son suficientes para garantizar la calidad del corpus.

5.2 Experimentos

Un sistema capaz de detectar tuits violentos es de gran relevancia en el contexto actual de ataques constantes a través de redes sociales. Para probar la validez del conjunto de datos VIL, se realizaron dos experimentos que lo utilizan como base para entrenar modelos de lenguaje y evaluar el rendimiento de los mismos para predecir si un tuit es violento o no.

Para llevar a cabo estos experimentos se utilizan dos modelos de lenguaje en español (BETO y RoBERTa_base), basados en arquitectura *transformers* descritos en Canete et al. (2020) y Gutiérrez-Fandiño et al. (2021) respectivamente.

BETO está basado en el modelo de lenguaje BERT, diseñado para representar relaciones bidireccionales profundas a partir de texto sin etiquetar, utilizando mecanismos de atención (Devlin et al., 2018). Para la creación de BETO se realizaron una serie de optimizaciones similares a las llevadas a cabo para obtener el modelo RoBERTa (Liu et al., 2019). En este caso se entrena utilizando textos en español de la enciclopedia libre (Wiki-

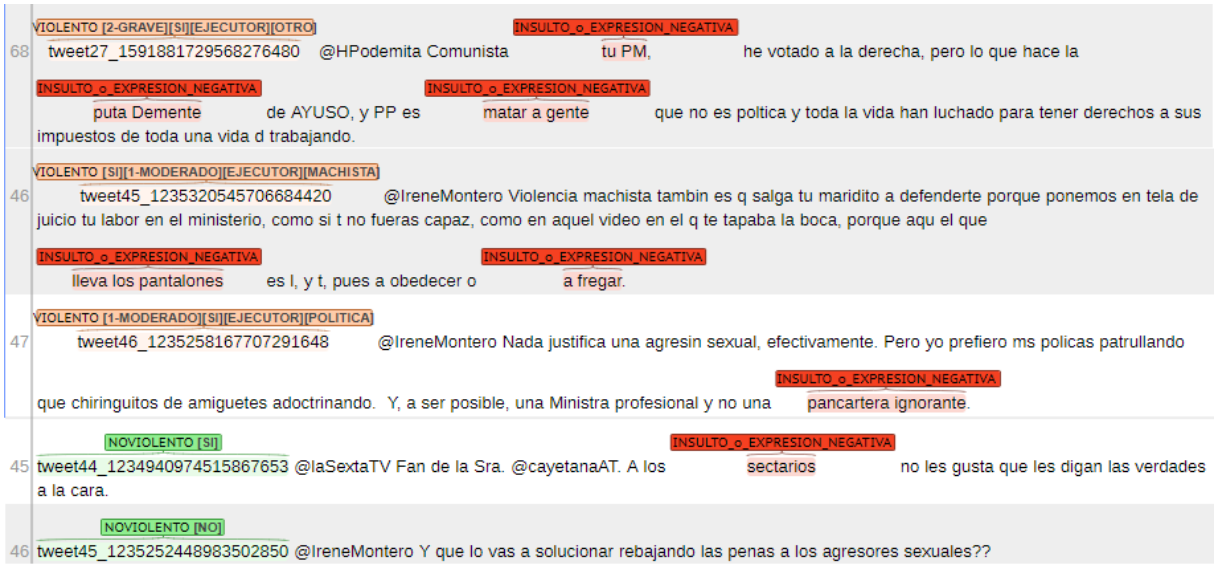

Figura 1: Ejemplos de tuits anotados en el corpus VIL.

pedia) y todas las fuentes del Proyecto OPUS (Tiedemann, 2012).

En el caso de RoBERTa_base en español se basa en el modelo de lenguaje RoBERTa. Para la obtención de este modelo se utiliza un total de 570 GB de texto limpio y recopilado por la Biblioteca Nacional de España de 2009 a 2019 (Gutiérrez-Fandiño et al., 2021).

En el contexto de estos experimentos los modelos BETO pre-entrenado[4] y RoBERTa_base[5] se utiliza en modo ajuste fino para ajustarlo a la tarea de detección de tuits violentos. En el primer experimento se entrenan los modelos de lenguajes utilizando como secuencia de entrada únicamente el texto del tuit. Por su parte, el segundo experimento concatena el texto del tuit con las frases anotadas como insultos para entrenar el modelo. Finalmente para llevar a cabo los experimentos se utilizó la biblioteca Simple Transformers[6] con la siguiente configuración de hiperparámetros en todos los experimentos: tasa de aprendizaje de 2e-5, tamaño de lote de 2 y número de iteraciones para entrenar 3.

6 *Resultados*

Los experimentos llevados a cabo en este artículo se pueden replicar descargando el código del siguiente repositorio GitHub: `http://bit.ly/3YGfVVs`.

[4] `http://bit.ly/3Feu5q6`
[5] `http://bit.ly/3FcBvu6`
[6] `https://simpletransformers.ai/`

Con la configuración inicial de hiperparámetros se realizó un ajuste fino preliminar para evaluar el comportamiento del modelo para predecir la partición de validación. La figura 2 muestra el comportamiento de las curvas de perdida, así como la métrica F_1m en cada iteración de entrenamiento. Esta figura corresponde con el ajuste fino del modelo BETO, sin embargo con el modelo RoBERTa_base el comportamiento es similar.

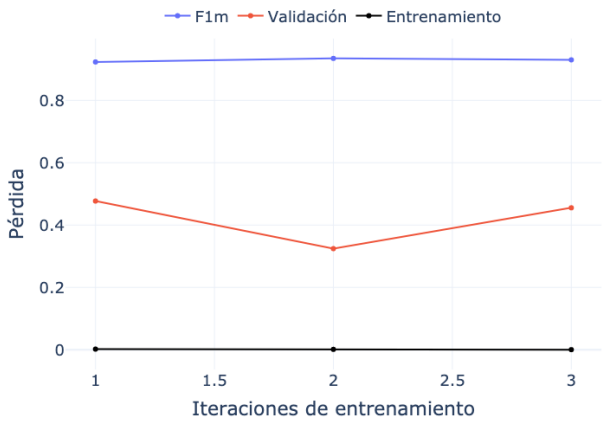

Figura 2: Curva de pérdida utilizando el conjunto de entrenamiento y validación durante el entrenamiento.

Como se puede apreciar en la figura 2, la pérdida en la curva de validación (línea roja) desciende de 0.47 a 0.32 de la primera iteración a la segunda, por el contrario en la tercera iteración esta aumenta hasta 0.45. Por su

parte la curva de la pérdida de entrenamiento (línea negra) disminuye en cada iteración. La disminución en la pérdida en el entrenamiento en todas las iteración así como el aumento para pérdida en validación (de la iteración 2 a la 3) es una evidencia contundente de que el modelo se empezó a sobreajustar. Por último la curva que representa la evolución de la métrica macro-promedio F_1 (F_1m) alcanza su máximo valor (93.47 %) en la segunda iteración, lo que se corresponde con el comportamiento de la curva de validación. Después de este análisis se decidió utilizar el modelo entrenado durante dos iteraciones para predecir la partición de prueba.

La tabla 2 muestra los resultados de las métricas puntuación F_1 por clase y el F_1m prediciendo la partición de prueba. Los valores se expresan en modo de porcentaje.

Los experimentos BETO (texto del tuit) y RoBERTa-base model (texto del tuit) obtienen un buen rendimiento en la predicción de si un tuit es o no violento. Estos solo utilizan como entrada al modelo de entrenamiento y predicción el texto del tuit. Los resultados alcanzados son similares a los obtenidos por Mathew et al. (2021) sobre el corpus HateXplain en idioma inglés para la detección de mensajes violentos, en este trabajo también se entrenan clasificadores basados en BERT. Para el caso del castellano, lo más cercano al trabajo aquí presentado sería la evaluación de SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (Basile et al., 2019), en la subtarea B donde se clasifica la agresividad de los mensajes, obteniéndose valores para el castellano de alrededor del 70.5 %.

Por su parte, los experimentos que concatenan el texto del tuit con los insultos anotados en el tuit —BETO (texto del tuit y los insultos) y RoBERTa_base model (texto del tuit y los insultos)— mejoran en todas las métricas a los sistemas bases que solo utilizan los textos del tuit. Este hallazgo indica la pertinencia de preanotar insultos en los textos y luego clasificarlos.

Teniendo en cuenta los modelos de lenguajes utilizados, con el modelo BETO se obtienen mejores resultados en el experimento que utiliza los insultos anotados como entrada al modelo. Sin embargo, para el experimento que solo utiliza el texto como entrada al modelo RoBERTa_base, se mejora la métrica F_1m en un punto porcentual con respecto al experimento utilizando el otro modelo. En los escenarios mencionados anteriormente las diferencias son bajas, sin embargo se considera que los resultados difieren de los esperados debido a que se pensaba que el modelo RoBERTa_base obtendría los mejores resultados en ambos experimentos debido a que fue entrenado sobre un conjunto de datos mucho más extenso y utilizando optimizando el proceso de entrenamiento.

7 Conclusiones y trabajo futuro

Este trabajo se ha realizado con el fin de encontrar mejoras en la detección de los mensajes violentos en redes sociales, utilizando un esquema de anotación de grano fino para obtener un corpus de mensajes violentos con indicadores de nivel de violencia, rol, presencia de insultos y tipo de violencia. El corpus VIL ha sido utilizado para entrenar clasificadores basados en modelos de lenguaje *transformers*. Estos clasificadores obtienen resultados significativos cuando se utiliza el texto del tuit concatenado con las frases con insultos anotadas. En próximos trabajos esperamos utilizar mecanismo de Human in the Loop y Active Learning para obtener un dataset a gran escala con mayor cantidad de mensajes violentos de tipo *Grave*, debido que en esta primera versión solo se cuenta con 60 mensajes de este tipo, además, vamos a trabajar con un mayor numero de dominios para aumentar los distintos tipos de violencia.

Agradecimientos

Esta investigación ha sido financiada por MCIN/AEI/ 10.13039/501100011033 y la Unión Europea NextGenerationEU/PRTR a través de los proyectos “TRIVIAL” (PID2021-122263OB-C22) and “SocialTrust” (PDC2022-133146-C22). También cuenta con el apoyo de la Generalitat Valenciana a través del proyecto “NL4DISMIS” (CIPROM/2021/21).

Bibliografía

Alonso, L. y V. J. Vázquez. 2017. *Sobre la libertad de expresión y el discurso del odio: Textos críticos*. Athenaica ediciones universitarias.

Arcila-Calderón, C., J. J. Amores, P. Sánchez-Holgado, y D. Blanco-Herrero. 2021. Using shallow and deep learning to automatically detect hate motivated by

	F_1		
	Violento	**NoViolento**	F_1m
BETO (texto del tuit)	87.32	86.92	87.12
BETO (texto del tuit y los insultos)	**97.18**	**96.89**	**97.03**
RoBERTa-base model (texto del tuit)	88.96	87.26	88.11
RoBERTa-base model (texto del tuit y los insultos)	**96.39**	**96.64**	**96.51**

Tabla 2: Resultados de los experimentos utilizando el corpus VIL.

gender and sexual orientation on twitter in spanish. *Multimodal technologies and interaction*, 5(10):63.

Badjatiya, P., S. Gupta, M. Gupta, y V. Varma. 2017. Deep learning for hate speech detection in tweets. En *Proceedings of the 26th international conference on World Wide Web companion*, páginas 759–760.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, y M. Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. En *Proceedings of the 13th international workshop on semantic evaluation*, páginas 54–63.

Bassignana, E., V. Basile, y V. Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. En *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volumen 2253, páginas 1–6. CEUR-WS.

Bruns, A. 2019. After the 'apicalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11):1544–1566.

Burnap, P. y M. L. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.

Canete, J., G. Chaperon, R. Fuentes, y J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Dadvar, M., D. Trieschnigg, R. Ordelman, y F. d. Jong. 2013. Improving cyberbullying detection with user context. En *European Conference on Information Retrieval*, páginas 693–696. Springer.

del Arco, F. M. P., M. D. Molina-González, L. A. Ureña-López, y M.-T. Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fernández, J., F. Llopis, P. Martínez-Barco, Y. Gutiérrez, y Á. Díez. 2017. Analizando opiniones en las redes sociales. *Procesamiento del Lenguaje Natural*, 58:141–148.

Flores, J. y M. Casal. 2008. *Ciberbullying. Guía rápida para la prevención del acoso por medio de las nuevas tecnologías*.

Fortuna, P. y S. Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Frenda, S., A. T. Cignarella, V. Basile, C. Bosco, V. Patti, y P. Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Frenda, S., V. Patti, y P. Rosso. 2022. Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering*, páginas 1–22.

Gitari, N. D., Z. Zuping, H. Damien, y J. Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, y M. Villegas.

2021. Spanish language models. *arXiv preprint arXiv:2107.07253*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, y V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martins, R., M. Gomes, J. J. Almeida, P. Novais, y P. Henriques. 2018. Hate speech classification in social media using emotional analysis. *Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018*, páginas 61–66, 12.

Mathew, B., P. Saha, S. M. Yimam, C. Biemann, P. Goyal, y A. Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. En *Proceedings of the AAAI Conference on Artificial Intelligence*, volumen 35, páginas 14867–14875.

McMenamin, G. R. 2017. *Introducción a la lingüística forense: un libro de curso*. Press at California State University, Fresno.

Nielsen, L. B. 2002. Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social Issues*, 58:265–280, 1.

Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad, y Y. Chang. 2016. Abusive language detection in online user content. En *Proceedings of the 25th international conference on world wide web*, páginas 145–153.

Olteanu, A., C. Castillo, J. Boy, y K. Varshney. 2018. The effect of extremist violence on hateful speech online. En *Proceedings of the international AAAI conference on web and social media*, volumen 12.

Ott, B. L. 2017. The age of twitter: Donald j. trump and the politics of debasement. *Critical studies in media communication*, 34(1):59–68.

Plaza-Del-Arco, F.-M., M. D. Molina-González, L. A. Ureña-López, y M. T. Martín-Valdivia. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

Plaza-del Arco, F. M., A. B. P. Portillo, P. L. Úbeda, B. Gil, y M.-T. Martín-Valdivia. 2022. Share: A lexicon of harmful expressions by spanish speakers. En *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, páginas 1307–1316.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, y V. Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Qian, J., M. ElSherief, E. Belding, y W. Y. Wang. 2019. Learning to decipher hate symbols. *arXiv preprint arXiv:1904.02418*.

Rosenthal, S., P. Atanasova, G. Karadzhov, M. Zampieri, y P. Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Salado, M. R. 2022. Análisis lingüístico del discurso de odio en redes sociales. *VISUAL REVIEW. International Visual Culture Review/Revista Internacional de Cultura Visual*, 9(Monográfico):1–11.

Sánchez-Junquera, J., P. Rosso, M. Montes, B. Chulvi, y others. 2021. Masking and bert-based models for stereotype identication. *Procesamiento del Lenguaje Natural*, 67:83–94.

Sarkar, D., M. Zampieri, T. Ranasinghe, y A. Ororbia. 2021. Fbert: A neural transformer for identifying offensive content. *arXiv preprint arXiv:2109.05074*.

Song, B., C. Pan, S. Wang, y Z. Luo. 2021. Deepblueai at semeval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 1130–1134.

Sood, S. O., E. F. Churchill, y J. Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63:270–285, 2.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, y J. Tsujii. 2012. Brat:

a web-based tool for nlp-assisted text annotation. En *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 102–107.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, páginas 2214–2218, Istanbul, Turkey, may. European Language Resources Association (ELRA).

WeAreSocial y Hootsuite. 2022. Digital report espaÑa 2022: Nueve de cada diez espaÑoles usan las redes sociales y pasan casi dos horas al dÍa en ellas.

Wiegand, M., J. Ruppenhofer, A. Schmidt, y C. Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 1046–1056.

Xu, J.-M., K.-S. Jun, X. Zhu, y A. Bellmore. 2012. Learning from bullying traces in social media. En *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, páginas 656–666.