DETECCIÓN DE MENSAJES DE ODIO EN TWITTER: UN MODELO BASADO EN BERT PARA LA CLASIFICACIÓN DEL DISCURSO DE ODIO EN ESPAÑOL

Gloria del Valle; Escuela Politécnica Superior, Universidad Autónoma de Madrid; *Lara Quijano-Sánchez,* Escuela Politécnica Superior, Universidad Autónoma de Madrid; *Jesús Gómez,* Oficina Nacional de Lucha Contra los Delitos de Odio, Dirección General de Coordinación y Estudios, Secretaría de Estado de Seguridad (Ministerio del Interior, España).

RESUMEN

Las redes sociales son una fuente prácticamente inagotable de datos que podemos utilizar para medir fenómenos que ocurren dentro de ellas, como la generación y viralización de contenido ofensivo y de odio por parte de los usuarios en Twitter. Este trabajo propone ahondar en los esfuerzos para la detección de odio y su prevención en la red social Twitter, estudiando la inclusión del modelo BERT. Para ello se crea HaterBERT, el cual mejora la clasificación textual de modelos anteriores creados expresamente para el idioma español.

PALABRAS CLAVE: PLN, Discurso de odio, Twitter, Análisis de la red social, BERT.

ABSTRACT

Social networks have become an endless and inexhaustible source of data so we can use the resources they offer to analyse phenomena that occur within them, such as the generation and viralization of offensive and hateful content by users on Twitter. This project comprises a proposal for a BERT-based approach for hate speech detection on Twitter. For that purpose, we develop HaterBERT, which improves textual classification of previous models created specifically for the Spanish language.

KEYWORDS: NLP, Hate Speech, Twitter, Social Network Analysis, BERT.

1. Introducción

Diariamente se publican una gran cantidad de mensajes en las redes sociales que tienen como objetivo promover y alimentar el odio contra determinadas personas o un grupo de individuos, siendo este uno de los problemas más graves de la era digital. Este fenómeno se puede encontrar en todo tipo de red social, sin embargo, es especialmente en Twitter donde los usuarios, que expresan libremente sus opiniones sin ningún tipo de censura o filtro, encuentran más facilidad para transmitir mensajes de carácter ofensivo, utilizando en muchas ocasiones cuentas anónimas. Por tanto, la lucha contra distintas conductas discriminatorias basadas en prejuicios como misoginia, xenofobia, *ciberbullying* y racismo entre otros, es, sin duda alguna, un reto social necesario para construir una mejor sociedad.

Sin embargo, enfrentarse al problema de la detección del discurso de odio en español es un verdadero desafío. Numerosos estudios han abordado esta problemática en inglés (POLETTO, BASILE, SANGUINETTI, BOSCO, & PATTI, 2021) y desde la perspectiva plurilingüe (ALURU, MATHEW, SAHA, & MUKHERJEE, 2020), sin embargo, el español es uno de los idiomas más olvidados en este problema (PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA, 2021; PEREIRA-KOHATSU, QUIJANO-

SÁNCHEZ, LIBERATORE, & CAMACHO-COLLADOS, 2019) y a su vez uno de los más hablados en todo el mundo. Se trata de la segunda lengua materna tras el chino mandarín, y la cuarta lengua en hablantes tras el chino mandarín, inglés e hindi, con 586 millones de hablantes en todo el mundo.

Es por ello que afrontar el reto en español supone un desafío en el campo de la investigación. Además, pocos son los estudios que han tratado el tema y los conjuntos de datos (*datasets*) para entrenar los modelos, de ahí que la necesidad de trabajar en esta problemática es cada vez mayor dada la relevancia a nivel sociocultural. De hecho, en los últimos años, el número de delitos de odio en España lleva una tendencia alcista según el *Informe 2019 sobre la evolución de los delitos de odio en España* (MINISTERIO DEL INTERIOR, 2019), se registran 1.706 hechos, un 6,8% más que en 2018, principalmente de racismo y xenofobia. Asimismo, son la discriminación, las amenazas e injurias los hechos delictivos que más se repiten, siendo Internet (54,9%) y las redes sociales (17,2%) los medios más empleados para la ejecución de los mismos.

Tras una amplia revisión de la tecnología existente más puntera, lo que se denomina el *estado del arte*, sobre la detección del discurso de odio se ha detectado que más de 32 estudios, con sus correspondientes modelos asociados en muchos casos, son en la red social Twitter (POLETTO, BASILE, SANGUINETTI, BOSCO, & PATTI, 2021), dentro de los cuales se observan 10 para la detección de odio en inglés y 3 en español. Por tanto, se detecta una falta de investigación en esta rama de la ciencia computacional aplicada a retos sociales, por lo que es necesario ahondar más en la construcción de un buen modelo en español.

Por todo ello, el objetivo de este trabajo es la creación de un sistema inteligente que mejora la clasificación textual del discurso de odio dentro de la red social Twitter. Concretamente se presenta HaterBERT, un modelo basado en BERT, transformador que reconoce el discurso de odio (*hate speech*) en Twitter. Este algoritmo mejora el anterior de base llamado *HaterNet* (PEREIRA-KOHATSU, QUIJANO-SÁNCHEZ, LIBERATORE, & CAMACHO-COLLADOS, 2019), un sistema inteligente que combina un *Long-Short Term Memory* (LSTM) con *un Multilayer perceptron* (MLP) que se desarrolló en colaboración con la Oficina Nacional de Lucha Contra los Delitos de Odio (ONDOD), dependiente de la Secretaría de Estado de Seguridad (Ministerio del Interior, España). Como se verá a lo largo de esta comunicación, el algoritmo presentado en este trabajo mejora de un 3% a un 27% los algoritmos creados anteriormente (PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA, 2021; ALURU, MATHEW, SAHA, & MUKHERJEE, 2020; PEREIRA-KOHATSU, QUIJANO-SÁNCHEZ, LIBERATORE, & CAMACHO-COLLADOS, 2019).

2. Revisión del estado del arte

Como se ha mencionado previamente, los estudios realizados en español son escasos. En la Tabla 1 que se muestra a continuación se reflejan las publicaciones realizadas para el idioma español.

Dataset	Paper	Fecha	Mejor modelo	Validación	F1-score
HaterNet	(PEREIRA-KOHATSU, Y OTROS, 2019)	Oct. 2019	LSTM+MLP	LOOCV	0.611
HaterNet	(ALURU, Y OTROS, 2020)	Abr. 2020	mBERT	70-20-10	0.733
HatEval (es)					0.734
HaterNet	(PLAZA-DEL-ARCO, Y OTROS, 2021)	Mar. 2021	BETO (BERT)	10k-Fold	0.772
HatEval (es)					0.776

Tabla 1: Resumen de las principales características de los modelos realizados para el idioma español en la literatura. Se indican los *datasets* utilizados para cada uno de los estudios, así como los modelos utilizados, con su validación y puntuación correspondientes.

En el estudio realizado por PEREIRA-KOHATSU, QUIJANO-SÁNCHEZ, LIBERATORE, & CAMACHO-COLLADOS (2019) se contrarresta el discurso de odio desde la perspectiva de la víctima. Se trata del primer sistema inteligente que monitoriza y visualiza el odio en la red social. Tras el análisis de diferentes aproximaciones, el mejor modelo que se obtuvo estaba basado en LSTM+MLP que realiza un preprocesamiento del texto de entrada enriquecido por la técnica TF-IDF. En este preprocesamiento se dividen los datos de entrada en palabras, *emojis* y *embeddings* (representación de palabras en vectores) del tweet, siendo estos últimos obtenidos por la técnica *word2vec*. Además, estos autores introdujeron el *dataset* de HaterNet (ver en Tabla 2). Hace dos años, cuando se realizó el estudio, no se contempló la posibilidad de incluir transformadores para la detección del odio, ni tampoco existía una primera versión de BERT en español, la cual llegó en 2020 (CAÑETE, Y OTROS, 2020). Sin embargo, sí se habían probado en diferentes estudios el buen funcionamiento de modelos basados Deep Learning (BADJATIYA, GUPTA, GUPTA, & VARMA, 2017) muchos de ellos combinados con preprocesamientos basados en *word2vec embeddings*, *n-grams* (subsecuencia de *n* elementos de una secuencia dada) y diferentes recursos léxicos.

En un estudio distinto se probó la eficacia de cuatro modelos en diferentes idiomas (ALURU, MATHEW, SAHA, & MUKHERJEE, 2020): MUSE + CNN-GRU, Translation + BERT, LASER + LR y mBERT con una amplia optimización de hiperparámetros. Además, se comprobó la eficacia de utilizar mBERT en numerosos idiomas, a pesar de no ser el más apropiado en cuestión para cada uno de ellos.

Los resultados de realizar una traducción previa a BERT no son del todo lejanos a los resultados con mBERT, pero BERT está entrenado para el inglés y la precisión depende mucho de la calidad de la traducción. Por otro lado, se observó también que la utilización de transformadores es mucho más útil para *datasets* con suficiente información, mientras que para los corpus más pequeños los resultados con LASER + LR pueden ser más prometedores.

Es en 2021 cuando se ha realizado la primera publicación oficial en el dominio que nos atañe con BETO, según PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA (2021) alcanzando la mejor puntuación en las tareas de clasificación hasta la fecha, demostrando que la utilización de un BERT en español presenta una mejor adecuación para este idioma. En esta investigación, también se estudiaron los resultados de otros modelos tradicionales basados en Machine Learning y Deep Learning, además de la comparativa con XLM y mBERT. Tras ello, demostraron que las redes neuronales CNN, LSTM y BiLSTM son mejores entendiendo el contexto que los modelos clásicos de Machine Learning, no obstante, los transformadores siguen siendo los más adecuados hasta la fecha.

A raíz de estos resultados dispares, se plantea la construcción de un modelo basado en transformadores, en concreto BERT, estudiando diversas modificaciones en su diseño y diferentes configuraciones de hiperparámetros, lo que servirá como algoritmo base sobre el cual se trabajará en un futuro.

3. Metodología

Para entender el marco social y jurídico de este proyecto se debe perfilar lo que se ha de entender por discurso de odio. En este trabajo se entenderá el concepto en el marco de un delito de odio según lo define la OSCE (2014):

Un delito de odio es toda infracción penal, incluidas las cometidas contra las personas o la propiedad, dónde el bien jurídico protegido, se elige por su, real o percibida, conexión, simpatía, filiación, apoyo o pertenencia a un grupo. Un grupo se basa en una característica común de sus miembros, como su "raza", real o percibida, el origen nacional o étnico, el lenguaje, el color, la religión, la edad, la discapacidad, la orientación sexual, u otro factor similar.

De esta manera, el proyecto se define como un problema de clasificación binaria, en odio (*hate)* o no odio (*non-hate)*.

A continuación, se explica el diseño de HaterBERT, un modelo que se presenta de manera novedosa en este documento para la clasificación textual del discurso de odio en español. Es preciso recalcar que este modelo está basado en BERT (Véase DEVLIN, CHANG, LEE, & TOUTANOVA, 2019).

Para la creación del algoritmo se han utilizado las librerías de Tensorflow (Keras) y Pytorch. Además, se ha hecho uso de la librería de HuggingFace (transformers), la cual pone a disposición diversas herramientas para NLP y transformadores pre-entrenados, lo que ha sido imprescindible para añadir a HaterBERT el modelo pre-entrenado de BETO (CAÑETE, Y OTROS, 2020), BERT para el idioma en español.

De la misma forma se ha necesitado entender cómo realizar un ajuste de BERT (*Fine-Tuning)* adecuado al problema de la clasificación textual. Para poder abordar el problema, se ha elegido como referencia la implementación de BERT Classifier de ALURU, MATHEW, SAHA, & MUKHERJEE (2020), ya que resulta ser intuitiva de ajustar y proporciona buenos resultados para el problema plurilingüe, si bien se realizan los cambios oportunos en el ajuste de hiperparámetros.

Además, PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA (2021) también realizaron un estudio en español con BETO. En su publicación no se detalla en concreto temas sobre la implementación ni el *Fine-Tuning*, si bien se conoce que eligen el modelo BETO sin sensibilidad con las palabras en mayúsculas. Sin embargo, en este estudio se ha partido de la idea de que la sensibilidad de las mayúsculas puede aportar más información sobre la polaridad del tweet, por ende, al problema de la detección de odio, ya que los usuarios pueden utilizarlas para expresar más énfasis en sus ideas. También se encuentran diferencias en la optimización de hiperparámetros, cuya optimización se detallará en la sección de resultados.

Datasets

La variedad de *datasets* es pequeña. Tan solo existen dos *datasets* en español, si bien recientemente se publicó la tarea Profiling Hate Speech Spreaders on Twitter, la cual ofrecía en ficheros agrupados 200 tweets por cada 200 autores, siendo estos últimos el objetivo a etiquetar. Dado que los tweets no estaban etiquetados, se tuvo que descartar el *dataset* por contextualizar un problema diferente al presente trabajo.

Los *datasets* en español provienen de HaterNet (PEREIRA-KOHATSU, QUIJANO-SÁNCHEZ, LIBERATORE, & CAMACHO-COLLADOS, 2019) (ver en Tabla 2) y HatEval en español (BASILE, Y OTROS, 2019) (ver en Tabla 3). Ambos, publicados en 2019 con apenas tres meses de diferencia (primero HatEval y después HaterNet), presentan una clasificación binaria del discurso del odio (*hate* vs *non-hate*) con un número similar de tweets.

El *dataset* de HaterNet proporciona un conjunto de datos completo y etiquetado. Contiene 6000 tweets etiquetados procedentes de 2 millones de tweets descargados entre febrero y diciembre del 2017 sin etiquetar. El *dataset* fue filtrado aplicando seis diccionarios de *hate speech* y uno de insultos. Tras ello, fue etiquetado a mano por cuatro expertos procedentes de diferentes escenarios, utilizando el voto por mayoría. Según la *Kappa de Fleis* este *dataset* tiene una puntuación de 0.588, lo que quiere decir que se encuentra entre un valor de acuerdo moderado bastante fiable. Los detalles del *dataset* se encuentran presentados en la Tabla 2.

Clase	Descargados	Seleccionados	Etiquetados
0	-	-	4433
1	-	-	1567
Total	2 millones	8710	6000

Tabla 2: Distribución del *dataset* de HaterNet, donde se refleja que un 73.89% del corpus pertenece al etiquetado *non-hate* y un 26.11% de *hate*.

El segundo *dataset* fue propuesto en la tarea publicada por BASILE, Y OTROS (2019), la cual consistía en la detección de discurso de odio contra dos objetivos: mujeres e inmigrantes. Los datos fueron extraídos principalmente desde julio hasta noviembre de 2017 y desde julio hasta septiembre de 2018 utilizando la plataforma *crowdsourcing* Figure Eight (F8). Este *dataset* reportó suficiente fiabilidad, con una puntuación de 0.89, lo que indica un valor de acuerdo casi perfecto. Posteriormente se reforzó el etiquetado

por dos expertos, hablantes del castellano, y utilizando finalmente voto por mayoría, aunque no consta puntuación final sobre su fiabilidad.

Clase	Train	Test	Total
0	2981	940	3861
1	2079	660	1567
Total	5000	1600	6000

Tabla 3: Distribución del *dataset* de HatEval (es), donde se refleja que un 58.5% del corpus pertenece al etiquetado *non-hate* y un 41.5% de *hate*.

Este *dataset*, ver en Tabla 3, se encuentra divido en tres conjuntos: entrenamiento, prueba y validación (*Stratified Sample*), juntando el entrenamiento y validación en *train* y apartando los datos de prueba en *test.*

Preprocesamiento

Con el objetivo de poder aprovechar los recursos de pre-trained BERT (BERT pre-entrenado) es necesario utilizar el *tokenizer* llamado *BertTokenizer*, proporcionado por la librería *transformers*, ya que dicho transformador tiene un vocabulario fijo específico y una forma particular de transformar las palabras en *tokens* y en máscaras, si bien se realizan las siguientes modificaciones para cada entrada de texto en el *encode*:

- *Tokenizar* la oración.
- Añadir el *token* [CLS] al principio de la oración.
- Añadir el *token* [SEP] al final de la oración.
- Asignar los *tokens* a sus IDs correspondientes.

Después de ello se realiza un *padding* para garantizar que todas las secuencias tengan la misma longitud, rellenando 0 al final de cada secuencia hasta que cada secuencia tenga la misma longitud que la secuencia más larga.
A continuación, para cada oración se crea la máscara de atención para los identificadores correspondientes. Se decide que:

- Si el ID es 0, entonces es *padding* y se activa en la máscara como un 0.
- Si el ID mayor o igual que 0, entonces es un *token* y se activa a 1.

Tras ello, se convierten las salidas en tensores para poder ser tratados. De esta manera son aptas para generar el *DataLoader* de Pytorch, lo que ayuda a ahorrar memoria durante el entrenamiento y aumenta la velocidad del mismo.

Clasificador

Afortunadamente, la librería de *transformers* pone a disposición una serie de interfaces diseñadas para tareas NLP, adaptando las entradas y las salidas dependiendo de la necesidad del problema. En este caso, es necesario utilizar *BertForSequenceClassification*, clase que contiene una capa de entrada adaptada para

secuencias de texto u oraciones. De esta manera se puede utilizar BERT para el análisis de sentimiento, en concreto para el problema de la detección de discurso de odio (ver en Figura 1).

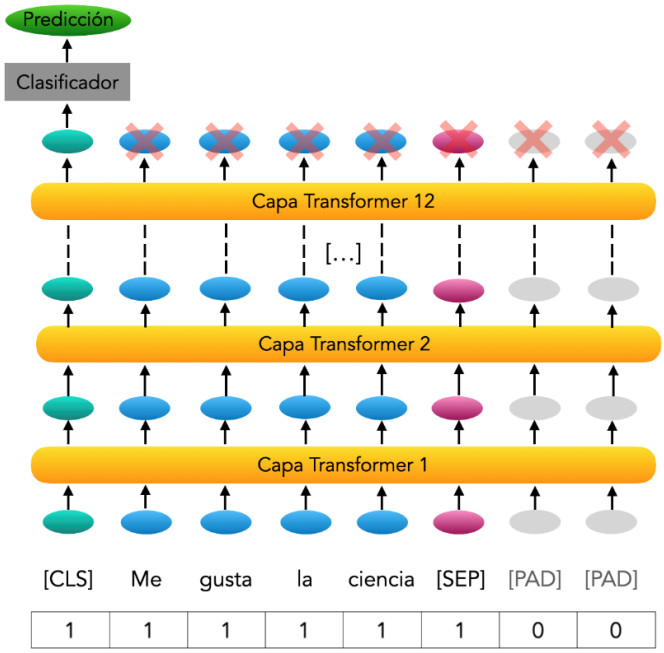

Figura 1: Ilustración del formato del preprocesamiento de una frase para el *Fine-Tuning* con BertForSequenceClassification. Se observa que la secuencia de dígitos corresponde a la máscara de atención y que a la frase se le ha realizado un *padding* para tener el mismo tamaño que todas las demás oraciones del *dataset*.

Análogamente, se puede indicar en el *path* qué modelo BERT utilizar de la librería de transformers. Este será BETO (dccuchile/bert-base-spanish-wwm-cased) (CAÑETE, Y OTROS, 2020).

Para poder ajustar el clasificador es imprescindible utilizar un optimizador. Se utiliza *AdamW*, una muy buena opción para BERT que además utiliza el propio BERT pre-entrenado. Se hablará de los hiperparámetros probados referentes a *AdamW* en la sección 4 de resultados.

4. Resultados

En esta sección se detallan y analizan los resultados obtenidos en los diferentes experimentos realizados sobre HaterBERT. En primer lugar, se muestra la optimización de hiperparámetros, siguiendo tras ello la exposición de los diferentes resultados obtenidos con una comparativa con el estado del arte en español (con su configuración de hiperparámetros correspondiente).

Es importante resaltar que las valoraciones de los resultados se realizan tomando en cuenta como principal métrica *F1-score*, ya que los falsos negativos y los falsos positivos son más cruciales en diversos temas como la detección de odio, si bien se valoran otras métricas.

Con el fin de recoger los mejores resultados posibles, se realizan diferentes experimentos con distintos hiperparámetros. Los hiperparámetros probados se muestran en la Tabla 4.

Hiperparámetros	Opciones
Épocas	[2,3,4,5]
Learning rate	[2e-5, 3e-5, 5e-5]
Random seed	**2018**, 2019, 2020, 2021, 2022, 2023
Batch size	[16, 32]
Epsilon	[1e-6, 1e-8]
Max length	256

Tabla 4: Detalle de los hiperparámetros probados para el modelo de HaterBERT.

En la Tabla 5 se detallan los resultados obtenidos con el *dataset* de HaterNet (6000 tweets, ver Tabla 1) además de la comparativa del rendimiento obtenido respecto a PEREIRA-KOHATSU, QUIJANO-SÁNCHEZ, LIBERATORE, & CAMACHO-COLLADOS (2019).

Modelo	Autor	Validación	Hiperparámetros	F1-score
LSTM+MLP	(PEREIRA-KOHATSU, Y OTROS, 2019)	LOOCV	-	0.6110
HaterBERT	Esta propuesta	LOOCV	Épocas: 5, Batch Size: 32, Learning Rate: 5e-5, Epsilon: 1e-6	0.9989

Tabla 5: Resultados y comparativa con el estado del arte de HaterBERT con PEREIRA-KOHATSU, QUIJANO-SÁNCHEZ, LIBERATORE, & CAMACHO-COLLADOS (2019). Se utiliza el *dataset* de HaterNet (ver Tabla 2).

Se puede comprobar que se alcanzan buenos resultados para el *dataset* y que la utilización de BETO mejora considerablemente el problema. Si bien cabe valorar la relevancia de los resultados, se ha de entender que *LOOCV* emplea *n-1* datos para entrenar el modelo, lo que implica prácticamente todo el *dataset*. En términos prácticos lo que realiza es un mayor ajuste del modelo a los datos disponibles y reduciendo el sesgo, lo que puede conllevar a un mayor riesgo de *overfitting* (sobre-ajuste) y de varianza. Es por ello que se realizan más comparaciones respecto al estado del arte con otros tipos de validación: *Stratified Sample* con diferentes divisiones, y *k-fold Cross Validation.*

En la Tabla 6, se detallan los resultados obtenidos con una división *Stratified Sample* del *dataset* de HaterNet de 70% para el entrenamiento, 20% para el *test* y 10% para la validación del modelo. Además, se comparan los resultados con el estudio de ALURU, MATHEW, SAHA, & MUKHERJEE (2020), donde utilizan mBERT.

Nuevamente se observa que los resultados mejoran con la utilización de BETO, además de que a una mayor entrada del *dataset* se mejora la salida. Es preciso destacar que, aunque mBERT pueda ofrecer buenos resultados a nivel general, el problema que tiene

es que está pre-entrenado sobre un conjunto de corpus monolingües de diferentes idiomas, por lo que no proporciona un mecanismo de detección del idioma en cuestión, y el *token* se puede confundir con otro idioma fácilmente. Sin embargo, BETO fue pre-entrenado con un conjunto de datos específicamente en el idioma español, por lo que es mucho más apropiado en conjuntos de datos en español.

Modelo	Autor	Validación	Hiperparámetros	F1-score
mBERT	(ALURU, Y OTROS, 2020)	70-20-10	-, Max length: 128	0.7329
HaterBERT	Esta propuesta	70-20-10	Épocas: 5, Batch Size: 32, Learning Rate: 5e-5, Epsilon: 1e-6	0.7667

Tabla 6: Resultados y comparativa con el estado del arte de HaterBERT con ALURU, MATHEW, SAHA, & MUKHERJEE (2020). Se utiliza el *dataset* de HaterNet (ver Tabla 2).

También se realiza una comparativa con el estudio de PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA (2021), el cual se muestra en la Tabla 7. En ella se realizan diferentes experimentos con una validación de *10k-Fold CV* sobre el *dataset* de HaterNet. Este método es también un proceso iterativo al igual que *LOOCV*, aunque emplea menos observaciones como entrenamiento. Sin embargo, obtiene una estimación más precisa del error de *test* gracias a un mejor balance entre el sesgo y la varianza, ya que emplea *k-1* grupos para entrenar el modelo y el restante para la validación. Gracias a esto, se obtiene una mejor perspectiva de los resultados obtenidos, valorando así la calidad de los mismos.

Modelo	Autor	Validación	Hiperparámetros	F1-score
BETO	(PLAZA-DEL-ARCO, Y OTROS, 2021)	10k-Fold	Épocas: 2, Batch Size: 16, Learning Rate: 2e-5, Max length: 80	0.6580
HaterBERT	Esta propuesta	10k-Fold	Épocas: 5, Batch Size: 32, Learning Rate: 5e-5, Epsilon: 1e-6, Max length: 256	0.9778

Tabla 7: Resultados y comparativa con el estado del arte de HaterBERT con PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA (2021). Se utiliza el dataset de HaterNet (ver Tabla 2).

En el estudio de PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA (2021), también se realizaron pruebas sobre el *dataset* de HatEval en español con el mismo método de validación (*10K-Fold CV*). En la Tabla 8 se muestra la comparativa de la presente propuesta con su estudio.

Modelo	Autor	Validación	Hiperparámetros	F1-score
BETO	(PLAZA-DEL-ARCO, Y OTROS, 2021)	10k-Fold	Épocas: 3, Batch Size: 16, Learning Rate: 2e-5, Max length: 80	0.7553

HaterBERT	Esta propuesta	10k-Fold	Épocas: 3, Batch Size: 16, Learning Rate: 2e-5, Max length: 256	0.8673

Tabla 8: Resultados y comparativa con el estado del arte de HaterBERT con PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA (2021). Se utiliza el dataset de HatEval (es) (ver Tabla 3).

A diferencia de ALURU, MATHEW, SAHA, & MUKHERJEE (2020) y PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA (2021), se han realizado unas configuraciones diferentes de hiperparámetros, las cuales se detallan en las tablas 5, 6, 7 y 8. Además de esto, se ha elegido un *max length* de 256, lo que hace que el algoritmo tarde más tiempo en procesar la entrada, pero también puede mejorar la clasificación. También es destacable mencionar que en PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA (2021) no se menciona qué tipo de ajuste sobre el preprocesamiento de BERT realizan, si bien se conoce la utilización de BETO y su diferente ajuste de hiperparámetros. Se puede extrapolar que el preprocesamiento basado en la implementación de ALURU, MATHEW, SAHA, & MUKHERJEE (2020) mejora los resultados de la clasificación en el idioma español, si bien en este caso se utiliza BETO en vez de mBERT, como se ha mencionado previamente.

5. Conclusiones

En un mundo polarizado, las redes sociales muestran ser un arma de doble filo con la aparición de fenómenos como el discurso de odio. En el presente trabajo se ha detectado y analizado la presencia del mismo en la red social Twitter. Para ello se ha construido HaterBERT, un modelo de inteligencia artificial que usa técnicas de Procesamiento del Lenguaje Natural, PLN (NLP) y el transformador BERT, que mejora de un 3% a un 27% los resultados de los actuales modelos que han analizado el discurso de odio en español (ALURU, MATHEW, SAHA, & MUKHERJEE, 2020; PLAZA-DEL-ARCO, MOLINA-GONZÁLEZ, UREÑA-LÓPEZ, & MARTÍN-VALDIVIA, 2021; PEREIRA-KOHATSU, QUIJANO-SÁNCHEZ, LIBERATORE, & CAMACHO-COLLADOS, 2019).

En el futuro se plantea el reto de incluir un clasificador de imágenes para el multimedia que comparten los usuarios, ya que en muchas ocasiones los usuarios utilizan comunicación no verbal en sus mensajes ofensivos.

También se podría estudiar el contexto del propio tweet en sí, esto es, conocer a qué responde exactamente en caso de ser una respuesta, y además, obtener todos los tweets conectados en caso de ser un hilo. Para ello se podría crear un árbol de respuestas con diversas herramientas que existen en Python.

De manera paralela, se podría estudiar la aportación de optimizadores como LAMB (NVLAMB, YOU, Y OTROS, 2019) con el fin de escalar el entrenamiento con BERT con las GPU de NVIDIA, ya que el tiempo de entrenamiento de estos modelos es alto a medida que el tamaño del *dataset* incrementa.

Bibliografía

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation* (pp. 54-63). Association for Computational Linguistics.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. *PML4DC at ICLR 2020*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805.*

Ministerio del Interior, España. (2019). *Informe 2019 sobre la evolución de los delitos de odio en España.* Ver aquí.

Office for Democratic Institutions and Human Rights (ODIHR), O. S. C. E. (2014). *What is hate crime*. OSCE. https://hatecrime.osce.org/what-hate-crime.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and Monitoring Hate Speech in Twitter. *Sensors,* 19(*21*), 4654.

Plaza-del-Arco, M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications, 166*, 114120.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 1-47.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., ... & Hsieh, C. J. (2019). Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.