ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.1077

Aggression and Hate in Spanish Text Messages. Identification Using a Pre-Trained Transformer Model.

César Espin-Riofrio, MSc.[1], Josue San Martín Torres, Ing.[1], Helen Rodríguez-Soria, Ing.[1], Verónica Mendoza Morán, MSc.[1], Angélica Cruz Chóez, MGs.[1], Arturo Montejo-Ráez, PhD.[2]
[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, roberto.sanmartint@ug.edu.ec, helen.rodriguezs@ug.edu.ec, veronica.mendozam@ug.edu.ec, angelica.cruzc@ug.edu.ec
[2]Universidad de Jaén, España, amontejo@ujaen.es

Abstract– ***Nowadays, social networks have given rise to the free expression of opinions and thoughts in real time, however, this can also lead to negative interactions, such as bullying, discrimination and other aggressive and hateful behaviour. To address this issue, different Natural Language Processing (NLP) methods and techniques exist. In this paper, a quasi-experimental investigation was carried out using the pre-trained Pysentimiento Transformer model to detect the presence of aggression and hate in Spanish text messages on the social network Twitter. Data extraction and processing tools were used to ensure the quality of the data before it was run through the model. A web interface was also designed to present the information obtained through graphs and tables, allowing a clear assessment of the detection of aggressive and hateful content in text messages through various analysis criteria. It is shown that it is possible to detect aggression and hatred in text messages using a Transformer model pre-trained for the task and use it to create systems or applications that detect and quantify these symptoms in messages written by people.***

Keywords-- Aggressiveness and hatred, Transformer models, Natural Language Processing.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Agresividad y Odio en Mensajes de Texto en Español. Identificación Usando un Modelo Transformer Pre-Entrenado.

César Espin-Riofrio, MSc.[1], Josue San Martín Torres, Ing.[1], Helen Rodríguez-Soria, Ing.[1], Verónica Mendoza Morán, MSc.[1], Angélica Cruz Chóez, MGs.[1], Arturo Montejo-Ráez, PhD.[2]
[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, roberto.sanmartint@ug.edu.ec, helen.rodriguezs@ug.edu.ec, veronica.mendozam@ug.edu.ec, angelica.cruzc@ug.edu.ec
[2]Universidad de Jaén, España, amontejo@ujaen.es

Resumen– ***En la actualidad, las redes sociales han dado lugar a la libre expresión de opiniones y pensamientos en tiempo real, sin embargo, esto también puede conducir a interacciones negativas, como el acoso, la discriminación y otros comportamientos agresivos y de odio. Para abordar esta problemática, existen diferentes métodos y técnicas de Procesamiento de Lenguaje Natural (PLN). En el presente artículo se llevó a cabo una investigación cuasi experimental utilizando el modelo Transformer pre-entrenado Pysentimiento para detectar la presencia de agresividad y odio en los mensajes de texto en español de la red social Twitter. Se emplearon herramientas para la extracción y procesamiento de los datos, asegurando así la calidad de estos antes de pasar por el modelo. También se diseñó una interfaz web para presentar la información obtenida mediante gráficos y tablas, lo que permitió una evaluación clara de la detección de contenido agresivo y de odio en los mensajes de texto a través de varios criterios de análisis. Se demuestra que es posible detectar agresividad y odio en mensajes de texto mediante un modelo Transformer pre-entrenado para la tarea, y utilizarlos para crear sistemas o aplicaciones que detecten y cuantifiquen estos síntomas en mensajes escritos por las personas.***

Palabras clave-- Agresividad y odio, modelos Transformer, Procesamiento de Lenguaje Natural.

I. INTRODUCCIÓN

En los últimos años, el uso de las redes sociales se ha convertido en una parte esencial de la vida cotidiana porque han experimentado un rápido crecimiento. Plataformas como Twitter permiten que las personas sean capaces de expresar de manera libre sus pensamientos, ideas y opiniones de los diferentes temas, así mismo estas pueden ser leídas, replicadas y comentadas por otras personas en todo el mundo. Sin embargo, el auge de las redes sociales ha traído una serie de impactos negativos en la sociedad, ya que, en ocasiones, las interacciones que se dan en las plataformas pueden expresar agresividad u odio y desencadenar muchos problemas. Bajo este contexto, el Procesamiento de Lenguaje Natural (PLN) se ha convertido en una herramienta clave para analizar el comportamiento de los usuarios en las redes sociales. En particular, el análisis de sentimientos y emociones en los mensajes publicados en Twitter puede ayudar a identificar patrones de comportamiento y detectar la agresividad y el odio en las interacciones.

El objetivo principal de este trabajo de investigación es detectar la presencia de agresividad y odio en los mensajes de texto en español, utilizando un modelo Transformer pre-entrenado y evaluar su rendimiento para este tipo de tarea. Detectar estos comportamientos a tiempo es crucial, debido a que la red social de Twitter es una plataforma donde los usuarios comparten experiencias y opiniones y a su vez hay usuarios que exponen su inconformidad tendiendo a ser ofensivos provocando depresión en las víctimas. Al obtener un mejor conocimiento de estos patrones, se pueden desarrollar estrategias más efectivas para fomentar un diálogo saludable y combatir la propagación de mensajes dañinos en las redes sociales, contribuyendo a un entorno más positivo y respetuoso para todos los usuarios.

Se llevará a cabo una investigación bibliográfica de las contribuciones científicas relevantes referentes a la detección de agresividad y odio en mensajes de texto, con el propósito de determinar los métodos y técnicas más utilizadas por los autores en este tema.

Experimentamos con el modelo Pysentimiento basado en Transformer que está calificado para analizar sentimientos y clasificarlos en positivo, negativo y neutral, siendo también capaz de detectar emociones como alegría tristeza, ira y miedo [1]. Pysentimiento[2]es un modelo de procesamiento de lenguaje natural basado en la variante del modelo RoBERTa, que ha sido entrenado con el corpus TASS 2020, que contiene aproximadamente 5.000 tweets en varios dialectos del idioma español. Este modelo ha demostrado una precisión del 75% en la detección de mensajes de texto con contenido de odio y agresividad.

De las investigaciones bibliográficas realizadas por diferentes autores sobre la identificación y clasificación de mensajes de texto en el idioma español, el estudio realizado por [3] se centra en el desarrollo de un sistema capaz de detectar Tweets agresivos en el idioma español de una manera binaria, donde este sería clasificado como agresivo o no. Para la elaboración de este sistema emplearon la técnica de aprendizaje automático Support Vector Machines (SVM).

En cambio, [4] utilizaron un enfoque de conjunto de votos para construir un modelo clasificador de agresividad, donde se

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

procedió a combinar los algoritmos de Decisión Tree (DT), Support Vector Machines (SVM) y Multinomial Naive Bayes (Multinomial NB), aunque también tuvieron en consideración los algoritmos de Random Forest (RF) y Logistic Regression (LR), finalmente los descartaron por ser menos eficientes en la tarea de detección de Tweets cuyo contenido sea agresivo. [5] en su trabajo investigaron el uso de una arquitectura de red de cápsulas multidimensionales para detectar la agresión y toxicidad en mensajes de texto la cual demostró que el uso de esta, en comparación a una red de cápsula de dimensión fija, le proporciona mejores resultados al momento de identificar agresión y toxicidad.

[6] mediante el uso combinatorio de algoritmos clasificadores como NB, SVM, LR y DT con el uso de secuencias de unigramas y bigramas, lograron observar que las puntuaciones individuales de cada clasificador para la tarea de detección de discursos de odio aumentan. El proceso para la detección de odio que abordaron [8] basada en algoritmos de Machine Learning como NB, DTM LR, RF, SVM, Long-Short Term Memory (LSTM), demostró que la clasificación basada en LSTM ofrece la mayor precisión para la tarea de detección de odio.

En el estudio realizado por [7] quienes emplean el uso del modelo ESGONLP-HSC para poder clasificar los textos de las redes sociales en tres clases que son el lenguaje neutral, ofensivo y de odio, dando como resultados valores de precisión mejorados en comparación con otros modelos como las redes neuronales profundas (DNN) y SVM.[8] utilizaron Multilingual Hatecheck (MHC) para mejorar la efectividad en la detección de la agresividad en diferentes idiomas, sin embargo, se ha observado que el modelo es sensible a las palabras y frases clave de cada idioma, lo que puede provocar errores en su identificación. El modelo híbrido de [9] para detectar publicaciones agresivas que contienen imágenes y texto en las redes sociales utiliza la técnica de RF y fue mejorado mediante la técnica de optimización de enjambre de partículas binarias (BPSO) la cual mejoró una precisión del algoritmo. [10] emplearon el Modelo Stands for Extreme Gradient Boosting (XGBoost) donde se observó que el rendimiento de este es pobre y necesita muchas modificaciones para mejorarlo.

[11] Utilizando dos modelos de deep learning LSTM y Bidirectional Long-Short Term Memory (Bi-LSTM) para la clasificación de contenidos en las redes sociales en discursos de odio y discursos normales, observaron que ambos modelos presentan una mínima diferencia en sus valores, aunque LSTM siempre fue el más efectivo para esta tarea. Por otra parte, [12] emplearon LR, convolutional neural networks (CNN) y Bi-LSTM para etiquetar mensajes de un conjunto de datos en tres categorías: agresivos, encubiertos agresivos y no agresivos, donde de manera independiente el mejor modelo fue Bi-LSTM obteniendo mejor puntuación que los otros dos. Para la identificación y categorización del lenguaje ofensivo en las redes sociales [13] emplearon tres modelos que fueron CNN, BiLSTM y SVM los cuales mostraron resultados muy similares para esta tarea donde quien tiene una pequeña ventaja son las CNN. El modelo combinatorio de los métodos de [14] es capaz de detectar comentarios agresivos en la red social Twitter, sin embargo, los resultados demuestran que el modelo más simple basado en Term Frequency times Inverse Document Frequency (TF-IDF) es más efectivo que el modelo más complejo que combina TF-IDF, Multi Layer Perceptron (MLP), CNN y LSTM. [15] mediante el uso de CNN, GRU y CNN+GRU para detectar discursos de odio, tuvieron resultados prometedores, pero CNN superó con éxito a los otros modelos. El uso de TF-IDF en las CNN, crea una mejora significativa a comparación a los modelos SVM y LR [16].

Por otro lado, [17] emplearon un modelo de aprendizaje automático para detectar la agresión y odio a todo lo relacionado a la mujer (misoginia) en el idioma español, enfocándose en la utilización del modelo pre-entrenado BERT para identificar las relaciones existentes em el tema de agresividad y misoginia. [18]Proponen un modelo que aprovecha el uso de varios clasificadores clásicos utilizando como características la combinación de medidas de estilometría con incrustaciones obtenidas a partir de un modelo RoBERTa español para la representación de textos. BERT [19] fue empleado debido a su capacidad de entender el significado de las palabras, el poder procesarlas y posteriormente analizar la información recolectada. [20] exploran la predicción de la complejidad léxica mediante la combinación de redes Transformer pre-entrenadas y ajustadas con distintos tipos de rasgos lingüísticos tradicionales.

[21]demostraron resultados prometedores para la identificación y categorización de eventos violentos en las redes sociales utilizando el modelo DistilBETO [22], que es una versión destilada de BETO que usa una arquitectura similar al modelo BERT-base, donde el corpus de entrenamiento es exclusivo en español, el cual puede ser ajustado finamente para una amplia gama de tareas de procesamiento de lenguaje natural, incluyendo la clasificación de texto, la extracción de información y la generación de texto. Sin embargo, la cantidad de datos con la cual se entrena el modelo es el factor que más influye en su rendimiento. En cambio, [23] mediante el uso del modelo BERT indicaron que su precisión en la clasificación de tweets agresivos, al igual que el caso anterior puede mejorar al utilizar una mayor cantidad de datos en su entrenamiento. En su estudio, [24] emplearon los modelos BERT y BETO para desarrollar un sistema de detección y clasificación de intenciones tóxicas en las redes sociales. Para adaptar los modelos a esta tarea, se realizaron ajustes específicos en cada uno de ellos. Los resultados demostraron que BETO fue el modelo que presentó un mejor desempeño para esta tarea en particular. [25] demostraron que al realizar un ajuste fino del modelo Transformer pre-entrenado BERT, y al aumentar el corpus de datos utilizados para su entrenamiento, se puede mejorar significativamente el rendimiento del modelo. Esto sugiere que el procesamiento cuidadoso de los datos es crucial para obtener resultados precisos en tareas de procesamiento de lenguaje natural utilizando modelos Transformer.

A continuación, detallamos la metodología empleada en el trabajo de investigación, así como las técnicas y métodos utilizados para la recolección y análisis de datos, seguido de la presentación de los resultados obtenidos con el modelo para la detección de agresividad y odio en los mensajes de texto. Luego e llevará a cabo una discusión y comparación de los resultados obtenidos. Finalmente, se presentarán las conclusiones y recomendaciones para futuros trabajos relacionados

II. METODOLOGÍA

Se aplica el método cuasi experimental que incluye la extracción de publicaciones en la red social Twitter y su evaluación con un modelo Transformer pre-entrenado como se muestra en la figura 1.

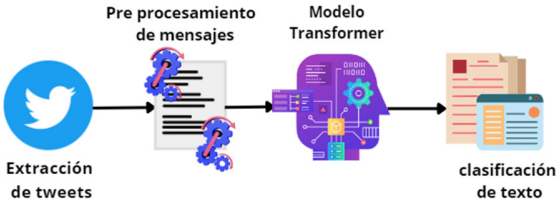

Fig. 1 Proceso de clasificación de texto

La figura 2 muestra el proceso llevado a cabo para la clasificación e identificación de agresividad y odio utilizando una interfaz web diseñada para presentar los resultados.

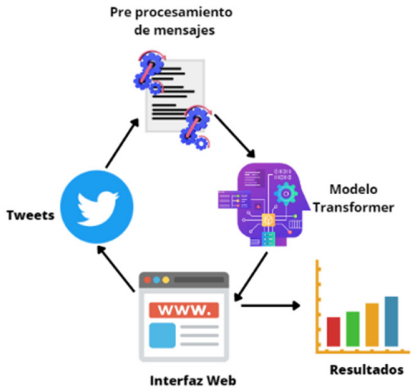

Fig. 2 Proceso de clasificación mediante una interfaz web.

A. *Extracción de Tweets*

Antes de empezar con la extracción de tweets, se creó una cuenta de desarrollador de Twitter que permite extraer publicaciones de los usuarios, por lo que se debe registrar y analizar qué tipo de nivel de acceso se requiere para la investigación.

Se utilizará la librería de Python llamada Tweepy que nos permitirá acceder a la API de Twitter. Para la recopilación de tweets se empleó el proceso indicado en Fig. 3.

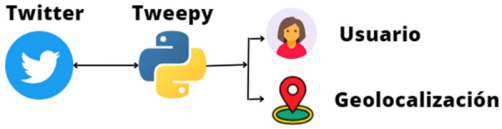

Fig. 3 Proceso de extracción de tweets

Se realizaron dos tipos de búsquedas específicas de mensajes de texto en español. Para la primera búsqueda se llevó a cabo una extracción de tweets de varios usuarios con el objetivo de efectuar un análisis posterior de los datos. Por lo cual, se verificó previamente que las cuentas de los usuarios seleccionados tuvieran su perfil y tweets públicos en la red social de Twitter. A continuación, se puede observar en la figura 4 los tweets que fueron extraídos de un determinado usuario.

	tweets
0	Boscán le metió hasta el brazo. Mis hijos me d...
1	Por error puse que la Clínica Alcívar mató a C...
2	Esto da asco. Este periodista español de @chir...
3	Él era el hombre de Cherrez en guayas\nAl que ...
4	Aquí con la familia Quiroz, personas buenas y ...
...	...
95	Mi consejo al presidente Lasso es que meta pre...
96	https://t.co/vgrV786qyr
97	Señora @FiscaliaEcuador ya tiene listo el alla...
98	Para el asesor del presidente Lasso que lo des...
99	Y esta gente del presidente Lasso es la que ha...

Fig. 4 Tweets extraídos de un usuario determinado.

En nuestra segunda búsqueda se extraen Tweets de una localidad con un radio de 80 km, permitiendo también la búsqueda sobre un tema. Se especifican los parámetros de búsqueda como lo son la latitud, longitud en el que se buscarán los tweets. La figura 5 muestra Tweets extraídos mediante el criterio de localidad.

	tweet
0	"Toda esta algarabía de la Asamblea no es comp...
1	Asamblea Nacional No Derogará Las Pensiones Vi...
2	Comisión Especializada Ocasional por la Verdad...
3	Asamblea Nacional no derogará pensiones vitali...
4	#Ecuador: Asamblea inhabilita a exministra de ...
...	...
95	#ATENCIÓN | El Presidente Lasso no asistirá a ...
96	.@VillaFernando_ asambleísta Nacional. Preside...
97	En #Cañar se encuentra un espacio de la Asambl...
98	Virgilio Saquicela presidente de la Asamblea N...
99	"No existen legisladores que hayan demostrado ...

Fig. 5 Tweets extraídos por localidad

B. *Procesamiento de los datos*

La información extraída de las redes sociales suele estar estructurada en un lenguaje informal y para superar los desafíos del proyecto en este aspecto, se utilizan expresiones regulares, las cuales son las encargadas de buscar y manipular la información dentro del conjunto de datos.

Este proceso incluye la eliminación de hashtags, menciones, URL, emoticones, tweets repetidos, espacios en blanco y la normalización de todo el texto a minúsculas, reduciendo así posibles errores en la identificación de agresividad y odio. En la figura 6 se puede apreciar los tweets luego de ser preprocesados.

	Mensaje
0	boscán le metió hasta el brazo mis hijos me di...
1	por error puse que la clínica alcívar mató cam...
2	esto da asco este periodista español de que se...
3	él era el hombre de cherrez en guayas al que l...
4	aquí con la familia quiroz personas buenas cri...
...	...
95	mi consejo al presidente lasso es que meta pre...
96	
97	señora ya tiene listo el allanamiento para est...
98	para el asesor del presidente lasso que lo des...
99	y esta gente del presidente lasso es la que ha...

Fig. 6 Tweets preprocesados.

C. *Clasificación de texto*

Acto seguido, se procede al análisis para la clasificación de los mensajes de texto, para ello se debe instalar las librerías necesarias del modelo Transformer Pysentimiento, en la figura 7 se ilustra como se instancia el modelo.

```
from pysentimiento import create_analyzer
```

Fig. 7 Descarga la biblioteca Pysentimiento

Luego, como se aprecia en la figura 8, al descargar la función se crea un objeto analyzer el cual permite utilizar la tarea de detección de odio y agresividad “hate_speech” para el idioma español que provee las tres posibles salidas odio, agresividad y direccionada (hateful, aggressive, targeted).

```
analyzer = create_analyzer(task="hate_speech", lang="es")
```

Fig. 8 Crear objeto analyzer

En última instancia, en la figura 9 se observa cómo se procede a analizar el texto con la función predict del modelo.

```
analyzer.predict("Los inmigrantes deberian ser deportados")
```

Fig. 9 Análisis de sentimientos del mensaje de texto

Los datos obtenidos del modelo Pysentimiento sobre ese texto se presentan en la figura 10, en el que etiqueta las salidas "hateful" y "aggressive" según lo determina, junto con su probabilidad de ocurrencia.

```
AnalyzerOutput(output=['hateful', 'aggressive'], probas={hateful: 0.963,
targeted: 0.021, aggressive: 0.768})
```

Fig. 10 Salida del modelo Pysentimiento

D. *Experimentación*

Utilizando el modelo Pysentimiento se evidencia, en la figura 11, como se detecta la agresividad y odio en los mensajes de texto de un determinado usuario, donde la columna de "Sentimiento" nos proporciona la salida del modelo más la puntuación de odio y agresividad.

	Mensaje	Sentimiento	Odio	Agresividad
0	boscán le metió hasta el brazo mis hijos me di...		2.058005	1.879044
1	por error puse que la clínica alcívar mató cam...		1.339606	1.334495
2	esto da asco este periodista español de que se...		3.355142	3.081783
3	él era el hombre de cherrez en guayas al que l...		1.568813	1.160378
4	aquí con la familia quiroz personas buenas cri...		3.169368	2.086074
...	...	...	...	...
95	mi consejo al presidente lasso es que meta pre...		13.503383	8.664530
96		odio - agresividad	66.211993	57.985008
97	señora ya tiene listo el allanamiento para est...	odio - ordinario - agresividad	84.985805	77.269316
98	para el asesor del presidente lasso que lo des...		4.462353	3.014222
99	y esta gente del presidente lasso es la que ha...		15.003161	11.765943

Fig. 11 Agresividad y odio de un usuario determinado

En figura 12, se pueden ver los Tweets analizados para una localidad y tema específico. Al igual que en el ejemplo anterior, el modelo indica los porcentajes de agresividad y odio en los mensajes de texto, así como la columna "Sentimiento" que muestra la salida del modelo.

	Mensaje	Sentimiento	odio	agresividad
0	toda esta algarabía de la asamblea no es comp...		1.324826	1.189874
1	asamblea nacional no derogará las pensiones vi...		2.619537	1.862345
2	comisión especializada ocasional por la verdad...		0.834427	1.032439
3	asamblea nacional no derogará pensiones vitali...		1.967902	1.701282
4	ecuador asamblea inhabilita exministra de sal...		3.656081	3.269294
...	...	...	...	...
95	atención el presidente lasso no asistirá la c...		1.222207	1.318003
96	asambleísta nacional presidente de la comisió...		1.087695	1.320226
97	en cañar se encuentra un espacio de la asamble...		1.481575	1.243359
98	virgilio saquicela presidente de la asamblea n...		1.102954	1.218984
99	no existen legisladores que hayan demostrado ...		0.804966	1.022198

Fig. 12 Agresividad y odio por tema y localidad

Como se mencionó con anterioridad, para nuestro trabajo investigativo se procedió al desarrollo de una interfaz web como complemento visual de los resultados del modelo para la detección de agresividad y odio utilizando Pysentimiento. En la figura 13 se muestra la página inicial de nuestra interfaz.

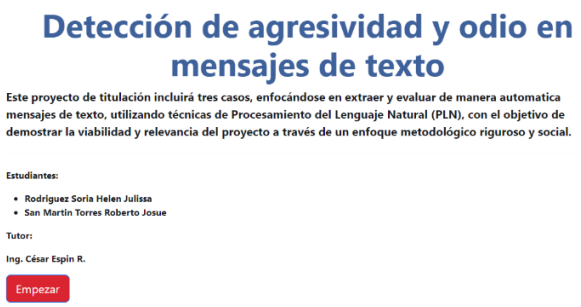

Fig. 13 Interfaz de presentación del modelo.

A continuación, en la figura 14 se muestra los tres casos experimentales en los que el modelo identifica y clasifica la detección de odio y agresividad en mensajes de texto.

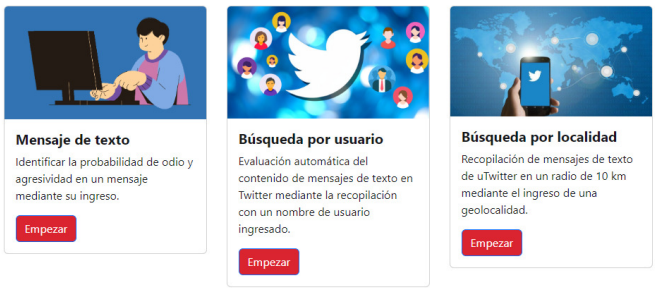

Fig. 14 Interfaz de casos experimentales

Nuestra primera experimentación permite el ingreso manual de un mensaje de texto en español. El modelo analiza la cantidad de agresividad y odio presentes en el mismo y muestra los resultados en una tabla que incluye los porcentajes correspondientes. Además, se muestra el tipo de salida del modelo, lo que proporciona información adicional sobre el análisis realizado.

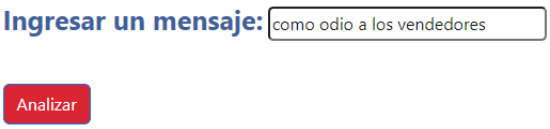

Fig. 15 Mensaje de texto ingresado manualmente.

En nuestro segundo caso, se podrá ingresar un usuario de la red social Twitter, así como la cantidad de mensajes que se quieran analizar del mismo, el modelo identificará la cantidad de mensajes que contengan agresividad y odio representándolos en un gráfico de barras, así como los dos tweets con mayor cantidad de odio y los dos tweets más agresivos.

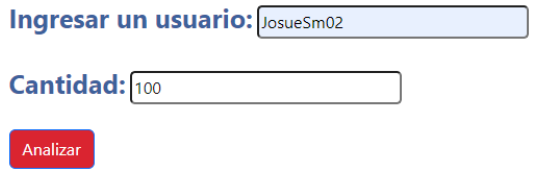

Fig. 16 Evaluación de mensajes de texto de un usuario de Twitter.

Para la búsqueda por localidad, se selecciona la ciudad y, si se desea ingresar, un tema para la búsqueda. De manera similar al segundo caso, el modelo mostrará los dos Tweets con mayor porcentaje de agresividad y odio, así como la cantidad de mensajes identificados como agresividad y odio en un gráfico de barras.

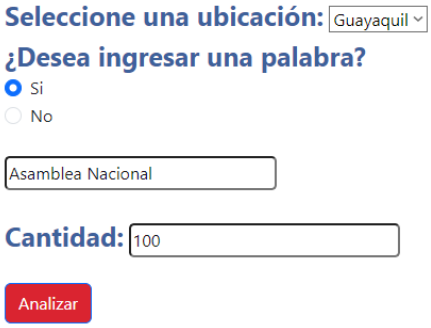

Fig. 17 Evaluación de mensajes de texto por localidad.

III. Resultados

Los resultados obtenidos en el presente trabajo de investigación utilizando el modelo Transformer Pysentimiento son los siguientes:

En el primer caso experimental, se ingresaron diferentes mensajes con la finalidad de obtener resultados variados en la salida del modelo.

En la Figura 18, para el mensaje de ejemplo, se obtuvo que el resultado del modelo es de una tendencia hacia el odio, ya que se observa que el porcentaje de probabilidad de odio es significativamente mayor al porcentaje de agresividad.

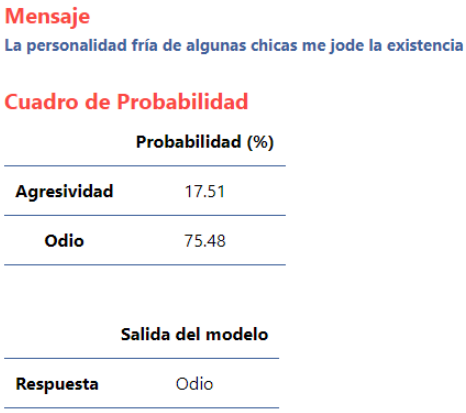

Fig. 18 Identificación de odio.

En algunas ocasiones, el modelo puede proporcionar dos etiquetas de sentimiento como se muestra en la Figura 19. Esto se debe a que ambos porcentajes de probabilidad asociados a cada etiqueta son considerados altos, lo que sugiere que el modelo ha detectado al mismo tiempo síntomas de odio y agresividad en el mensaje analizado.

Mensaje

pilas anda cobrarle esos sabidos prepotentes

Cuadro de Probabilidad

	Probabilidad (%)
Agresividad	64.7
Odio	89.2

	Salida del modelo
Respuesta	Odio y Agresividad

Fig. 19 Identificación de odio y agresividad

Por otro lado, se observa que en la figura 20 el modelo no clasifica al mensaje como odio ni agresividad, e indica porcentajes bajos en cada etiqueta sin mostrar una respuesta específica.

Mensaje

bendito sean los angeles

Cuadro de Probabilidad

	Probabilidad (%)
Agresividad	1.96
Odio	5.28

	Salida del modelo
Respuesta	

Fig. 20 No identifica ni odio ni agresividad.

Para nuestro segundo caso experimental, se buscaron usuarios de Twitter que tuvieran al menos 100 publicaciones en su perfil.

En el primer ejemplo tomamos el usuario @abdalabucaram y mediante el modelo obtuvimos los dos mensajes de texto con mayor porcentaje de odio y los dos mensajes con mayor porcentaje de agresividad como se aprecian en la figura 21.

Mensajes con mayor % de Odio

	Mensaje	Odio	Agresividad
83	hemos respaldado en la brutal acción del gobierno contra ella pero nos preguntamos porq no está presa la prima la romo porq no está presa la romo la ratera de los hospitales porq no está enjuiciado el cobarde traidor de moreno la bruja de su esposa	91.38	83.86
93	otra locura de las mafiosas feministas que perjudican la mujer	88.02	27.60

Mensajes con mayor % de Agresividad

	Mensaje	Odio	Agresividad
83	hemos respaldado en la brutal acción del gobierno contra ella pero nos preguntamos porq no está presa la prima la romo porq no está presa la romo la ratera de los hospitales porq no está enjuiciado el cobarde traidor de moreno la bruja de su esposa	91.38	83.86
6	atención ecuador se busca esta banda de mafiosos que daban órdenes para perseguir bucaram su familia estamos reuniendo dinero para premiar la persona que lo saque de paraguay lo traiga ecuador este hdlgp en forma especial la bruja de su mujer que se creía diosa	73.98	62.95

Fig. 21 Mensajes de odio y agresividad usuario @abdalabucaram

En la figura 22 se observa gráficamente la cantidad de mensajes que fueron identificados como agresividad y odio en las publicaciones de dicho usuario, el cual identificó un total de 7 mensajes de odio y 2 de agresividad.

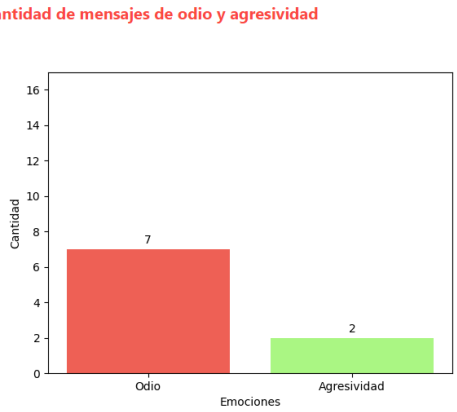

Fig. 22 Cantidad de mensajes de odio y agresividad usuario @abdalabucaram

Para nuestro siguiente ejemplo usamos al usuario @Alebamaaa. La Figura 23 muestra los dos mensajes con los mayores porcentajes de odio y los dos mensajes con mayor porcentaje de agresividad.

Mensajes con mayor % de Odio

	Mensaje	Odio	Agresividad
66	me dicen extranjera como si fuera un insulto cuando son los primeros por quererse salir del país	96.71	74.10
2	que vrg pero las que están encuera no dicen nada	93.74	58.69

Mensajes con mayor % de Agresividad

	Mensaje	Odio	Agresividad
66	me dicen extranjera como si fuera un insulto cuando son los primeros por quererse salir del país	96.71	74.10
2	que vrg pero las que están encuera no dicen nada	93.74	58.69

Fig. 23 Mensajes de odio y agresividad usuario @Alebamaaa

Además, el modelo clasificó 7 mensajes como odio y 3 como agresividad del usuario @Alebamaaa, tal y como se puede observar en la figura 24.

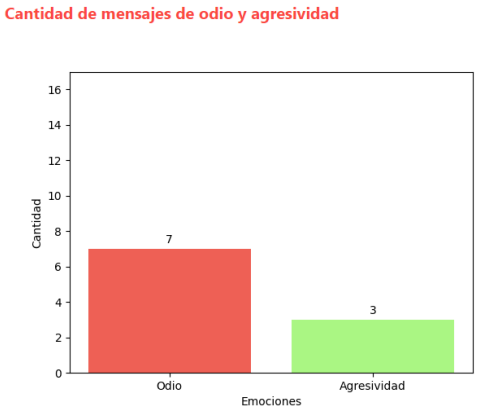

Fig. 24 Cantidad de mensajes de odio y agresividad usuario @Alebamaaa

Para nuestro tercer caso, se realizó una búsqueda por localidad con el tema "delincuencia" que se encontraba en tendencia en Ecuador al momento de la extracción.

Comenzando con la localidad de Guayaquil, en la figura 25 se presentan los dos mensajes con mayor probabilidad de odio, con un 94.94% y 87.23%, y los dos que presentan mayor probabilidad de agresividad, con porcentajes de 78.14% y 53.67%.

Mensajes con mayor % de Odio

	Mensaje	Odio	Agresividad
20	más cansados estamos la mayoría de ecuatorianos de soportar tanto ilegal venezolano que terminará en la delincuencia cuándo se los envían maduro	94.94	78.14
42	sra viteri tome sus medidas de seguridad para guayaquil sin miedo prohíba definitivamente dos personas en moto saque de las calles tanto indigente extranjero que sirven de campaneros para la delincuencia utilizan menores de edad	87.23	52.44

Mensajes con mayor % de Agresividad

	Mensaje	Odio	Agresividad
20	más cansados estamos la mayoría de ecuatorianos de soportar tanto ilegal venezolano que terminará en la delincuencia cuándo se los envían maduro	94.94	78.14
49	aspiraciones de poder se han tomado la delincuencia organizada ecuatoriana mundial para llegar gobernar del brazos de sus socios robolucionarios asesinos narcoterroristas bulgares delincuentes	79.24	53.67

Fig. 25 Odio y agresividad Guayaquil tema “delincuencia”

Como ilustra la figura 26, el modelo Transformer identificó 8 mensajes con contenido de odio y 3 mensajes con contenido de agresividad para la localidad Guayaquil.

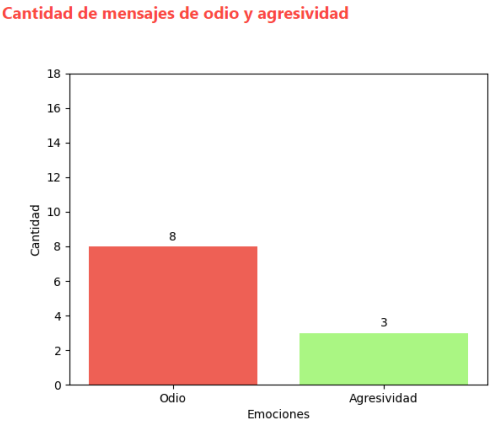

Fig. 26 Cantidad de odio y agresividad Guayaquil tema “delincuencia”

En la figura 27, se puede apreciar cómo Pysentimiento detectó que los dos mensajes con mayor porcentaje de odio presentan un 96.81% y 88.93%, mientras que en los mensajes de agresividad se identificaron valores de 93.29% y 50.22%, esto para la Quito.

Mensajes con mayor % de Odio

	Mensaje	Odio	Agresividad
51	eres una alcahueta de la delincuencia	96.81	93.29
13	no hay control de ingreso de extranjeros muchos delincuentes sicarios por tanto la delincuencia narcoterrorismo deben ser eliminados en su propia ley solo así tendremos un mejor país estoy seguro	88.93	50.22

Mensajes con mayor % de Agresividad

	Mensaje	Odio	Agresividad
51	eres una alcahueta de la delincuencia	96.81	93.29
13	no hay control de ingreso de extranjeros muchos delincuentes sicarios por tanto la delincuencia narcoterrorismo deben ser eliminados en su propia ley solo así tendremos un mejor país estoy seguro	88.93	50.22

Fig. 27 Odio y agresividad Quito tema “delincuencia”

En figura 28, el modelo detectó 12 mensajes de odio y 2 de agresividad para esta localidad.

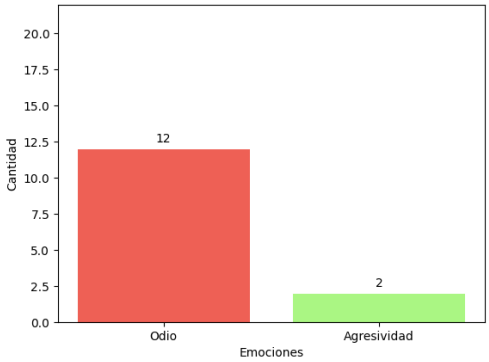

Fig. 28 Cantidad de odio y agresividad Quito tema “delincuencia”

Para Cuenca, en figura 29 podemos visualizar dos primeros mensajes con mayor probabilidad de odio con un 93.37% y 84.39% y dos que tienden a mostrar mayor probabilidad de agresividad con 60.89% y 53.46%.

Mensajes con mayor % de Odio

	Mensaje	Odio	Agresividad
2	chapas maricones cuidando quienes vienen destruir nuestro país mientras la delincuencia nos mata roba	93.37	60.89
18	desgraciadamente esmeraldas ha sido tomada por narcos delincuentes volvieron infierno el paraíso a las autoridades no les intwresa quiera dios algún día tengamos un presiente capaz de parar tanta delincuencia para regresar en paz nuestras hermosas playas	84.39	53.46

Mensajes con mayor % de Agresividad

	Mensaje	Odio	Agresividad
2	chapas maricones cuidando quienes vienen destruir nuestro país mientras la delincuencia nos mata roba	93.37	60.89
18	desgraciadamente esmeraldas ha sido tomada por narcos delincuentes volvieron infierno el paraíso a las autoridades no les intwresa quiera dios algún día tengamos un presiente capaz de parar tanta delincuencia para regresar en paz nuestras hermosas playas	84.39	53.46

Fig. 29 Odio y agresividad Cuenca tema "delincuencia"

Para finalizar, en figura 30 se observa que la emoción de odio es la que predominó, el modelo identificó 8 mensajes clasificados como odio y 2 de agresividad.

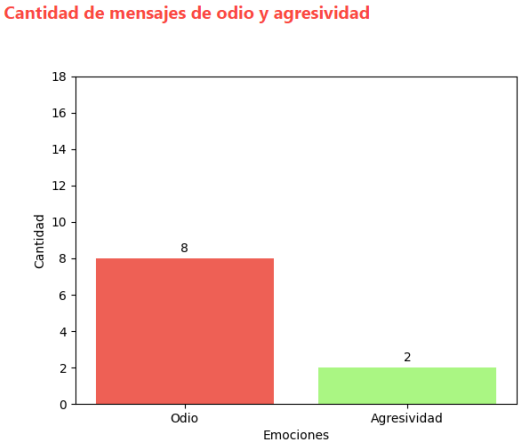

Fig. 30 Cantidad de odio y agresividad Cuenca tema "delincuencia"

IV. DISCUSIÓN

Para el proceso de extracción de Tweets, es importante tener en cuenta que existe un límite de 3000 tweets por intervalo de 15 minutos. Si se excede este límite, la API entrará en modo sleep y será necesario esperar un tiempo antes de poder extraer más información.

Por otro lado, en la extracción de geolocalización la cantidad de Tweets que se pueden extraer se encuentra limitada a un máximo de 100 por búsqueda. Existen varias alternativas para la extracción de tweets por geolocalización, por ejemplo, Twint que permite recolectar tweets cercanos a una ubicación específica. Por otro lado, la librería Geopy de Python en conjunto con Nominatim también son una opción interesante para buscar tweets por coordenadas o por proximidad a una localidad determinada.

En la etapa de procesamiento de datos, se identificaron mensajes de texto que no contienen información relevante para el análisis de sentimiento debido a que sólo incluyen elementos como URL, emoticones o hashtags. Estos mensajes son descartados en el proceso de análisis, ya que no aportan datos significativos para la identificación de contenido agresivo u odioso en los mensajes de texto.

Podemos comparar en Tabla 1 los resultados obtenidos en el tercer caso experimental para las tres localidades con el tema "delincuencia".

TABLA I
CANTIDAD DE MENSAJES DETECTADOS

Localidad	Odio	Agresividad
Guayaquil	8	3
Quito	12	2
Cuenca	8	2

Se aprecia que Quito presenta una mayor tendencia a expresiones de odio, mientras que la ciudad de Guayaquil muestra una mayor tendencia a expresiones de agresividad. Así mismo, tanto Guayaquil como Cuenca presentaron resultados similares en cuanto a la cantidad de mensajes detectados como odio, mientras que Quito y Cuenca obtuvieron resultados similares en cuanto a la cantidad de mensajes detectados como agresivos.

Cuando los resultados del modelo utilizado presentan porcentajes de agresividad y odio muy bajos, no son etiquetados como se muestra en la figura 20. Sin embargo, en todos los casos evaluados, se encontró que el modelo demostró una mayor precisión en la detección del sentimiento de odio, lo cual se evidenció en los valores más altos obtenidos en la clasificación de dicho sentimiento. Además, las limitaciones que tiene el modelo pueden ser atribuidos por factores como la falta de reconocimiento de ciertas palabras y que la precisión de Pysentimiento para identificar odio y agresividad es de 76.05% en la escala F1 y una desviación estándar de 0.5285. Por lo tanto, es importante seguir mejorando y perfeccionando el modelo para lograr una detección más precisa y equilibrada de ambos tipos de contenidos negativos en los mensajes de texto.

Es fundamental destacar que el enfoque principal de nuestro trabajo de investigación es la detección de agresividad y odio en mensajes de texto, y no el estudio poblacional de una ciudad ni la atribución a un usuario en particular. Los usuarios de Twitter seleccionados tienen perfil público y, al igual que las localidades, se utilizaron como datos complementarios para realizar los análisis correspondientes y probar las bondades del modelo.

V. CONCLUSIÓN

El estudio llevado a cabo mediante el uso del modelo Transformer pre-entrenado Pysentimiento demuestra que es

posible determinar síntomas de agresividad y odio analizando mensajes de textos en español publicados por un usuario en redes sociales. Los resultados obtenidos son muy prometedores, en nuestras pruebas el modelo fue capaz de detectar con mayor sensibilidad las emociones de odio en comparación con las de agresividad. No obstante, el modelo Pysentimiento aún presenta algunas limitaciones en la detección correcta de ambos sentimientos negativos mencionadas por sus autores con una precisión del 76.05%. Otro factor para considerar son las diferencias entre los distintos léxicos del idioma español según el país o región de uso, posiblemente Pysentimiento no analice correctamente diversas maneras de expresarse en nuestro idioma. Por lo tanto, es importante seguir entrenando modelos con datasets más diversos en cuanto las distintas variantes léxicas del idioma español. También, con nuestra interfaz web demostramos que es posible crear sistemas o aplicaciones basados en modelos Transformer que detecten síntomas de agresividad y odio en usuarios analizando sus mensajes de textos publicados dando una amplia aplicabilidad comercial y social sobre el tema.

VI. REFERENCIAS

[1] A. Arghittu, A. Aleksandric, H. I. Anderson, S. Melcher, S. Nilizadeh, and G. M. Wilson, "Spanish Facebook Posts as an Indicator of COVID-19 Vaccine Hesitancy in Texas," 2022, doi: 10.3390/vaccines10101713.

[2] "pysentimiento/robertuito-sentiment-analysis · Hugging Face." https://huggingface.co/pysentimiento/robertuito-sentiment-analysis (accessed Feb. 21, 2023).

[3] G. Ortiz and A. H. Gómez, "Detection of Aggressive Tweets in Mexican Spanish Using Multiple Features with Parameter Optimization," 2019.

[4] G. M. Molina, D. M. Plaza, V. T. Martín, and L. A. Ureña, "Ensemble Learning to Detect Aggressiveness in Mexican Spanish Tweets," 2019, Accessed: Jan. 11, 2023. [Online]. Available: http://www.ujaen.es

[5] L. E. Argota Vega, Magana. J. Reyes, Adorno. H. Gomez, and Enguix. G. Bel, "MineriaUNAM at SemEval-2019 Task 5: Detecting Hate Speech in Twitter using Multiple Features in a Combinatorial Framework," pp. 447–452, 2019.

[6] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies," *ACM Transactions on Internet Technology (TOIT)*, vol. 20, no. 2, Mar. 2020, doi: 10.1145/3369869.

[7] Y. Asiri, H. T. Halawani, H. M. Alghamdi, S. H. Abdalaha Hamza, S. Abdel-Khalek, and R. F. Mansour, "Enhanced Seagull Optimization with Natural Language Processing Based Hate Speech Detection and Classification," *Applied Sciences (Switzerland)*, vol. 12, no. 16, Aug. 2022, doi: 10.3390/app12168000.

[8] P. Röttger, H. Seelawi, D. Nozza, Z. Talat, and B. Vidgen, "MULTILINGUAL HATECHECK: Functional Tests for Multilingual Hate Speech Detection Models".

[9] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Multi-modal aggression identification using Convolutional Neural Network and Binary Particle Swarm Optimization," *Future Generation Computer Systems*, vol. 118, pp. 187–197, May 2021, doi: 10.1016/J.FUTURE.2021.01.014.

[10] A. Datta, S. Si, U. Chakraborty, and S. Kumar Naskar, "Spyder: Aggression Detection on Multilingual Tweets," pp. 11–16, 2020, Accessed: Jan. 26, 2023. [Online]. Available: https://www.smartinsights.com/social-media-

[11] C. Paul and P. Bora, "Detecting Hate Speech using Deep Learning Techniques," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, p. 2021, Accessed: Jan. 26, 2023. [Online]. Available: www.ijacsa.thesai.org

[12] V. G. Mladen and K. Jaň, "Combining Shallow and Deep Learning for Aggressive Text Detection," p. 188, 2018.

[13] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media," 2019, Accessed: Jan. 26, 2023. [Online]. Available: http://bit.ly/2FhLMVz

[14] S. G. L. de la Peña and P. Rosso, "Aggressive Analysis in Twitter using a Combination of Models," 2019.

[15] R. Alshalan and H. Al-Khalifa, "A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere," *Applied Sciences 2020, Vol. 10, Page 8614*, vol. 10, no. 23, p. 8614, Dec. 2020, doi: 10.3390/APP10238614.

[16] J. Chen, S. Yan, and K. C. Wong, "Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis," *Neural Comput Appl*, vol. 32, no. 15, pp. 10809–10818, Aug. 2020, doi: 10.1007/S00521-018-3442-0/METRICS.

[17] S. N. Safi, P. Patwa, S. Pykl, P. Mukherjee, A. Das, and T. Solorio, "Aggression and Misogyny Detection using BERT: A Multi-Task Approach," pp. 11–16, 2020, Accessed: Jan. 11, 2023. [Online]. Available: https://www.theverge.com/interface/2019/

[18] C. Espin-Riofrio, J. Ortiz-Zambrano, and A. Montejo-Ráez, "SINAI at PoliticEs 2022: Exploring Relative Frequency of Words in Stylometrics for Profile Discovery," *CEUR Workshop Proc*, vol. 3202, 2022.

[19] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019, Accessed: Jan. 11, 2023. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[20] J. A. Ortiz-Zambrano, C. Espin-Riofrio, and A. Montejo-Ráez, "Combining Transformer Embeddings with Linguistic Features for Complex Word Identification," *Electronics (Switzerland)*, vol. 12, no. 1, pp. 1–10, 2023, doi: 10.3390/electronics12010120.

[21] A. Lambebo Tonja, M. Arif, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Detection of Aggressive and Violent Incidents from Social Media in Spanish using Pre-trained Language Model," 2022, Accessed: Jan. 22, 2023. [Online]. Available: http://ceur-ws.org

[22] J. Cã Nete, S. Donoso, F. Bravo-Marquez, A. Carvallo, and V. Araujo, "ALBETO and DistilBETO: Lightweight Spanish Language Models", Accessed: Feb. 20, 2023. [Online]. Available: https://github.com/ckiplab/

[23] M. Guzman-Silverio, Á. Balderas-Paredes, and A. Pastor López-Monroy, "Transformers and Data Augmentation for Aggressiveness Detection in Mexican Spanish," 2020, Accessed: Jan. 22, 2023. [Online]. Available: https://www.cimat.mx/es/adri

[24] S. E. Romero, R. Kleinlein, C. Luna-Jiménez, and J. M. Montero, "GTH-UPM at DETOXIS-IberLEF 2021: Automatic Detection of Toxic Comments in Social Networks".

[25] V. Gómez-Espinosa, V. Muñiz-Sanchez, and A. Pastor López-Monroy, "Transformers Pipeline for Offensiveness Detection in Mexican Spanish Social Media," 2021, Accessed: Jan. 22, 2023. [Online]. Available: https://github.com/keredson/wordninja