**ISSN 2529-9824**

Artículo de Investigación

# Cómo entrenar a un algoritmo para detectar el lenguaje de odio en el conflicto Israel-Palestina

## How to train an algorithm to detect hate speech in the Israel-Palestine conflict

**Antonio Rico Sulayes:** Universidad de las Américas Puebla, México.
antonio.rico@udlap.mx

**Fecha de Recepción:** 24/12/2024

**Fecha de Aceptación:** 25/01/2025

**Fecha de Publicación:** 30/01/2025

## Cómo citar el artículo

Rico Sulayes, A. (2025). Cómo entrenar a un algoritmo para detectar el lenguaje de odio en el conflicto Israel-Palestina [How to train an algorithm to detect hate speech in the Israel-Palestine conflict]. *European Public & Social Innovation Review*, 10, 01-16. https://doi.org/10.31637/epsir-2025-1199

**Resumen**

**Introducción:** La detección automática de comportamientos nocivos en la comunicación digital, como el lenguaje de odio hacia algún grupo social, es abordada por el análisis de sentimientos, tarea del procesamiento del lenguaje natural. Detectar el lenguaje de odio en el conflicto Israel-Palestina es particularmente complejo, ya que se puede observar un discurso de odio, y otro de apoyo, hacia cada uno de los dos grupos involucrados. **Metodología:** Como un algoritmo necesita aprender a reconocer este comportamiento a partir de ejemplos, aquí se ensambló un corpus con comentarios extraídos de redes sociales y relacionados con el conflicto armado. Además, se crearon las reglas para un etiquetado cuádruple, propio del discurso ideológico, donde un grupo dice tener la verdad frente a otro, cuyas creencias califica de ideología, mientras el otro grupo defiende lo contrario **Resultados:** Este trabajo muestra el nivel de validez alcanzado en el acuerdo entre etiquetadores respecto de varios niveles de polaridad, incluyendo apoyo, odio y neutral. **Conclusiones:** Si bien se alcanzó una validez sustancial con dos niveles de polaridad opuestos, la inclusión de un nivel de neutralidad aumenta la complejidad y reduce el coeficiente de validez. Al final, se discuten las posibles aplicaciones del corpus en términos de seguridad e inteligencia.

**Palabras clave:** discurso ideológico; lenguaje de odio; corpus lingüísticos; comunicación digital; acuerdo entre etiquetadores; análisis de sentimientos; lingüística computacional; conflicto Israel-Palestina.

**Abstract**
**Introduction:** Automatic detection of toxic behavior in digital communication, such as hate speech targeting social groups, is explored in sentiment analysis, a task of natural language processing. Detecting hate speech in the Israel-Palestine conflict is particularly complex because there are both, a hate speech towards and a support speech for each of the two groups involved. **Methodology:** Since an algorithm needs to learn to recognize this behavior using examples, a corpus was created using comments related to the conflict in social media. A set of rules were also designed for a four-way tagging process. Ideological discourse is characterized by a group that argues they have the truth in opposition to another group whose beliefs they see as an ideology, while the opposite view is held by the other group. **Results:** This article shows the validity level achieved in the interrater agreement with various polarity levels, including support, hate and neutral. **Conclusions:** A high level of validity is reached with two opposite polarity levels, but including a level of neutrality increases the complexity and reduces the validity coefficient. At the end, potential applications of the created corpus in the context of security and intelligence are discussed.

**Keywords:** ideological discourse; hate speech; corpora; digital communication; interrater agreement; sentiment analysis; computational linguistics; Israel-Palestine conflict.

## 1. Introducción

El conflicto Israel-Palestina tiene aspectos sociales, políticos y económicos, así como implicaciones culturales, morales y religiosas, que lo convierten en un fenómeno estrechamente complejo para su simple descripción y definición. Lo delicado del asunto aumenta cuando pretendemos estudiar el fenómeno y argumentar sobre alguna propuesta de ayuda o solución.

Reconociendo esta complejidad, este artículo se centra únicamente en un aspecto que es característico de este tipo de conflicto en general, el discurso ideológico que se forma en torno al problema. En particular, este estudio explora la posibilidad de contar con los recursos necesarios para detectar de manera automática las varias formas que puede tomar este discurso entre quienes apoyan o reprueban a los grupos sociales involucrados en el conflicto.

Para los estudiosos del lenguaje (van Dijk, 2011), algunos contextos sociales como los conflictos políticos, por ejemplo, se van a caracterizar por un discurso donde dos grupos que se enfrentan van a hacer algo muy similar. Cada grupo presenta a los suyos como buenos y a los del grupo opuesto como malos. Esta última acción, comúnmente genera un discurso que busca menoscabar, humillar o incluso agredir al grupo contrario, convirtiéndose en lo que se suele llamar un lenguaje de odio (Gagliardone *et al*., 2015).

La aparición de un lenguaje de odio puede ser difícil de reconocer para los humanos mismos porque, aunque muchas veces el discurso de odio se puede manifestar como un discurso explícitamente contestario y abiertamente violento, muchas otras veces utiliza recursos de negación o soslayo, para no mostrar su aversión por los otros de manera tan rotunda. Las complejas manifestaciones que puede tener el lenguaje de odio, a mi leal saber y entender, no han sido abordadas por los estudiosos de la detección automática de comportamientos nocivos en la comunicación.

El procesamiento del lenguaje natural ha sido una de las áreas, de entre las varias que estudian al lenguaje, que se ha interesado particularmente por el lenguaje de odio (Davidson *et al*., 2017; Fortuna y Nunes, 2019; Poletto *et al*., 2021; Waseem, 2016). La detección automática de comportamientos nocivos en la comunicación digital es uno de los objetivos principales del análisis de sentimientos y de la minería de textos.

Para realizar una detección automática de un comportamiento nocivo en la comunicación digital, se pueden o bien usar listas de palabras o expresiones previamente identificadas como características de este tipo de comportamiento (Lingiardi *et al*., 2020), o bien enseñar o entrenar a un algoritmo de aprendizaje automático a detectar este comportamiento (Nobata *et al*., 2016). Para este último tipo de acercamiento al problema, un algoritmo de inteligencia artificial necesita aprender qué constituye dicho comportamiento a partir de ejemplos del mismo en la comunicación espontánea.

Alineándose con el segundo método, el presente trabajo tiene como objetivo mostrar las dificultades y los logros al crear un corpus con ejemplos del lenguaje de odio que caracteriza al conflicto Israel-Palestina. Como se argumenta aquí, las dificultades se derivan principalmente de la complejidad discursiva que es propia de este tipo de problema social.

En esta investigación, se extrajeron comentarios de redes sociales que tenían que ver con el conflicto armado. Después se identificó a algunos individuos con presencia pública que producían o promovían un discurso de odio hacia alguno de los dos grupos involucrados y se extrajeron automáticamente comentarios de sus conversaciones para generar un corpus.

Después de eliminar los comentarios fuera de tema, se etiquetaron los comentarios restantes haciendo una distinción entre el discurso de apoyo y de odio hacia los dos grupos en conflicto. También se identificaron comentarios que describen el conflicto desde un punto de vista neutro. En el proceso de etiquetación se generó un conjunto de reglas y pautas para ayudar a los etiquetadores a identificar los tres niveles de polaridad: apoyo, odio y neutro. Después de realizar una segunda etiquetación, se compararon las etiquetas para ver si los humanos podían alcanzar un nivel de acuerdo significativo que validara los datos.

En los resultados de este trabajo, se discuten los logros en términos de la validación de las etiquetas por medio de una comparación del trabajo de diferentes etiquetadores. Con las dos etiquetas que representan los niveles de polaridad extremos, *odio* y *apoyo*, se alcanzó un nivel de acuerdo sustancial (cercano a un nivel casi perfecto). Además, con las etiquetas que identifican al grupo meta del comentario, *Israel* y *Palestina*, también se alcanzó un nivel de acuerdo sustancial.

Dos aspectos se discuten a manera de conclusión. Por un lado, no todo el discurso que aparece en las redes sociales toma un sesgo de odio o apoyo hacia alguno de los dos grupos involucrados en el conflicto. También existe un discurso neutro, comúnmente periodístico, que habría que poder identificar y separar del discurso en el que sí se asume una postura.

Con un corpus aumentado con la categoría *neutro*, también se ha logrado un acuerdo sustancial entre los etiquetadores, pero el coeficiente de kappa de Cohen, que mide el acuerdo entre los etiquetadores, disminuye al añadir esta tercera etiqueta. Esto muestra la dificultad de diferenciar esta neutralidad incluso entre los humanos mismos. Por otro lado, al final se discuten brevemente las posibles aplicaciones de nuestro corpus en términos de seguridad e inteligencia.

## 2. Metodología

Entrenar a una máquina para que determine si un texto expresa odio o apoyo es, en su conceptualización más básica, una tarea de clasificación de texto. Es decir, la máquina debe decidir si el texto pertenece a una categoría de textos que expresan odio o a la categoría opuesta, de textos que expresan apoyo. La clasificación automática de textos es parte de la llamada minería de textos o recuperación de información, y ésta es parte del procesamiento del lenguaje natural (Manning, Raghavan y Schütze, 2008).

Cuando las categorías en que se clasifica un texto representan un sentimiento (de manera muy genérica una emoción expresada en un binomio positivo-negativo), se suele hablar de análisis de sentimientos (Jurafsky y Martin, 2023). Para poder realizar una clasificación de textos, existen dos posibles procedimientos, llamados supervisado y no supervisado, respectivamente (Manning, Raghavan y Schütze, 2008). El primero es el método más común, y se le llama supervisado porque lo que usa la máquina como punto de partida es un corpus de entrenamiento en el cual los humanos han decidido previamente qué ejemplos pertenecen a una etiqueta y cuáles a otra, al menos en un contexto dicotómico como el antes descrito.

Ahora bien, poco a poco los estudios de análisis de sentimientos se han ido haciendo más complejos en sus clasificaciones, y para ello han construido corpus que rebasan el binomio positivo-negativo. De esta manera, se va reconociendo la complejidad de algunos de los fenómenos lingüísticos que están bajo estudio. Por ejemplo, Nobata *et al*. (2016), que tienen como propósito general detectar el lenguaje abusivo (lo que aquí hemos llamado un comportamiento lingüístico nocivo), solicitaron a sus anotadores humanos decidir si un comentario era *limpio* o *abusivo*, y si resultaba ser el segundo, tenían que dar una subclasificación en *lenguaje de odio*, *derogatorio* o *profanidad*.

Usando el corpus de una competencia, de Paula y Schlicht (2021) desarrollan un sistema para detectar leguaje nocivo, de tipo xenofóbico, y si el lenguaje es nocivo intentan calificarlo como *ligeramente nocivo*, *nocivo* o *muy nocivo*. Ejemplos de este tipo de subclasificaciones o categorizaciones más granulares hay muchos. Lo que no hay, a mi leal saber y entender, es un trabajo que intente operacionalizar la conceptualización del lenguaje de odio en el análisis del discurso a través de una categorización que distinga la complejidad de este fenómeno desde la perspectiva de esta área de la lingüística aplicada.

### *2.1. Los conflictos políticos y el discurso ideológico*

Para los analistas del discurso, el estudio del discurso político los llevó a observar una configuración discursiva peculiar en los conflictos entre grupos sociales. Esta configuración se deriva del concepto de ideología. El concepto de ideología es particularmente interesante desde el punto de vista lingüístico porque su connotación negativa no es fija, sino que depende de la perspectiva del enunciador. La ideología no es más que una etiqueta derogatoria utilizada por los hablantes para designar la verdad de los otros, cuya verdad es opuesta a la propia (van Dick, 2011).

Este concepto es particularmente socorrido cuando se está hablando de asuntos importantes como, por ejemplo, la constitución de las naciones o etnias, su derecho a los recursos (como el territorio), o la justificación de un ataque o agresión al otro. Un aspecto más de la complejidad del discurso ideológico es que, comúnmente, cada uno de los grupos opuestos en un *confrontamiento* social utiliza el mismo recurso de llamar ideología a la verdad de sus contrarios, reservando el término de verdad para la suya exclusivamente.

De esta manera, en una configuración simple donde se oponen dos grupos, el discurso ideológico está guiando por estrategias discursivas que buscan uno de cuatro propósitos. Cuando un grupo se enfrenta a otro, su discurso intenta

1) enfatizar sus cosas buenas,

2) desenfatizar las malas, y al mismo tiempo,

3) enfatizar lo malo en el otro grupo y

4) desenfatizar lo bueno que tiene (Jiwani y Richardson, 2011; van Dick, 2011).

Estos cuatro objetivos opuestos constituyen lo que van Dick (2011) llama un cuadrado ideológico. Llevado al máximo de detalle en un conflicto de dos grupos, como el aquí estudiado, un cuadrado ideológico produce ocho posibles acciones discursivas, las cuales se muestran en la Figura 1, a continuación.

**Figura 1.**

*Propósitos de las estrategias discursivas ideológicas en el conflicto Israel-Palestina*

| **Cuadrado Ideológico en el conflicto Israel-Palestina** | |
|---|---|
| Mi grupo | El grupo contrario |
| **Enfatizar lo bueno**<br>Israel enfatiza lo bueno de Israel<br>Palestina enfatiza lo bueno de Palestina. | **Desenfatizar lo bueno**<br>Israel desenfatiza lo bueno de Palestina<br>Palestina desenfatiza lo bueno de Israel. |
| **Desenfatizar lo malo**<br>Israel desenfatiza lo malo de Israel<br>Palestina desenfatiza lo malo de Palestina. | **Enfatizar lo malo**<br>Israel enfatiza lo malo de Palestina<br>Palestina enfatiza lo malo de Israel. |

**Fuente:** Elaboración propia (2024).

Si fuéramos exhaustivos, la etiquetación de un corpus de comentarios sobre el conflicto Israel-Palestina debería incluir al menos ocho etiquetas, una por cada de las dos acciones que contiene cada vértice del cuadrado ideológico presentado en la Figura 1. Aquí mostraremos una solución más simple para nuestro etiquetado, pero que se inspira en la complejidad del cuadrado ideológico de la Figura 1.

Una simplificación resulta necesaria ya que cada una de las ocho acciones discursivas mostradas en la Figura 1 puede expresarse por medio de estrategias discursivas diversas, que si fueran identificadas individualmente producirían un etiquetado excesivamente complejo. Estas estrategias discursivas, que se presentan a continuación, más bien han servido en esta investigación para establecer las reglas y pautas de etiquetado que se discuten dos subsecciones más adelante.

Los cuatro objetivos de un cuadrado ideológico simple, o de un solo grupo, pueden expresarse en el discurso real de diversas maneras, a veces de manera directa, expresando abiertamente el rechazo de un grupo contrario, a veces de manera indirecta, disfrazando el ataque o realizándolo de manera más velada. En los estudios del análisis del discurso existe al menos un grupo de estrategias para cada una de estas dos formas de enfrentamiento (Jiwani y Richardson, 2011).

Las estrategias de focalización o perspectivización, que incluyen a los nombramientos, serían del primer tipo. Es decir, se trata de estrategias más bien explícitas o abiertamente de rechazo. Por el contrario, las estrategias argumentativas, particularmente aquellas que incluyen una negación de responsabilidades, son un ejemplo de las segundas, en las que se busca ocultar el ataque o la agresión.

Las estrategias de focalización o perspectivización, explícitas en la forma que expresan el odio o rechazo del otro, incluyen cuatro estrategias de nombramiento. Según Jiwani y Richardson (2011), estas cuatro estrategias son:

1) la primitivización ("la animalidad / el salvajismo del atentado…"),

2) la problematización superlativa ("el ataque dejó a la ciudad paralizada / sin servicios…"),

3) la somatización mental negativa ("disparaban como locos…") y

4) la criminalización ("el ejército es un asesino…").

Frente a estas estrategias abiertas en su rechazo, las argumentativas de negación de responsabilidades buscan expresar una postura de manera más indirecta o parcialmente oculta.

Las estrategias de negación de responsabilidades para Jiwani y Richardson (2011) son:

1) la negación aparente ("no estoy en contra de los palestinos, pero ellos tienen la culpa…"),

2) la concesión aparente ("algunos israelitas dicen la verdad, pero los sionistas mienten…"),

3) la empatía aparente ("estoy con los palestinos que han perdido familiares, pero la situación de Israel es peor…"),

4) la ignorancia aparente ("no conozco las cifras reales de muertos, pero los medios siempre exageran en favor de Palestina…"),

5) la inversión ("Israel es la verdadera víctima del pueblo Palestino…") y

6) la transferencia ("yo le creo a Israel, pero los medios dicen que ha mentido…").

La dificultad de observar comentarios como los anteriores es que incluso a un humano le resulta muchas veces difícil decidir cuál de los grupos es el objetivo del comentario y si el sentimiento expresado es de odio o apoyo. Todas las estrategias antes mencionadas están listadas, por categoría, en la Tabla 1 a continuación.

**Tabla 1.**

*Estrategias discursivas características del discurso ideológico*

| Estrategias del discurso ideológico | |
|---|---|
| **Focalización** | **Argumentativas** |
| nombramientos | negación de responsabilidades |
| primitivización | negación aparente |
| problematización superlativa | concesión aparente |
| somatización mental negativa | empatía aparente |
| criminalización | ignorancia aparente |
| | inversión |
| | transferencia |

**Fuente:** Elaboración propia (2024).

Haciéndole explícitas a los etiquetadores las estrategias listadas en la Tabla 1, he intentado hacer más clara la tarea y mejorar el acuerdo entre los mismos. Todo esto tiene como finalidad última obtener el coeficiente de validez más alto posible en los datos. Después de presentar el proceso de obtención del corpus aquí utilizado en la siguiente subsección, la última subsección de la Metodología muestra cómo estas estrategias discursivas ayudaron a guiar el proceso de etiquetación.

### *2.2. Recolección de un corpus sobre el conflicto Israel-Palestina*

Dos son los elementos primordiales de cualquier sistema de clasificación automática de texto, sobre todo si la clasificación es supervisada o basada en juicios humanos previos (Rico-Sulayes, 2018). Por un lado, es necesario contar con un corpus con ejemplos de los diferentes tipos de texto que el sistema debe reconocer.

Por otro lado, se debe seleccionar un algoritmo de clasificación adecuado, comúnmente de carácter estocástico o basado en estadísticas e identificación de patrones, para procesar los datos observados en el corpus y hacer predicciones sobre nuevos textos. Este trabajo se enfoca en el reto más importante implícito en los dos requisitos antes mencionados, la obtención de datos que tengan validez.

Cabe señalar, que en ocasiones el investigador puede utilizar un corpus ensamblado por alguien más para alcanzar una tarea de clasificación automática. Personalmente, he seguido este tipo de formato en otras investigaciones (por ejemplo, Rico-Sulayes, 2020; Rico-Sulayes y Monsalve-Pulido, 2022) y antes hablé de que el estudio en de Paula y Schlicht (2021) justo utiliza un corpus proporcionado en una competencia y desarrollado por terceros.

Sin embargo, como he mencionado antes, no conozco ningún corpus que haya sido construido teniendo en cuenta la complejidad del cuadrado ideológico de los analistas del discurso. En este sentido, el conflicto Israel-Palestina ofrece una oportunidad perfecta para explorar este fenómeno discursivo en la creación de un corpus para una tarea de análisis de sentimientos y esto es lo que he hecho trabajando con un grupo de tesistas y estudiantes.

El primer método de recolección que decidí utilizar con algunas estudiantes con quienes comencé a trabajar este tema fue la recolección de comentarios de la plataforma de YouTube. Concretamente comenzamos utilizando una solución en línea para la descarga automática de comentarios asociados con videos específicos, *YouTube Data Tools* (Rieder, 2015).

También utilizamos una solución programática, que implementé en Python explotando las librerías de Google API Client (2024). Sin embargo, algo que resultó evidente de manera inmediata era que esta plataforma no generaba comentarios con el tipo y nivel de polarización que estábamos buscando.

Entonces, decidimos identificar y seguir a usuarios de alta popularidad en la red social X, antes Twitter. De esta manera, nos dimos cuenta de que este tipo de usuarios tenían la capacidad de promover entre sus seguidores un lenguaje de odio mucho más abierto y agresivo contra el grupo contrario. Este fenómeno no es extraño. De hecho, como se mencionó antes, los estudios del discurso ideológico se basaron en gran medida en el estudio del discurso político.

En el análisis del discurso político, los lingüistas han observado que los actores de la vida política tienden a legitimar a su grupo y deslegitimar a sus oponentes con el tipo de estrategias discursivas antes expuestas (Chilton y Schäffner, 2011). En la comunicación digital de las redes sociales, estos usuarios de gran popularidad se convierten en actores políticos que exhiben y promueven entre sus seguidores este tipo de discurso polarizado, y se convierten en un nicho para la proliferación del lenguaje de odio en los conflictos políticos.

Partiendo de las búsquedas "Israel vs Palestina" y "Guerra Palestina e Israel", se filtraron los resultados en X para identificar un conjunto de usuarios que apoyaban abiertamente a uno de los dos grupos. Además, se añadieron algunas cuentas de medios de información dedicadas a la cobertura del conflicto. Esto último permitió que recuperáramos comentarios de carácter más bien neutro, que es la otra categoría que nos interesaba.

Con esto, se ensambló una lista de 90 usuarios, 30 para cada una de tres categorías, pro-Israel, pro-Palestina y neutral. Utilizando el servicio de descarga automática de comentarios *Export Social Media Comments* se hicieron descargas de comentarios de estas cuentas y se ensambló una primera base de datos con 1.300 comentarios aproximadamente.

Como las descargas automáticas incluyen todos los comentarios asociados con una cuenta, sin ningún tipo de filtro, la base de datos antes mencionada tuvo que procesarse para dejar sólo aquellos comentarios que realmente se relacionaban con nuestro tema.

Utilizando las reglas que se detallan en la siguiente subsección, se limpió la base de datos y al final nos quedamos con 362 comentarios, incluyendo la categoría neutral. Sin esta categoría, 257 comentarios con una postura de odio o apoyo se distribuyen en un cuadrado ideológico simple según lo muestra la Figura 2, a continuación.

**Figura 2.**

*Distribución de los comentarios del corpus en un cuadrado ideológico simple*

| | **Comentarios de un corpus sobre el conflicto Israel-Palestina** | |
|---|---|---|
| pro-Palestina | **Odio hacia Israel**<br>106 comentarios | **Apoyo hacia Palestina**<br>69 comentarios |
| pro-Israel | **Odio hacia Palestina**<br>51 comentarios | **Apoyo a Israel**<br>31 comentarios |

**Fuente:** Elaboración propia (2024).

Aunque la distribución de comentarios entre los cuatro vértices no es perfecta, pues raramente las comunicaciones digitales se corresponden con las categorías que las investigaciones tienen como objetivo, la división más general, pro-Palestina y pro-Israel, muestra que existen abundantes ejemplos para ambas posturas. La primera alcanza un total de 175 comentarios y la segunda, 82 comentarios.

En una clasificación automática existen medidas del rendimiento de un algoritmo que neutralizan los desbalances de los datos en cierta medida. Sin embargo, también es posible eliminar algunos comentarios de la categoría más abundante, *odio hacia Israel*, para que el algoritmo reciba un corpus más balanceado. De una u otra manera, los números alcanzados son un buen punto de partida para poder hacer el entrenamiento de un algoritmo en la tarea que aquí nos ocupa. La Tabla 2 a continuación muestra algunos ejemplos de comentarios que han sido categorizados en cada uno de los cuatro vértices del cuadrado ideológico de la Figura 2.

**Tabla 2.**

*Ejemplos de comentarios categorizados en cada uno de los vértices del cuadrado ideológico*

| Postura | Vértice | Cód | Ejemplo de comentario |
|---|---|---|---|
| pro-Palestina | Odio a Israel | OI | "@usuario @usuario @usuario Los sionistas cometieron 12 masacres durante las últimas 24 horas, matando a 140 palestinos, 40 de ellos niños, e hiriendo a otros 210 civiles! ¡La "prensa" sigue ocultando cada día estas imágenes! https: URL" |
| | Apoyo a Palestina | AP | "@usuario @usuario @usuario No se tú, pero los palestinos soportan por 76 años constantemente la represión brutal, encarcelamiento de hombres, mujeres, ancianos y niños, torturas, asesinatos, destrucción de viviendas, etc." |
| pro-Israel | Odio a Palestina | OP | "@usuario ¡Qué horror! Ahora estarán felices los propalestinos, mostrarán el martir ejemplar de su amor por la tragedía y la muerte violenta en pos de la causa de Mahoma. La locura y la barbarie al extremo hacen fácil presa de individuos con debilidades mentales." |
| | Apoyo a Israel | AI | "@usuario Felicidades al ejército israelí por el buen trabajo exterminado a los terroristas gazaties. Qué no quede un sólo túnel y refugio para los terroristas palestinos. ¡¡No más un 7 de octubre, no más!! Netanyahu has lo qué tengas que hacer para evitarlo." |

**Fuente:** Elaboración propia (2024).

Los ejemplos mostrados en la Tabla 2 conservan la ortografía del comentario original y sólo han sido editados anonimizando las referencias a usuarios o páginas de internet específicas y eliminando los espacios sobrantes. La siguiente subsección explica en detalle las reglas que se utilizaron para llegar a la primera etiquetación. Esta etiquetación utilizó los códigos de los cuatro vértices del cuadrado ideológico, los cuales aparecen en la tercera columna de izquierda a derecha en la Tabla 2, como OI, AO, OP y AI.

### *2.3. Reglas de etiquetado del lenguaje de odio en el conflicto Israel-Palestina*

De los 1.300 comentarios que se recolectaron originalmente, se hizo una limpieza eliminando los comentarios que estaban fuera de tema o que no eran entendibles por su complejidad o escritura no normativa. Las reglas a las que se llegó para esa primera eliminación se muestran en la Tabla 3 a continuación.

**Tabla 3.**

*Reglas de eliminación de comentarios de la descarga automática original*

| Reglas de eliminación |
|---|
| 1. El comentario no se relaciona con el tema |
| 2. El comentario contiene tan sólo una hiperliga |
| 3. El comentario contiene únicamente emoticones |
| 4. No es posible reconocer al grupo meta del comentario |
| 5. El comentario es demasiado confuso |
| 6. El comentario está en otro idioma que no es el español |
| 7. El comentario es ambiguo o no se entiende su postura |

**Fuente:** Elaboración propia (2024).

Cabe señalar que la mayor cantidad de comentarios que se quitaron fueron eliminados por la aplicación de la primera regla de la Tabla 3, es decir, porque estaban fuera de temática. Como se mencionó ya, los programas de descarga automática de comentarios no discriminan el tema. Esto es particularmente cierto en el tipo de descarga que nosotros hicimos, que estaba basada en cuentas de usuarios. En este esquema, todo comentario asociado con una cuenta es descargado, incluso cuando el usuario y sus seguidores cambian de tema. Sin embargo, decidimos hacer este tipo de descarga porque creemos que el efecto de estos usuarios en la alta polarización del discurso que generan o promueven supera el efecto negativo de tener que borrar comentarios.

Después de la aplicación de las reglas en la Tabla 3, el corpus contaba con 362 comentarios. La aplicación de una regla más para identificar comentarios neutrales produjo 105 de estos comentarios. El resto de los 257 comentarios fueron anotados aplicándoles una etiqueta para el grupo social meta en el comentario I (Israel) y P (Palestina) y una etiqueta para categorizar el sentimiento expresado hacia dicho grupo, A (apoyo) y O (odio). Con la combinación de estas dos etiquetas se producen los cuatro códigos para los vértices del cuadrado ideológico, OI (odio a Israel), AP (apoyo a Palestina), OP (odio a Palestina), y AI (apoyo a Israel).

Las estrategias discursivas de la Tabla 1, cuatro abiertamente en contra del grupo opuesto y 6 más bien indirectas, fueron el punto de partida, pero no las quisimos forzar sobre el etiquetado de manera artificial. Al final, el objetivo principal era hacer consciente a los etiquetadores sobre la complejidad de las estrategias discursivas ideológicas, y no generar un etiquetado excesivamente complejo y difícil de aplicar. Los resultados en la validez del corpus y sus categorías, después de realizar dos etiquetados independientes, son presentados en la siguiente sección.

## 3. Resultados

La validez de un estudio tiene que ver con su adecuación o acercamiento a la verdad, es decir, un estudio es válido si está libre de errores (Villasís-Keever *et al*., 2018). Las fuentes de errores en un estudio pueden categorizarse en tres: los sesgos de selección, de medición y de confusión. De estas tres fuentes de errores o sesgos, los problemas de medición afectan particularmente a los datos.

Los datos utilizados en una investigación podrían ser no válidos o verdaderos por una de tres razones: errores del sujeto de investigación, del instrumento para la medición de las variables o del evaluador de las mediciones. Particularmente cuando se tienen datos que una persona del equipo investigador debe evaluar y categorizar, podría haber errores en dicha evaluación. Para evitar que haya una evaluación incorrecta de los datos se debe:

1) establecer lineamientos claros para la evaluación,

2) capacitar a los evaluadores y

3) medir la variabilidad de las evaluaciones por diferentes sujetos.

Las dos primeras acciones, definir nuestros criterios de etiquetación y presentárselos a los etiquetadores, han sido descritas en la sección anterior. La medición del acuerdo entre los etiquetadores es el paso final que se presenta en esta sección para argüir sobre la validez de los datos recolectados.

Cuando se busca medir la variabilidad en una medición cualitativa o categórica, del tipo positivo-negativo, se utiliza un coeficiente Kappa (Villasís-Keever *et al*., 2018). En el contexto del análisis de sentimientos, que es el objetivo último de esta investigación, se suele usar el coeficiente Kappa de Cohen (Spasić y Buerki, 2020). Este coeficiente se calcula utilizando la fórmula siguiente (DATAtab, 2024).

(1)

$$K = \frac{\mathrm{Po} - \mathrm{Pe}}{1 - \mathrm{Pe}}$$

En (1), *Po* representa el acuerdo observado, es decir, la proporción de coincidencias, la cual se obtiene dividiendo el número de coincidencias entre el número total de juicios o etiquetas producidas. Por su parte, *Pe* representa el acuerdo esperado como producto del azar, o la probabilidad hipotética de que las coincidencias sean resultado del azar. Para mostrar el cálculo de estas dos figuras y la obtención de un coeficiente Kappa de Cohen, es mejor utilizar un ejemplo.

Para evaluar el acuerdo de las etiquetadoras que trabajaron en esta investigación respecto del binomio del sentimiento A (apoyo) y O (odio), es necesario graficar sus juicios en una matriz de confusión. La Tabla 4, a continuación, muestra la matriz de confusión para estas etiquetas con las coincidencias y discrepancias de las etiquetadoras.

**Tabla 4.**

*Matriz de confusión para el etiquetado A-O*

| Etiquetas | A | O | Total |
|---|---|---|---|
| **A** | 89 | 21 | 110 |
| **O** | 12 | 135 | 147 |
| **Total** | **101** | **156** | **257** |

**Fuente:** Elaboración propia (2024).

En la segunda línea de la segunda columna, de izquierda a derecha, la cifra 89 representa el número de veces que las dos etiquetadoras le asignaron una etiqueta de *apoyo* a un mismo comentario, mientras que la tercera línea de la tercera columna, con la cifra 135, contiene el número de coincidencias en la asignación de una etiqueta de *odio* para un comentario.

Las cifras 21 y 12, por su parte, muestran el número de desacuerdos, en los que una etiquetadora asignó la etiqueta *odio* y la otra la de *apoyo*. La cuarta columna incluye la suma de las filas y la cuarta fila, la suma de las columnas. Mientras el cálculo de *Po* es simple (el número de coincidencias, 89+ 135 = 224, entre el total de juicios, 257, es igual a 0,87), el cálculo de *Pe* requiere de los totales de columnas y filas, según muestra el procedimiento mostrado en (2).

(2)

$$\mathrm{Pe} = \frac{110}{257} \times \frac{101}{257} + \frac{147}{257} \times \frac{156}{257} =$$

$$\mathrm{Pe} = 0{,}42 \times 0{,}39 + 0{,}57 \times 0{,}60 =$$

$$\mathrm{Pe} = 0{,}16 + 0{,}34 = 0{,}5$$

$$\mathrm{Pe} = 0{,}5$$

El cálculo de *Pe* requiere la división de las sumas de todas las filas (110 y 101) y las columnas (147 y 156), por el total de todos los juicios (257), y la multiplicación de todos estos cocientes. Una vez obtenido *Pe*, el despeje de la fórmula (1) es mostrado en (3) a continuación.

(3)

$$K = \frac{0{,}87 - 0{,}5}{1 - 0{,}5}$$

$$K = \frac{0{,}37}{0{,}5} = 0{,}74$$

$$K = 0{,}74$$

Al final, el coeficiente Kappa de Cohen obtenido se compara con una escala estandarizada, donde el acuerdo entre etiquetadores es juzgado de la siguiente manera: < 0 (sin acuerdo), 0–0,20 (ligero), 0,21–0,40 (regular), 0,41–0,60 (moderado), 0,61–0,80 (sustancial) y 0,81–1,00 (casi perfecto) (DATAtab, 2024; GraphPad, 2024; Meta Analysis, 2024). De esta manera, el acuerdo alcanzado en las etiquetas *apoyo* y *odio* ha sido sustancial.

La Tabla 5 a continuación muestra los resultados de la etiquetación del grupo meta del comentario, I (Israel) y P (Palestina). Utilizando la formula (1) y siguiendo los procedimientos mostrados en (2) y (3), se obtiene un acuerdo sustancial en esta etiquetación.

**Tabla 5.**

*Matriz de confusión para el etiquetado I-P*

| Etiquetas | I | P | Total |
|---|---|---|---|
| **I** | 118 | 22 | 140 |
| **P** | 19 | 98 | 117 |
| **Total** | **137** | **120** | **257** |

**Fuente:** Elaboración propia (2024).

A partir de las coincidencias y discrepancias mostradas en la Tabla 5, se obtuvo un acuerdo observado, *Po*, de 0,84 (con las coincidencias, 118+ 98 = 216, divididas entre el total de juicios, 257). Por su lado, el cálculo del acuerdo que podría ser resultado del azar, *Pe*, obtuvo 0,48, con lo que el coeficiente Kappa de Cohen final fue de 0,69.

De esta manera, las etiquetas que producen los códigos de los cuatro vértices del cuadrado ideológico (OI, AO, OP y AI) alcanzan un nivel de acuerdo sustancial. Además, este acuerdo es cercano al casi perfecto en el caso del binomio A-O, que ultimadamente representa el sentimiento que se espera detectar con un algoritmo de aprendizaje automático.

## 4. Conclusiones

Este artículo comenzó expresando el objetivo de producir un corpus con validez científica que permitiera entrenar a un algoritmo para detectar el lenguaje de odio y apoyo característico del conflicto Israel-Palestina. Para ello, se diseñó un esquema de etiquetación que se inspiró en el discurso ideológico que aparece con este tipo de conflicto político, según el análisis del discurso. La sección anterior ha mostrado que la etiquetación ha alcanzado niveles de validez sustancial, no muy por debajo del nivel casi perfecto para la etiquetación del binomio *odio-apoyo*.

Para ponderar este logro habría que contextualizar un poco más la posible aplicación de este tipo de recurso. Si el corpus obtenido le permite a un algoritmo decidir si existe en la red la presencia de un discurso polarizado respecto de un conflicto político, habría también que distinguir la presencia de un discurso neutral, como los discursos informativos, que simplemente reportan la existencia del conflicto.

En un ejercicio de evaluación de la validez del corpus que incluye al trinomio *odio-apoyo-neutro*, el nivel de acuerdo entre las etiquetadoras obtuvo un coeficiente Kappa de Cohen de 0,69, el cual es inferior al del binomio *odio-apoyo*, 0,74, pero que sigue exhibiendo un nivel sustancial de acuerdo. De esta manera, el corpus presenta una validez aceptable en la distinción no sólo del discurso polarizado, sino también de la presencia del mismo respecto de un discurso neutro de carácter informativo.

Ahora bien, más allá de las cifras presentadas, sobre las cuales se ha argumentado ya que son muy buenas, falta contestar cuál podría ser una aplicación tangible de este tipo de recurso. Después de todo, hay que considerar que el corpus está en español, no en hebreo, ni en árabe.

Si uno tiene un clasificador que puede detectar la presencia de un discurso sobre un conflicto político, y puede detectar la polarización radical en el mismo, más allá de la presencia de un discurso neutro e informativo, no sólo se puede decir cuándo los miembros de una sociedad se están polarizando, sino que también se podría intentar prevenir ataques o agresiones de algún grupo que se asocia con una postura hacia el grupo que se asocia con la contraria.

Esto podría darse incluso en una comunidad externa al conflicto armado. De esta manera, este tipo de recurso podría tener utilidad para otras sociedades donde el impacto del enfrentamiento es más indirecto, pero no sin consecuencias. Este tipo de detección sería útil, por ejemplo, en comunidades con presencia de migrantes que se declaran como simpatizantes de algún grupo del conflicto, particularmente en sociedades que tienen un historial de ataques relacionados con estos grupos.

## 5. Referencias

Chilton, P. y Schäffner, C. (2011). Discourse, Ethnicity and Racism. En T. A. van Dijk. (ed.) *Discourse Studies: A Multidisciplinary Introduction* (2a ed.) (pp. 303-330). SAGE.

DATAtab (2024). *Cohen's Kappa*. https://datatab.net/tutorial/cohens-kappa

Davidson, T., Warmsley, D., Macy, M. y Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955

de Paula, A. F. M. y Schlicht, I. B. (2021). AI-UPV at IberLEF-2021 DETOXIS task: Toxicity Detection in Immigration-Related Web News Comments Using Transformers and Statistical Models. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) CEUR-WS*, 2943, 547-566. https://doi.org/10.48550/arXiv.2111.04530

exportcomments.com (2024). *Export Social Media Comments*. https://exportcomments.com/

Fortuna, P. y Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys, 51*(4), 1-30. https://doi.org/10.1145/3232676

Gagliardone, I., Gal, D., Alves, T. y Martinez, G. (2015). *Countering online hate speech*. United Nations Educational, Scientific and Cultural Organization.

Google API Client (2024). *Google API Python client library* (version 1.0.290.0). [Software]. https://github.com/googleapis/google-api-python-client

GraphPad (2024). *Quantify agreement with kappa*. https://www.graphpad.com/quickcalcs/kappa1.cfm

Jiwani, Y. y Richardson, J. E. (2011). Discourse, Ethnicity and Racism. En T. A. van Dijk. (ed.) *Discourse Studies: A Multidisciplinary Introduction* (2a ed.) (pp. 241-262). SAGE.

Jurafsky, D. y Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Language Natural Processing, Computational Linguistics, and Speech Recognition* (3a ed.). Available on line: https://web.stanford.edu/~jurafsky/slp3/

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M. y Brena, S. (2020). Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & Information Technology, 39*(7), 711-721. https://doi.org/10.1080/0144929X.2019.1607903

Manning, C. D., Raghavan, P. y Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Meta-analysis (2024). *Cohen's kappa free calculator*. https://meta-analysis.actilab.onl/cohen-kappa-free-calculator/#risultati

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. y Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*, 145-153. https://doi.org/10.1145/2872427.2883062

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C. y Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation, 55*(2), 477-523. https://doi.org/10.1007/s10579-020-09502-8

Rico-Sulayes, A. (2018). *Authorship Attribution on Crime-Related Social Media: Research on the darknet in forensic linguistics*. Aracne Editrice.

Rico-Sulayes, A. (2020). General Lexicon-Based Complex Word Identification Extended with Stem N-grams and Morphological Engines. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), Spain, CEUR-WS,* 2664, 15-23.

Rico-Sulayes, A. y Monsalve-Pulido, J. (2022). A Proposal and Comparison of Supervised and Unsupervised Classification Techniques for Sentiment Analysis in Tourism Data. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), Spain, CEUR-WS, 3202*, 1-8.

Rieder, B. (2015). *YouTube Data Tools* (Version 1.42) [Software]. https://ytdt.digitalmethods.net/

Spasić, L. W. y Buerki, A. (2020). Idiom-Based Features in Sentiment Analysis: Cutting the Gordian Knot. *IEEE Transactions on Affective Computing, 11*(2), 189-199. http://dx.doi.org/10.1109/TAFFC.2017.2777842

Villasís-Keever, M. Á., Márquez-González, H., Zurita-Cruz, J. N., Miranda-Novales, G. y Escamilla-Núñez, A. (2018). El protocolo de investigación VII. Validez y confiabilidad de las mediciones. *Revista Alergia México, 65*(4), 414-421. https://doi.org/10.29262/ram.v65i4.560

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138-142. https://doi.org/10.18653/v1/W16-5618

van Dijk, T. A. (2011). Discourse and Ideology. En T. A. van Dijk (ed.), *Discourse Studies: A Multidisciplinary Introduction* (2a ed.) (pp. 379-407). SAGE.

**Agradecimientos:** La presente investigación nace en el marco del proyecto RECOVER, ref. no. ProyExcel_00540, 2022-2025, Projects of Excellence, Andalusian Regional Government (Junta de Andalucía), PAIDI 2020.

La recolección de datos y la primera etiquetación han sido realizadas en la Universidad de las Américas Puebla, como parte de la investigación de Hannia Ramos Solís, estudiante de la licenciatura en Psicología e integrante del Programa de Honores para estudiantes de alto rendimiento académico, bajo la supervisión del autor de este artículo. La segunda etiquetación y los cálculos del nivel de confiabilidad han sido llevados a cabo con la colaboración de Flor de María Moreno González, integrante de un proyecto terminal de servicio social de la licenciatura de Idiomas en la misma institución. El trabajo de ambas ha sido invaluable para esta investigación.

## AUTOR:

**Antonio Rico Sulayes**
Universidad de las Américas Puebla, México.

Antonio Rico Sulayes es doctor en Lingüística Computacional por Georgetown University, en Washington, DC. Ha sido profesor de lingüística y computación en México, Colombia y los Estados Unidos. También ha trabajado como lingüista computacional para instituciones como la Organización Mundial de la Salud y para varios contratistas del Pentágono. Es autor de más de 40 artículos y de 3 libros de investigación. Ha dictado conferencias en una decena de países y es miembro fundador de la Asociación Mexicana de Procesamiento del Lenguaje Natural. Actualmente es profesor-investigador tiempo completo de la Universidad de las Américas Puebla, en México. Su investigación se enfoca en las tecnologías del lenguaje, la lingüística forense y los estudios léxicos.
antonio.rico@udlap.mx

**Índice H:** 6
**Orcid ID:** https://orcid.org/0000-0003-0932-4733
**Scopus ID:** https://www.scopus.com/authid/detail.uri?authorId=51462128400
**Google Scholar:** https://scholar.google.com/citations?user=6HfpVBIAAAAJ
**ResearchGate:** https://www.researchgate.net/profile/Antonio-Rico-Sulayes/stats