revista SIGNO Y PENSAMIENTO

Vol. 44
Publicación continua
2025

ISSN: 2027-2731 (En línea) | ISSN: 0120-4823 (Impreso - Descontinuado)

Artículos

Análisis crítico del modelo BERT en la construcción de controversias públicas: el caso Hamás-Israel [*]

A Critical Analysis of the BERT Model in the Construction of Public Controversies: The Case of Hamas-Israel

Análise crítica do modelo BERT na construção de controvérsias públicas: o caso do Hamas-Israel

Víctor M. Hernández L. [a]
Fundación Universitaria del Área Andina, Bogotá, Colombia
victorm.hernandez@urosario.edu.co
ORCID: https://orcid.org/0000-0003-2631-4350

Jaime E. Cuellar
Pontificia Universidad Javeriana, Bogotá, Colombia
ORCID: https://orcid.org/0009-0004-0858-2823

DOI: https://doi.org/10.11144/Javeriana.syp44.acmb

Recibido: 15 mayo 2025
Aceptado: 10 agosto 2025
Publicado: 14 octubre 2025

Resumen:

Este artículo propone una reflexión crítica sobre el uso de BERT, un modelo de procesamiento de lenguaje natural, como herramienta para clasificar posturas en controversias públicas. A través del estudio de caso del conflicto Hamás-Israel, se analiza cómo BERT actúa no solo como clasificador, sino como un dispositivo de inscripción que coconstruye versiones simplificadas de fenómenos sociales complejos. El corpus analizado se compone de más de 250 000 comentarios en español publicados en YouTube entre octubre de 2023 y enero de 2024. Mediante una muestra anotada manualmente y un modelo ajustado en TensorFlow, se entrenó a BERT para identificar siete categorías discursivas. Aunque el modelo alcanzó una precisión del 92,76#%, fue incapaz de predecir correctamente la categoría Anti-Hamás, revelando límites y sesgos en el modelo. El artículo articula herramientas de inteligencia artificial con la teoría del análisis de controversias y la teoría de los estudios de Ciencia, Tecnología y Sociedad (STS, por su sigla en inglés), para mostrar cómo los modelos de lenguaje no solo identifican patrones, sino que también participan activamente en la producción de sentido. Se concluye que el uso de estos modelos debe ser permanentemente interrogado desde marcos éticos, teóricos y epistémicos, pues más que reflejar la realidad, la construyen.

Palabras clave: BERT, Análisis de controversias, inteligencia artificial, conflicto Hamás-Israel, ciencia, tecnología y sociedad (STS).

Abstract:

This article offers a critical reflection on the use of BERT, a natural language processing model, as a tool for classifying stances in public controversies. Using the case study of the Hamas-Israel conflict, it analyzes how BERT functions not only as a classifier but also as an inscription device that co-constructs simplified versions of complex social phenomena. The corpus consists of over 250,000 Spanish-language comments published on YouTube between October 2023 and January 2024. Through a manually annotated sample and a fine-tuned TensorFlow model, BERT was trained to identify seven discursive categories. Although the model achieved an accuracy of 92.76 %, it failed to correctly predict the Anti-Hamas category, revealing limitations and biases within the model. The article brings together artificial intelligence tools with both controversy analysis theory and Science and Technology Studies (STS) to show that language models do not merely identify patterns, they also actively participate in the production of meaning. The article concludes that the use of such models must be continuously questioned through ethical, theoretical, and epistemic lenses, as they do not simply reflect reality but help construct it.

Keywords: BERT, Controversy Analysis, Artificial Intelligence, Hamas-Israel Conflict, Science and Technology Studies (STS).

Resumo:

Este artigo propõe uma reflexão crítica sobre o uso do BERT, um modelo de processamento de linguagem natural, como uma ferramenta para classificar posições em controvérsias públicas. Por meio do estudo de caso do conflito Hamas-Israel, ele analisa como o BERT atua não apenas como um classificador, mas também como um dispositivo de inscrição que co-constrói versões simplifi cadas de fenômenos sociais complexos. O corpus analisado consiste em mais de 250 000 comentários em espanhol publicados no YouTube entre outubro de 2023 e janeiro de 2024. Usando uma amostra anotada manualmente e um modelo ajustado no TensorFlow, o BERT foi treinado para identificar sete categorias discursivas. Embora o modelo tenha alcançado 92,76 %

Notas de autor

[a] Autor de correspondencia. Correo electrónico: victorm.hernandez@urosario.edu.co

de precisão, ele não conseguiu prever corretamente a categoria Anti-Hamas, revelando limites e vieses no modelo. O artigo articula ferramentas de inteligência artificial com a teoria da análise de controvérsias e a teoria dos Estudos de Ciência, Tecnologia e Sociedade (STS) para mostrar como os modelos de linguagem não apenas identificam padrões, mas também participam ativamente da produção de significado. Conclui-se que o uso desses modelos deve ser permanentemente questionado a partir de estruturas éticas, teóricas e epistêmicas, pois eles constroem, em vez de refletir, a realidade.
Palavras-chave: BERT, análise de disputas, inteligência artificial, conflito Hamas-Israel, ciência, tecnologia e sociedade (STS), Ciência, Tecnologia e Sociedade (STS).

Introducción

En un escenario cada vez más atravesado por tecnologías de procesamiento del lenguaje natural (PLN), modelos de inteligencia artificial a partir de redes neuronales como Bidirectional Encoder Representations from Transformers (BERT) se han consolidado como herramientas ampliamente utilizadas para tareas de clasificación textual. Este modelo, que procesa lenguaje natural teniendo en cuenta el contexto de escritura, surge como un insumo para modelos que hoy conocemos como Generative Pre-trained Transformer (GPT). La capacidad de BERT para captar relaciones contextuales complejas entre palabras lo convierte en un candidato atractivo para analizar fenómenos comunicativos y sociales complejos, como lo son las controversias públicas. Sin embargo, el uso de estas aplicaciones conlleva preguntas fundamentales: ¿puede una red neuronal representar fielmente una controversia? ¿Qué pierde y qué gana el análisis social al apoyarse en modelos preentrenados y algoritmos de clasificación?

Para este artículo se propone una reflexión de BERT como modelo de clasificación de textos, utilizando el estudio de caso del conflicto Hamás-Israel. Se reflexiona críticamente sobre BERT como dispositivo de inscripción en contextos de polarización. A partir de un estudio de caso, la controversia que emergió tras el ataque de Hamás a Israel el 7 de octubre de 2023 y la posterior respuesta militar israelí, se examina el comportamiento del modelo ante un corpus altamente polarizado, compuesto por más de 250 000 comentarios en español publicados en YouTube durante los tres primeros meses del conflicto. Por lo cual, la pregunta central es ¿qué controversia del conflicto Hamás-Israel es posible construir a partir de un modelo como BERT?

El 7 de octubre de 2023, la noticia de que Hamás había lanzado misiles desde la Franja de Gaza a un grupo de ciudadanos israelíes avivó la guerra con Israel ("Qué es Hamás, el grupo islamista militante que lanzó un ataque sin precedentes contra Israel", 2023). Israel respondió bombardeando zonas de la Franja, marcando el inicio del escalamiento del conflicto alrededor de las zonas israelíes y palestinas ("Guerra entre Israel y Gaza, 8 de octubre - Más de 1.100 muertos en el ataque de Hamás y la represalia de Israel", 2023). Lo cierto es que, aunque recrudecido, el conflicto tiene más de medio siglo de historia y se origina luego de que actores israelíes de carácter sionista se establecieron en el territorio que antes de la Segunda Guerra Mundial se denominaba Palestina y fundaron el Estado de Israel.

A partir de los acontecimientos de octubre de 2023, numerosos expertos, organizaciones y medios de comunicación han seguido de cerca los hechos. Lo mismo ha ocurrido en redes sociales, en las que personas en todo el mundo han expresado sus opiniones sobre el conflicto. Por una parte, algunos argumentan la justa insurgencia por parte de grupos armados palestinos que consideran ilegítimo el poder de Israel y lo señalan de ser un "Estado colonialista" y "usurpador de tierras palestinas". Por otro lado, existen personas que mencionan que "Israel tiene derecho a defender su territorio", justificando la ofensiva a partir del lanzamiento de misiles de Hamás y calificando los bombardeos israelíes subsecuentes como parte de la dinámica de la guerra. La naturaleza actual, polarizante y diversa de estas posiciones hace manifiesta la existencia de una controversia desde la perspectiva de Zielinski *et al*. (2018).

En lugar de situar el foco analítico en el conflicto mismo, nuestro interés se dirige hacia las operaciones técnicas y epistémicas que implica aplicar un modelo como BERT para clasificar posturas en disputas públicas

como esta. El estudio de caso de Hamás-Israel funciona como un medio para explorar en qué medida un modelo como BERT permite (o no) captar las complejidades discursivas, emocionales y políticas involucradas en controversias altamente polarizadas. Además, permite ver cómo las decisiones técnicas, como el etiquetado manual, el diseño del conjunto de entrenamiento o la definición de categorías, participan activamente en la producción de sentido y en la construcción de la controversia misma.

A lo largo del texto se argumentará que BERT no solo identifica patrones, sino que también coconstruye una versión de la controversia basada en supuestos lingüísticos y estructurales que deben ser interrogados críticamente. En consecuencia, este artículo invita a pensar los modelos de inteligencia artificial no como espejos del mundo, sino como dispositivos que requieren de una mirada ética y epistémica situada, como lo propone Mackenzie (2017).

Revisión de literatura

Autores como Law (2015) reflexionan sobre la importancia de los métodos en la tecnociencia y sostienen que estos no solo configuran lo social, sino que, en consonancia con los planteamientos de los STS, participan activamente en la conformación de la realidad misma. En esta línea, en el presente artículo analizamos las implicaciones del modelo BERT como un agente configurador de lo social. Por su parte, Latour (1979) introduce el concepto de *dispositivo de inscripción*, entendido como cualquier elemento capaz de transformar una sustancia material en una figura, diagrama o trazo que deje una huella de la sustancia original. Aunque en este caso no se parte de una sustancia material en sentido estricto, proponemos reflexionar sobre cómo BERT transforma los comentarios en YouTube, produciendo representaciones que no solo dejan una huella de los datos originales, sino que también contribuyen a la construcción de la controversia que se busca analizar. En consonancia con estas ideas, Mackenzie (2017) sostiene que el aprendizaje automático consiste en procesos computacionales que generan enunciados transformados en gráficas, cifras y otros formatos, los cuales reconfiguran la producción de conocimiento. Esta dinámica implica que dichos sistemas no solo describen la realidad, sino que también producen una ontología estadística que la constituye.

Dado que el proceso mediante el cual se hace la reflexión del modelo BERT es un análisis de controversias, vemos importante llevar a cabo un desarrollo del concepto. El *análisis de controversias* ha sido abordado desde distintas disciplinas, especialmente por las ciencias sociales y las ciencias computacionales. Desde las ciencias sociales, Marres (2015) y Venturini y Munk (2022) proponen el mapeo de controversias como una metodología que busca describir los debates públicos, o que, en términos de Marres (2015), "implica el uso de técnicas computacionales para detectar, analizar y visualizar la comparación pública sobre asuntos de actualidad" (p. 3). Venturini y Munk (2022) ubican de igual forma el mapeo de controversia como una metodología propia de la teoría del actor red (TAR) propuesta por el filósofo y sociólogo Bruno Latour. Aunque el método se sitúa en la investigación de los debates sociotécnicos y en general en estudios sociales de la ciencia, cada vez se utiliza más para mapear cualquier tipo de controversia (Marres, 2015). Para Venturini y Munk (2022), los elementos metodológicos fundamentales de un mapeo de controversia son seguir los actores, ver las proporcionalidades y los pesos de la controversia, no olvidar nuestra posición dentro de la controversia, estudiar los actores dentro de sus contextos propios, realizar visualizaciones legibles y dejar los datos abiertos para futuras investigaciones (p. 7). Para Marres (2015), los algoritmos o en general las tecnologías utilizadas para analizar las controversias en línea ejercen una influencia significativa en el desarrollo de las mismas. Por este motivo, enfatiza en la necesidad de abordar el sesgo digital en el análisis de controversias, proponiendo un cambio del análisis de controversias al *mapeo de temas*, el cual implica utilizar métodos computacionales en el análisis de una controversia, teniendo en cuenta sus limitaciones y sesgos, además de cuidar el proceso de etiquetado de los datos.

Desde las ciencias computacionales, las controversias se han enfocado más en el desarrollo técnico. Desde el campo del PLN, se han desarrollado modelos computacionales para detectar y clasificar controversias mediante técnicas como el análisis de sentimiento, el *topic modeling* y los *word embeddings* (Németh, 2023). Algoritmos como K-Vecinos Más Cercanos, Support Vector Machines o redes neuronales han mostrado buenos resultados en tareas de clasificación (Godara y Kumar, 2019; Dori-Hacohen y Allan, 2015; Habernal y Gurevych, 2015), pero es con modelos como BERT que se ha logrado una mejora sustancial en la identificación de posturas de la controversia misma (De Zárate *et al*., 2020).

Existen diversos casos que se ha trabajado con herramientas computacionales para dar cuenta de controversias en línea. Diversos estudios han abordado el conflicto en redes sociales como Reddit, X y YouTube, con un enfoque predominante en la narrativa Israel-Palestina, más que Israel-Hamás. En Reddit, Nushin *et al*. (2024) analizan sentimientos y tópicos geopolíticos mediante VADER, aprendizaje automático y Latent Dirichl Location (LDA), concluyendo que la mayoría de los usuarios mantienen una postura neutral; mientras que Guerra *et al*. (2024) se enfocan en opiniones extremas ligadas a eventos específicos, como los bombardeos al Hospital Al Quds y al campo de refugiados de Jabalia. En X, Qiu *et al*., (2024) emplean LDA para la identificación de temas, lexicones emocionales y modelos *zero-shot* para mostrar preocupaciones centradas en el daño y la traición, identificando una carga emocional fuerte tanto en mensajes anti-Hamás como anti-Israel. Alamsyah *et al*. (2024) revelan una marcada polarización a través del análisis de redes, con una mayoría de nodos pro-Palestina. En cuanto a YouTube y X, Rico-Sulayes (2025) desarrolla un corpus para detectar lenguaje de odio y apoyo mediante cuatro categorías que inspiraron este documento; Maathuis y Kerkhof (2024) utilizan Large Language Models (LLM) para interpretar narrativas complejas en los debates sobre el conflicto, y Liyih *et al*. (2024) aplican el aprendizaje profundo para clasificar sentimientos, destacando así el valor de los comentarios en YouTube como un insumo clave para analizar fenómenos sociales de esta índole. Aquí, entonces, se evidencia el despliegue de elementos computacionales que se ha implementado para entender el problema Israel-Palestina. En este texto queremos reformular el asunto hacia una polarización Hamás-Israel y nombrarlo como *controversia*.

Lo anterior conduce a que BERT se ha consolidado como un modelo ampliamente utilizado por su capacidad para captar el contexto completo de las palabras en un texto. A diferencia de modelos previos que procesan el lenguaje en una sola dirección, el modelo BERT lee simultáneamente hacia adelante y hacia atrás, lo que le permite reconocer matices y ambigüedades propios del discurso (Devlin *et al*., 2019). Esta propiedad resulta especialmente útil para el análisis de controversias, ya que facilita la identificación de posiciones contrarias, de tonalidades implícitas y de desplazamientos en el sentido de los enunciados. El análisis de controversias se convierte así en un medio ideal para analizar críticamente el modelo BERT, ya que sirve para explorar hasta qué punto sus representaciones contextualizadas permiten identificar o construir temas y narrativas más recurrentes, distinguir matices de postura y polarización en el discurso, y trazar la evolución temporal de las distintas voces involucradas.

El caso Israel-Hamás ofrece condiciones particularmente idóneas para esta exploración. La gran cantidad de contenido textual generado en plataformas como X, YouTube y Reddit proporciona un corpus extenso y diverso, ideal para entrenar y evaluar modelos de lenguaje con representaciones profundas. La rápida evolución temática del conflicto, desde el ataque inicial hasta las discusiones geopolíticas y humanitarias posteriores, exige un análisis contextualizado bidireccional, algo para lo cual BERT está especialmente diseñado. Además, la fuerte polarización, con posturas marcadamente pro-Israel y pro-Palestina, favorece tareas como la detección de postura (*stance detection*), en la que BERT ha mostrado altos niveles de precisión (Liyih *et al*., 2024). El conflicto también abarcó múltiples fases narrativas y una complejidad emocional significativa, lo que lo convierte en un caso propicio. Finalmente, al tratarse de un debate con profundas implicaciones mediáticas y geopolíticas, las explicaciones generadas por BERT se convierten en un punto clave de reflexión, pues evidencian las consecuencias sociales que pueden derivarse del uso de estos modelos.

Metodología

Este estudio adopta un enfoque metodológico que combina elementos del mapeo de controversias (Venturini, 2010; Venturini y Munk, 2022) con herramientas del PLN, en particular, con un modelo BERT ajustado para clasificación multicategoría. Nuestro propósito no fue únicamente obtener un modelo de alta precisión, sino interrogar los procesos mediante los cuales la controversia se traduce en datos, anotaciones, categorías y vectores numéricos. De esta manera, la metodología cuantitativa de este texto tiene en cuenta los siguientes puntos.

1) Se realizó una recolección de comentarios escritos en español en vídeos de canales de YouTube de prensa que resultan de la búsqueda "Hamás - Israel - Gaza - Palestina" y que han sido publicados desde el 7 de octubre del año 2023 al 7 de enero de 2024. La elección de esta temporalidad es relevante porque da cuenta de los meses iniciales de la controversia reciente y de sus reacciones. La selección de YouTube responde a su relevancia como espacio de deliberación pública en tiempo real y a su accesibilidad para la recolección de datos a partir de la API de YouTube.

Se enlistó un total de 289 vídeos de YouTube en español resultado de las búsquedas: "Hamas español", "Israel español", "Gaza español" y "Palestina español". Esto con el fin de atender a vídeos que tuviesen los principales actores de la controversia y asegurar el lenguaje de los mismos. Los vídeos, en su mayoría, provienen de páginas de noticias internacionales como *DW, CNN, Euronews, El País*, entre otros. En pequeña proporción, aparecen vídeos de creadores de contenido sin afiliación periodística y vídeos que pretenden educar y explicar contextos desde y para la comunidad general; esta diversidad de videos y su selección aleatoria busca reflejar la naturaleza múltiple de la controversia y del público interesado en ella. La captura de los comentarios se realizó con YouTube Data API v3 el 8 de enero de 2024. La base de datos generada contiene 270 573 filas y 6 columnas, en las que se incluye autor, fecha de publicación, número de *likes*, texto del comentario, ID del vídeo y estatus (si es público o no).

2) Se llevó a cabo un preprocesamiento de la base de datos utilizada en el modelo. Esto incluye la limpieza de comentarios vacíos, de comentarios con letras o de emojis sin referencia a la controversia, además de comentarios realizados por fuera de las fechas propias del periodo elegido.

3) Se tomó una base de subdatos de 500 observaciones elegidas aleatoriamente dentro de la base de datos total. En esta investigación, el proceso de la elección de las etiquetas que se iban a utilizar surgió de una identificación no automatizada de entidades y de disposiciones textuales dentro de cada uno de los comentarios revisados. Las etiquetas enlistadas son el resultado de un preanálisis de los comentarios que resultan en tipos de clasificaciones sugeridas explícitamente por los mismos datos y escogidos por los investigadores, además de la revisión de literatura (Rico-Sulayes, 2025). Este proceso se denomina como *anotación basada en caracteres*, según Pustejovsky y Stubbs (2012). Así, la inclusión o exclusión de un comentario dentro de una etiqueta u otra está basada específicamente en su sintaxis gramatical y no tanto en su significado implícito. Lo anterior con el fin de tomar decisiones de clasificación lo suficientemente parecidas a las que la máquina podría llegar a realizar. Al final, cada una de las etiquetas dispone de un valor numérico asignado para mejor rendimiento del modelo a desarrollar.

El PLN se basa en anotaciones de sus textos para "evaluar las nuevas tecnologías del lenguaje humano y, fundamentalmente, para desarrollar modelos estadísticos fiables para la formación de estas tecnologías" (Ide y Pustejovsky, 2017, p. 2). Es por eso que el proceso de anotación se llevó a cabo de manera organizada y se pensó en cuanto a la estructura lingüística de lo que se quiere analizar. Finlayson y Erjavec (2017) proponen un esquema como primer paso de anotación o de etiquetaje de datos dentro de un proyecto que implique PLN. Los autores mencionan el esquema MATTER, por sus siglas en inglés, como el más adecuado, con algunas adecuaciones. Este esquema intenta establecer la ruta de la realización de una máquina de aprendizaje para PLN de la siguiente forma:

- M = modelado. Se establece el marco conceptual del proyecto.
- P = obtención (*procure*). Ver las anotaciones más apropiadas para el proceso. Paso agregado por los autores.
- A = anotación. La aplicación de anotaciones.
- TT = entrenamiento y testeo (*training and testing*). Entrenamiento y testeo del modelo de *machine learning*.
- E = evaluación.
- R = revisión.
- D = distribución. Compartir los resultados con la comunidad científica y el mundo en general. Paso agregado por los autores.

De esta manera, el proceso metodológico presente en este artículo cumple la estructura de "MATTER +PD".

Primero se creó una estructura conceptual del modelo, ejemplificado aquí como una revisión de antecedentes y de contextualización del fenómeno que se va a analizar. En segundo lugar, se estableció un proceso de generación de etiquetas, o *targets*, basados en la revisión de una cantidad estimada de comentarios y de análisis de contexto dispuestos por muestreo aleatorio. Se propuso como meta alcanzar 200 datos etiquetados por categoría para generar un total de 1400 datos. Este total de datos etiquetados corresponden a un aproximado de 0,55 % de los datos totales. La elección de la cantidad de comentarios se realizó teniendo en cuenta la potencia de la red neuronal que se iba a utilizar. Al ser BERT una red neuronal preentrenada en grandes cantidades de datos sin etiquetar, permite adquirir una comprensión profunda del lenguaje y de las relaciones contextuales entre palabras. Así se realizó el modelo, la evaluación y la distribución. Este último paso de la estructura coincide con lo mencionado por Venturini y Munk (2022) en cuanto a un mapeo de controversias dispuesto para la discusión pública.

Luego de analizar los datos, se llegó a una categorización multiinvestigador que fue posteriormente validada por un tercero. A continuación, se especificará en qué consiste cada etiqueta:

- *Comentarios no relacionados (NR - Label: 4)*: aquellos comentarios que no tienen relación con la controversia analizada. Esta categoría engloba comentarios de algunos vídeos que no solamente presentaban noticias relacionadas con el conflicto Hamás-Israel, sino también resúmenes de otras noticias internacionales. Así, se relacionaban con temas no relacionados con la controversia analizada. De la misma manera, algunos comentarios atacaban a personas entrevistadas por sus condiciones políticas sin conexión explícita con la controversia.
- *Comentarios sin postura (SP - Label: 3)*: comentarios que no toman postura explícita hacia Israel o Hamás, o que atacan a ambos bandos por igual. Aquí se ubican aquellos comentarios que atacaban explícitamente el conflicto en general con expresiones como "Paz para el mundo" o "La guerra no es el camino". Igualmente, en esta categoría se encuentran los comentarios que atacan explícitamente ambas posiciones del conflicto cuando mencionan que "Tanto Israel como Hamás deben pagar por lo que hacen". Esta categoría no entra en el análisis de controversia porque no toma una postura definida hacia alguno de los dos actores. Muchos de los comentarios de etiqueta SP alegorizan a Dios como único poder capaz de acabar con la guerra. En esta etiqueta también se ubican comentarios a vídeos subidos por noticieros o a presentadores de los mismos, a gobiernos alrededor del mundo, a ideologías económicas y políticas, y a otros elementos que no tenían que ver directa o explícitamente con la controversia analizada.
- *Comentarios en apoyo al pueblo palestino (Pro-Palestina - Label: 6):* comentarios que apoyan al pueblo palestino, por ejemplo, calificando la ofensiva como un genocidio. Aquí se engloban aquellos comentarios como "Palestina libre" o que se muestran explícitamente preocupados por la situación

de la población que habita la Franja de Gaza. Algunos ejemplos de esta etiqueta son "Viva Palestina!", "PAZ, JUSTICIA Y EQUIDAD también para PALESTINA...".

- *Comentarios apoyo al Estado de Israel (Pro-Israel - Label: 5)*: comentarios que apoyan a Israel y su defensa contra Hamás. Esta categoría acoge comentarios que explícitamente defienden el actuar de Israel, como "Dios bendiga a Israel", o que se muestran preocupados por la situación que vive la población israelí. De igual manera, señalan a Hamás por atacar primero y defienden la legítima defensa del pueblo de Israel. Aquí también se encuentran los comentarios más explícitamente violentos por el uso de "groserías" o palabras soeces, como "Estan atacando a Israel y los HP de Iran no quieren quedae defienda... [*sic*]".
- *Comentarios que señalan a Israel (Anti-Israel - Label: 1)*: comentarios que señalan a Israel de cometer actos inmorales y lo consideran como el principal culpable del conflicto. Algunos comentarios ubican a Benjamin Netanyahu, primer ministro de Israel, al gobierno y a políticos sionistas como los principales responsables. De la misma manera, esta etiqueta engloba aquellos comentarios que señalan explíticamente a Israel como culpable del genocidio palestino. Algunos ejemplos de esta etiqueta son "Yanquis e israelís genocidas!", "Netanyahu, como candidato a dictador...", "Israel es el malo de la historia!" [*sic*].
- *Comentarios que señalan a la población palestina (Anti-Palestina - Label: 2)*: comentarios que señalan que los asesinatos de la población palestina son causados por la misma o la señalan de ser terroristas. Dentro de esta etiqueta se encuentran aquellos comentarios que señalan a la población palestina como la culpable del conflicto, o que atacan su religión y sus creencias, además de su posición política. De la misma manera, se encuentran comentarios que tienen un tono agresivo hacia las personas de Gaza: "Por qué los palestinos no hacen nada contra Hamas", "Ellos atacan y luego son las víctimas" [*sic*].
- *Comentarios contra Hamás (Anti-Hamás - Label: 0)*: comentarios que explícitamente atacan a Hamás y lo señalan como un grupo terrorista. Aquí se encuentran comentarios que señalan a Hamás como el principal culpable del conflicto. "Hamás debe de ser destruido" o "Terroristas los de Hamás" son algunos ejemplos en los que se señala explícitamente a este grupo armado.

El proceso de anotación fue manual y fue realizado por investigadores sociales, quienes identificaron siete categorías discursivas: Pro-Palestina, Pro-Israel, Anti-Israel, Anti-Palestina, Anti-Hamás, Sin postura y No relacionado. Se etiquetaron 1400 comentarios mediante muestreo aleatorio estratificado, con un mínimo de 200 observaciones por categoría. Esta base se usó para entrenar el modelo BERT ajustado en TensorFlow.

Se siguió el esquema MATTER+PD (Modelado, Adquisición, Anotación, Entrenamiento, Evaluación, Revisión, Distribución), enfatizando que cada etapa implicó decisiones interpretativas. Las categorías no fueron impuestas *a priori*, sino que se derivan de un análisis inductivo de los comentarios, aunque su traducción a etiquetas numéricas implicó necesariamente una reducción semántica.

El modelo BERT empleado fue afinado sobre esta base anotada durante 15 épocas, alcanzando una precisión del 92,76 %. Se usaron representaciones contextualizadas del lenguaje (*embeddings*) y mecanismos de atención para identificar los *tokens* más relevantes en cada predicción. Aun así, el modelo fue incapaz de predecir comentarios Anti-Hamás en la fase de inferencia, pese a haberlos recibido en el entrenamiento, lo que evidencia problemas de ambigüedad semántica, sesgos en los datos y límites en la arquitectura del modelo.

Esta sección metodológica no pretende presentar un procedimiento neutral, sino explicitar cómo las decisiones técnicas, desde la selección de corpus hasta la configuración del clasificador, participan en la forma en que se representa la controversia. En vez de buscar la generalización, proponemos este diseño como un ejercicio reflexivo sobre el ensamblaje técnico-social que supone usar la inteligencia artificial para leer disputas humanas.

4) Se utilizó el modelo BERT para la clasificación de texto. Todo el código fue realizado en Python versión 3.

El modelo BERT utilizado corresponde a una versión ajustada específicamente a las necesidades del estudio, para esto se recurrió a TensorFlow. Este ajuste se realizó con el fin de optimizar el desempeño del modelo en la tarea específica, considerando las características particulares de los datos y los objetivos del análisis. En este modelo primero se *tokenizan* los datos, creando tuberías de entrada con TensorFlow junto con la API "tf.data" para su posterior entrenamiento y evaluación. Esta API se encuentra pre-entrenada y guardada como modelo "codificador para incrustaciones de texto con codificadores de transformador". Se espera un dict con tres en 32 Tensores como entrada: input_word_ids, input_mask, y input_type_ids" (Shrikant, 2023).

El tokenizador de BERT divide la cadena de texto en unidades básicas de significado o *tokens*. Posteriormente, las cadenas pasan a una máscara de entrada: una matriz que se utiliza para indicar qué *tokens* son importantes para la tarea que se está realizando. A su vez, los identificadores de tipo de entrada son números que se utilizan para representar el tipo de cada *token*. Luego de este proceso, pasan a la capa BERT, la parte central del modelo. Esta capa es una red neuronal que aprende a representar el significado de las palabras en contexto. Al final, todo se encapsula en Keras. Es precisamente la última capa de Keras la que asegura que el modelo funcione como clasificador. Al ser un modelo utilizado para otras actividades de clasificación, fue más fácil adecuar el código realizado en Python para que cumpliera con el entrenamiento y con el criterio de predicción descritos por el investigador.

Dentro de BERT se encuentra un complejo de redes neuronales que utilizan aprendizaje en contexto. Una de las características clave de BERT es que es bidireccional. Esto significa que el modelo puede aprender el significado de una palabra en función de las palabras que la preceden y la siguen. Esto le permite a BERT comprender mejor el contexto de una palabra y, por lo tanto, realizar mejor las tareas de PNL (Devlin *et al*., 2019). Sin embargo, algunos artículos mencionan que esto puede presentar sesgos, dependiendo de los datos de entrenamiento. Así lo mencionan De Zárate *et al*. (2020) cuando revisan distintos modelos de *machine learning* para encontrar controversias. Los autores mencionan que, aunque no exista tal controversia, el modelo es capaz de percibirla gracias a su potencia.

Una vez dispuesto el modelo se procedió a entrenarlo con los 1400 datos previamente etiquetados en 7 categorías. El modelo final fue entrenado en 15 épocas resultado de prueba y error. Cuando el entrenamiento culminó en la época número 15, el modelo presentó una precisión general (*accuracy*) de 92,76 % y un f1-score, que muestra la precisión de clasificación, de 90,08 %. Debemos explicar que estas métricas se toman sin la nula predicción de la categoría Anti-Hamás, algo de lo que hablaremos más adelante. La precisión del modelo (*accuracy*) refiere a la proporción de predicciones correctas realizadas con respecto al total de predicciones posibles. Tomamos como medida de referencia para evaluación del modelo la precisión (*accuracy*), ya que los datos de entrenamiento se encuentran balanceados. Posteriormente, se tomó el modelo entrenado y se inició el periodo de predicción para todos los datos. Después de realizar una limpieza de posibles comentarios previos al 7 de octubre del 2023, comentarios en blanco o con elementos que no consistían en una palabra o frase o *emoji* completo, se dejó un total de 253 925 comentarios con categoría predicha.

5) Se establecieron las discusiones generadas de los comentarios dentro de las clasificaciones generadas y dispuestas por cantidad, temporalidad y afinidad o "me gusta".

Resultados

Los resultados obtenidos mediante el modelo BERT permiten analizar cómo una arquitectura de red neuronal interpreta o reduce una controversia compleja a una serie finita de categorías. Lejos de presentar estas salidas como verdades objetivas, las entendemos como inscripciones que revelan tanto patrones recurrentes como zonas de ambigüedad y de falla.

La figura 1 muestra la cantidad de comentarios etiquetados por cada una de las categorías. Organizadas de mayor a menor, no es de extrañar que los comentarios que más se presentan son aquellos que no tienen que

ver con la controversia misma (Etiqueta 4, NR). Asimismo, los comentarios "sin postura" (Etiqueta 3, SP) puntúan en la predicción de los comentarios. Esto concuerda con los resultados obtenidos por Nushin *et al.* (2024) en Reddit. Este hallazgo podría reflejar una limitación del modelo, que, al asumir una alta frecuencia de categorías asociadas con la "neutralidad", en realidad estaría evidenciando su incapacidad para reconocer formas discursivas complejas.

Distribución general de las categorías

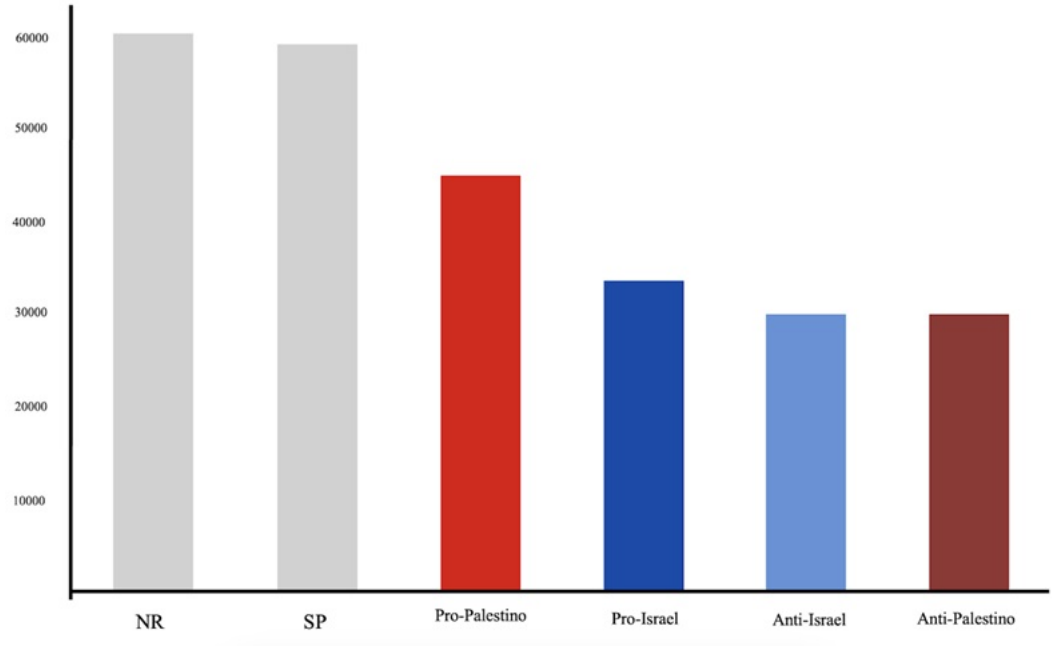

FIGURA 1.
Cantidad de comentarios etiquetados por cada categoría
Fuente: elaboración propia.

Entre las categorías que efectivamente toman postura en la controversia, se destacan las categorías Pro-Palestina (Etiqueta 6) y Pro-Israel (Etiqueta 5). Es aquí donde la controversia tiene lugar y se evidencian las dicotomías en los comentarios. Sin embargo, es notorio que la categoría Pro-Palestina tiene 34,16 % más de comentarios que la categoría Pro-Israel. Por último, se encuentran las categorías Anti-Israel (Etiqueta 1) y Anti-Palestina (Etiqueta 2) con casi la misma cantidad de datos: 29 884 y 29 820 comentarios etiquetados, respectivamente.

Invisibilidad de la categoría Anti-Hamás

Un hallazgo especialmente revelador fue la incapacidad del modelo para predecir comentarios Anti-Hamás, a pesar de haber sido entrenado con 200 comentarios explícitos de esa categoría. Esta ausencia no es menor: señala un límite crítico en la semántica operativa del modelo. Comentarios como "Hay que matar a todos los terroristas de Gaza y no dejar a nadie", aunque son explícitamente Anti-Hamás, fueron etiquetados como Pro-Israel o Anti-Palestina.

Este fenómeno podría explicarse por varias razones: baja representatividad de esa clase en el conjunto de entrenamiento, que no fue el caso, ya que se les suministró una cantidad suficiente de comentarios al igual que a las otras categorías; o interferencias semánticas con otras categorías similares, como Pro-Israel o Anti-

Palestina. En todo caso, esto ilustra que BERT no interpreta: infiere y ajusta en función de probabilidades, lo cual tiene implicaciones epistemológicas sobre qué discursos y categorías quedan invisibilizadas.

Dinámica temporal de la controversia

La figura 2 muestra, en un corte de dos semanas, el comportamiento de las diferentes categorías clasificadas. El número de comentarios relacionados con el conflicto Hamás-Israel se incrementó a más del doble en las dos primeras semanas posteriores al 7 de octubre. Es de destacar que el etiquetado automático del modelo evidencia una presencia constante de comentarios Pro-Palestinos a lo largo del tiempo. Sin embargo, se destacan algunos puntos importantes en el comportamiento de la controversia en los tres meses analizados.

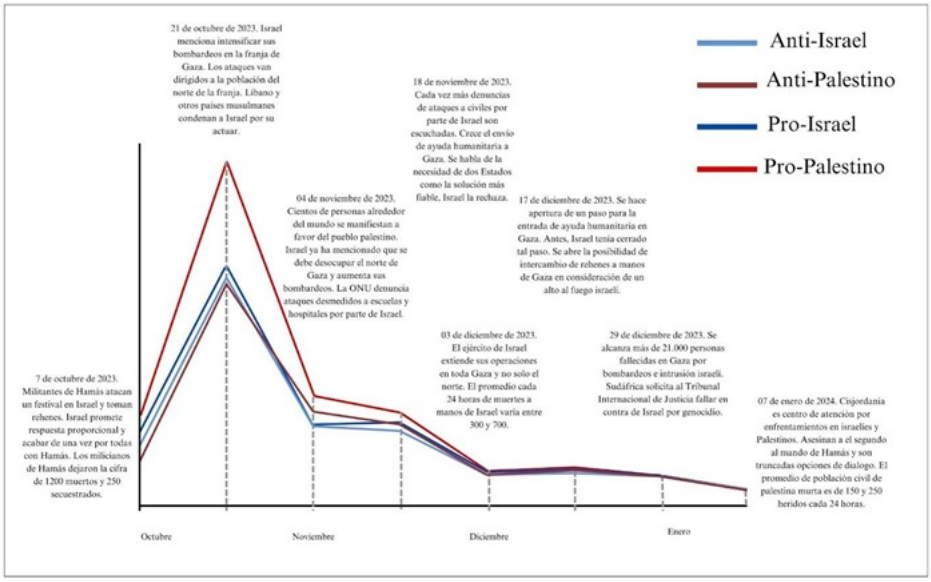

FIGURA 2.
Frecuencia de comentarios divididos cada dos semanas y su contexto
Fuente: elaboración propia.

Los comentarios Anti-Palestina no tuvieron una presencia significativa, excepto en la última semana del mes de octubre y la primera semana del mes de noviembre. En este punto, la noticia de lo ocurrido era de interés global y se argumentaba comúnmente la legítima defensa de Israel. Varios medios de comunicación se enfocaban en el contexto histórico de la insurgencia del grupo palestino. En este periodo es cuando la categoría Anti-Palestino aparece en segundo lugar, en cuanto a la cantidad de comentarios presentes en los vídeos.

Hay que destacar también que el *boom* mediático que significó el ataque de Hamás, y luego el ataque Israel, dejó consigo más comentarios realizados en etapas tempranas de la temporalidad de la controversia analizada que hacia el final de la misma. Respondiendo a esta dinámica, el análisis temporal se llevó a cabo según la fecha de realización del comentario y no de la publicación del vídeo.

Sin embargo, se da cuenta de un cambio en la composición de la controversia entre octubre y noviembre. Resulta interesante analizar que la cantidad de comentarios Anti-Israel desciende hacia la última posición y se mantiene así todo el mes de noviembre. En este periodo puntúan en segunda y tercera posición aquellos comentarios clasificados como Anti-Palestina y Pro-Israel. La dinámica del conflicto y su posición mediática hacen que la opinión pública tome una posición, atacando a mercenarios palestinos y defendiendo el actuar israelí.

Además, en la figura 3 se puede observar un ligero cambio hacia el final de las líneas. La figura 3 muestra que, pese a que se tomaron comentarios realizados solo en los primeros ocho días del mes de enero, el modelo

identifica más comentarios en la categoría Anti-Israel. Este cambio de comportamiento en la controversia puede significar un cambio en la opinión pública en general. Luego de más de 100 días del recrudecimiento del conflicto, las acciones realizadas por Israel en Gaza y la crisis humanitaria que vive el pueblo palestino aparecen más en el cubrimiento de los medios de comunicación.

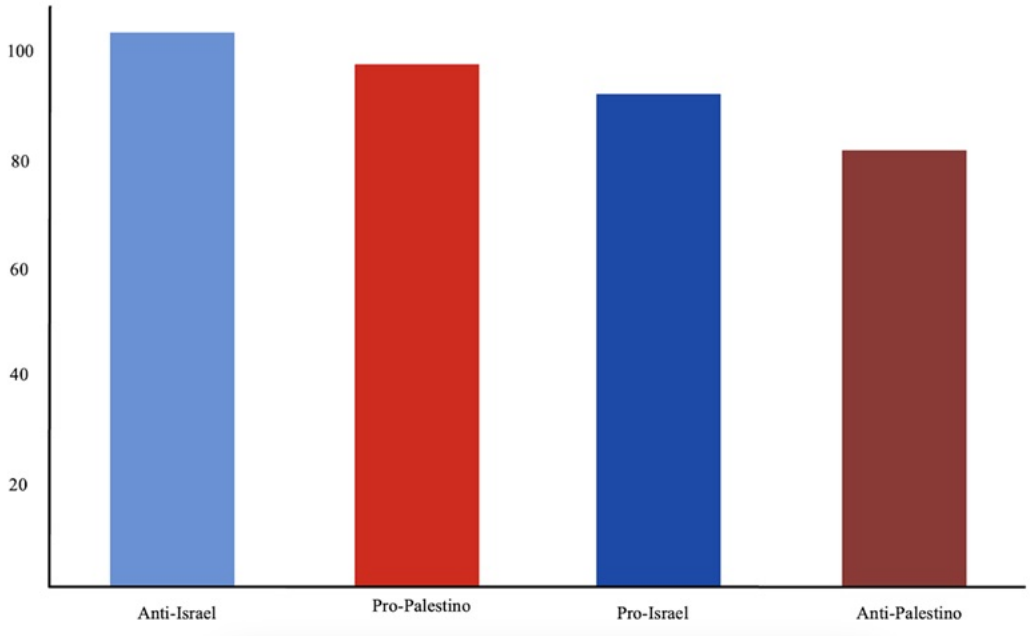

FIGURA 3.
Cantidad de comentarios para el mes de enero de 2024 dividido por categorías
Fuente: elaboración propia.

Se puede inferir que las crecientes cifras de fallecidos en Gaza y las denuncias realizadas por organismos internacionales, como la Organización de las Naciones Unidas (ONU), generaron un cambio en la opinión pública reflejada en los comentarios. Así, con el paso del tiempo, se da entrada a nuevos actores políticos y sociales a la controversia. Lo anterior demuestra que la controversia cumple con la característica de imposibilidad de estaticidad que, en este caso, está permeada por la opinión pública (Venturini y Munk, 2022). La modificación del discurso ante los crímenes de lesa humanidad perpetrados por Israel hacia mitad y finales de la temporalidad de la controversia analizada sugiere cómo la agenda de los medios de comunicación puede ocasionar un cambio en la controversia misma.

En este escenario, consideramos que coexisten dos posibles interpretaciones para el cambio observado en las controversias. La primera, ya mencionada, sugiere que dicho cambio responde a una transformación en la opinión pública, provocada por los acontecimientos previamente descritos. La segunda apunta a que la variación se relaciona con el propio funcionamiento del modelo (a partir del entrenamiento y diseño), así como con el algoritmo de recomendación de YouTube, y que ambos influyen directamente en los resultados de clasificación. Como señala Marres (2015), las tecnologías utilizadas para analizar controversias en entornos digitales no son neutrales, sino que ejercen una influencia significativa en la forma en que estas se desarrollan. Desde nuestra perspectiva, ambas dinámicas deben entenderse como fuerzas concurrentes en la configuración de las controversias detectadas.

Otro punto importante que hay que analizar es el promedio de *likes* por categoría. Exceptuando las categorías NR y SP, en promedio los comentarios con más *likes* hacen parte de las categorías Pro-Israel y Anti-Palestina. Esto quiere decir que, a pesar de que existen más comentarios Pro-Palestina, los comentarios más populares son aquellos que defienden a Israel o que atacan al pueblo palestino. Se concluye así que, aunque la participación de comentarios Pro-israelíes y Anti-palestinos son menos frecuentes que los comentarios

Pro-Palestinos y Anti-Israelí, estos son los que más reciben apoyo por parte de los usuarios de YouTube. Este hallazgo abre preguntas sobre la relación entre la visibilidad y el apoyo-participación en plataformas como YouTube, y sobre cómo los modelos de IA se pueden ver permeados por esto.

Afinidad expresada en likes

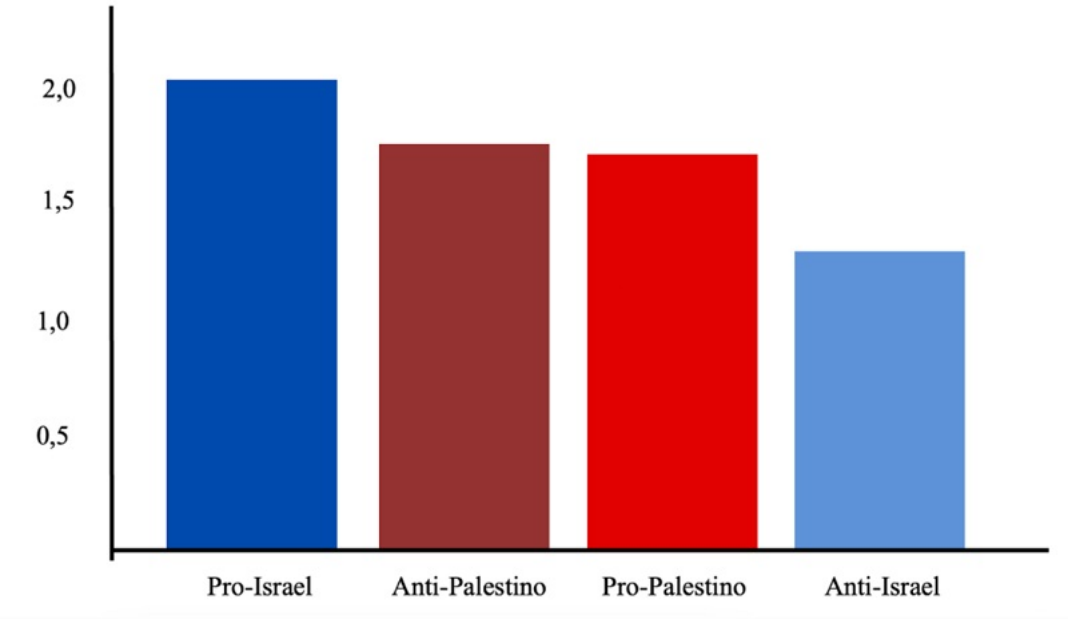

FIGURA 4.
Promedio de *likes* por cada categoría
Fuente: elaboración propia.

Conclusiones

Los resultados presentados deben leerse no solo como una medición cuantitativa del conflicto Hamás-Israel en redes sociales, sino como una muestra de la construcción de una controversia a partir de un modelo como BERT. La precisión estadística lograda (92,76#%) no debe ocultar el hecho de que los modelos de aprendizaje profundo no "descubren" controversias, sino que las producen. Según Law (2015) esto implica que este tipo de modelos están configurando lo social y la realidad misma, en este caso, lo que entendemos por el conflicto entre Israel y Hamás.

A lo largo del proceso de entrenamiento, clasificación y análisis, se hizo evidente que el modelo BERT actúa como un dispositivo de inscripción, en términos de Latour (1979), puesto que transforma en etiquetas lo que en el discurso humano son ambivalencias, ironías, contradicciones o silencios. La dificultad del modelo para identificar comentarios Anti-Hamás, pese a haber sido entrenado para eso, ilustra este punto de forma especialmente clara. No estamos ante una simple omisión técnica, sino ante una forma de invisibilización algorítmica de discursos complejos, que tiene implicaciones políticas. ¿Qué sucede cuando una red neuronal no logra distinguir entre una crítica a Hamás y una acusación general al pueblo palestino? ¿Qué tipo de lectura política impone el modelo cuando agrupa ambas cosas bajo la categoría de Anti-Palestina o Pro-Israel? Estas preguntas invitan a repensar el papel de los investigadores no solo como usuarios de modelos, sino como coconstructores de los sentidos que estos modelos producen (Mackenzie, 2017).

Además, es importante destacar que el modelo hereda los sesgos del corpus, del etiquetado y de su propio diseño. El uso exclusivo de videos en español en YouTube, el enfoque en medios tradicionales o la definición

de siete categorías cerradas son decisiones que repercuten en la construcción de la controversia, tal como lo menciona Marres (2015).

Este artículo no demerita el uso de BERT o de redes neuronales para el análisis de fenómenos sociales, de hecho, en cierta medida lo incentiva, porque su capacidad de procesamiento es destacada y constituye indudablemente una herramienta valiosa para el análisis de controversias. Lo que se propone es recomendar un uso crítico en el que su utilización debe ser permanentemente interrogada por marcos teóricos, metodologías reflexivas y criterios éticos. Esta investigación no responde a la pregunta ¿qué piensan los usuarios de YouTube sobre el conflicto Hamás-Israel?, sino más bien a ¿qué controversia es posible construir a partir de un modelo BERT? En ese desplazamiento reside el valor crítico de este trabajo.

Referencias

Alamsyah, A., Muharam, A. W. y Ramadhani, D. P. (2024). Polarized Narratives in Digital Spaces: A Social Network Examination of the Gaza Conflict. En *International Conference on Data Science and Its Applications* (ICoDSA) (pp. 527-532). IEEE. https://doi.org/10.1109/icodsa62899.2024.10651759

De Zarate, J. M. O., Di Giovanni, M., Feuerstein, E. Z. yy Brambilla, M. (2020). Measuring Controversy in Social Networks Through NLP. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). *Springer Nature, 12303* LNCS, 194-209. https://doi.org/10.1007/978-3-030-59212-7_14

Devlin, J., Chang, M. W., Lee, K. y Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, 1*, 4171-4186. https://doi.org/10.48550/arXiv.1810.04805

Dori-Hacohen, S., y Allan, J. (2015). *Automated Controversy Detection on the Web. Lecture Notes in Computer Science* (pp. 423-434). Springer International Publishing. https://doi.org/10.1007/978-3-319-16354-3_46

Finlayson, M. A. y Erjavec, T. (2017). Overview of Annotation Creation: Processes and Tools. En N. Ide y J. Pustejovsky (eds.), *Handbook of Linguistic Annotation* (pp. 167-191). Springer.

Godara, N. y Kumar, S. (2019). Opinion Mining using Machine Learning Techniques. *International Journal of Engineering and Advanced Technology, 9*(2), 4287-4292. https://doi.org/10.35940/ijeat.b4108.129219.

Guerra entre Israel y Gaza, 8 de octubre - Más de 1.100 muertos en el ataque de Hamás y la represalia de Israel. (2023, octubre 8). *El País.* https://elpais.com/internacional/2023-10-08/guerra-de-israel-en-gaza-en-directo.html

Guerra, A., Lepre, M. y Karakus, O. (2024). *Quantifying extreme opinions on Reddit amidst the 2023 Israeli-Palestinian conflict*. Elsevier. https://arxiv.org/abs/2412.10913v1

Habernal, I. y Gurevych, I. (2015). Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. En *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2127-2137). Association for Computational Linguistics. https://aclanthology.org/D15-1255/

Ide, N. y Pustejovsky, J. (2017). *Handbook of Linguistic Annotation*. Springer. https://doi.org/10.1007/978-94-024-0881-2

Latour, B. y Woolgar, S. (1979). *La vida en el laboratorio. La construcción de los hechos científicos*. Alianza Editorial.

Law, J. (2015). *STS as method.* The MIT Press

Liyih, A., Anagaw, S., Yibeyin, M. y Tehone, Y. (2024). Sentiment analysis of the Hamas-Israel war on YouTube comments using deep learning. *Scientific Reports, 14*(1). https://doi.org/10.1038/s41598-024-63367-3

Maathuis, C. y Kerkhof, I. (2024). Navigating Online Narratives on Israel-Hamas War with LLMs. *Proceedings of the International Conference on AI Research, 4(*1), 252-259. https://doi.org/10.34190/icair.4.1.2868

Mackenzie, A. (2017). *Archaeology of a Data Practice*. The MIT Press

Marres, N. (2015). Why Map Issues? On Controversy Analysis as a Digital Method. *Science, Technology, & Human Values, 40*(5), 655-686. https://doi.org/10.1177/0162243915574602

Németh R. (2023). A Scoping Review on the Use of Natural Language Processing in Research on Political Polarization: Trends and Research Prospects. *Journal of Computational Social Science, 6*(1), 289-313. https://doi.org/10.1007/s42001-022-00196-2.

Nushin, K. N., Zaman, M. S. U. y Ahmed, M. (2024). Analyzing Sentiment and Unveiling Geopolitical Perspectives: A Comprehensive Study of Reddit Comments on the Contemporary Israel-Palestine Conflict. En *International Conference on Electrical Engineering and Information y Communication Technology (ICEEICT)* (pp. 788-793). IEEE. https://doi.org/10.1109/iceeict62016.2024.10534429

Pustejovsky, J. y Stubbs, A. (2012). *Natural Language Annotation for Machine Learning*. O'REILLY.

Qiu, Y., Ye, Y., Zhang, X. y Luo, J. (2024). Moral Frameworks and Sentiment in Tweets: A Comparative Study of Public Opinion on the Israeli-Palestine Conflict. En *2021 IEEE International Conference on Big Data (Big Data)* (pp. 7122-7130). IEEE. https://doi.org/10.1109/bigdata62323.2024.10825487

Qué es Hamás, el grupo islamista militante que lanzó un ataque sin precedentes contra Israel. (2023, octubre 8). *BBC News Mundo*. https://www.bbc.com/mundo/articles/c907yed00x0o

Rico-Sulayes, A. (2025). Cómo entrenar a un algoritmo para detectar el lenguaje de odio en el conflicto Israel-Palestina. *European Public & Social Innovation Review, 10,* 01-16. https://doi.org/10.31637/epsir-2025-1199

Dori-Hacohen, S. y Allan, J. (2013). Detecting Controversy on the Web. En *Proceedings of the 22nd ACM international conference on Information y Knowledge Management (CIKM'13)*. (pp. 1845-1848). Association for Computing Machinery. https://doi.org/10.1145/2505515.2507877

Shrikant, N. (2023). BERT for text classification with TensorFlow. *GitHub*. https://github.com/shrikantnaidu/BERT-for-Text-Classification-with-TensorFlow/blob/main/Fine_Tune_BERT_for_Text_Classification_with_TensorFlow.ipynb

Venturini, T. (2010). Building on Faults: How to Represent Controversies with Digital Methods. *Public Understanding of Science, 21*(7), 796-812. SAGE Publications. https://doi.org/10.1177/0963662510387558

Venturini, T. y Munk, A. K. (2022). *Controversy mapping a field guide*. Polity Press.

Zielinski, K., Nielek, R., Wierzbicki, A. y Jatowt, A. (2018). Computing Controversy: Formal Model and Algorithms for Detecting Controversy on Wikipedia and in Search Queries. *Information Processing & Management,54*(1), 14-36. Elsevier BV. https://doi.org/10.1016/j.ipm.2017.08.005

Notas

* Artículo de investigación

Origen del artículo

Investigación realizada en el marco de la tesis de maestría del autor Víctor M. Hernández.

Cómo citar: Hernández L., V. M. y Cuellar, J. E. (2025). Análisis crítico del modelo BERT en la construcción de controversias públicas: el caso Hamás-Israel. *Signo y Pensamiento, 44*. https://doi.org//10.11144/Javeri ana.syp44.acmb