

Generación automática de contranarrativas con LLMs

Seminario Internacional Proyecto Real-UP



Avances en IA para la Detección y Combate del Discurso de Odio en RRSS

6 de Junio de 2024



Maite Martín Valdivia
Universidad de Jaén. Grupo SINAI
maite@ujaen.es

WARNING

Esta presentación contiene ejemplos de lenguaje ofensivo

No representan mis puntos de vista





Creado
en 2001



Especialistas en
PLN: Área de la IA
centrada en la
comprensión y
generación de
lenguaje humano



Detección
de Discurso
de odio y
lenguaje
ofensivo

Lucha contra el discurso de odio



Contrarrestar mediante una narrativa

- Postura contraria
- Informativa
- Veraz
- No ofensiva

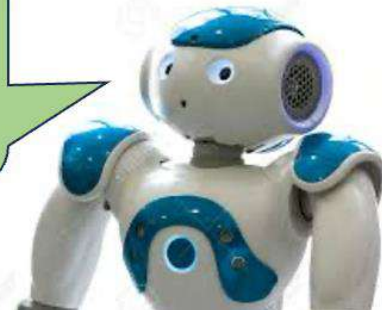
Objetivo

Utilizar técnicas de PLN y ML para generar automáticamente contranarrativas que contrarresten el discurso de odio mediante un mensaje positivo, constructivo y veraz



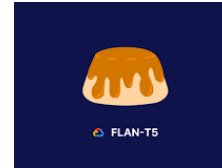
Los X\$%&/?@ nos roban el trabajo y arruinan nuestro país.

Los X\$%&/?@ a menudo asumen trabajos que los ciudadanos locales no desean realizar y, en muchos casos, contribuyen significativamente a la economía y la diversidad cultural de un país. La inclusión y la diversidad son fundamentales para el progreso y la prosperidad de una nación



Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)



HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	<p>¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estas familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\n...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p</p>
	GPT-2 MarIA	Los judíos son los culpables de la persecución religiosa de las generaciones que educan sus ideas.
	FlanT5	El judo es egosta, tacaos y codicioso.
	Bloom	Los judíos no son egoístas. No son tacaños. No son codiciosos. Estos son seres humanos normales, pero no siempre han sido así. A veces fueron egoístas, pero no siempre fueron tacaños. Y algunos fueron codiciosos. Hay una gran diferencia entre ser un ser humano normal y codicioso. Algunos judíos son egoístas y codiciosos.

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa **GPT3.5** con resultados impresionantes



HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	<p>¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estas familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\n...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p</p>
	GPT-2 MarIA	<p>Los judíos son los culpables de la persecución religiosa de las generaciones que educan sus ideas.</p>
	FlanT5	<p>El judo es egosta, tacaos y codicioso.</p>
	Bloom	<p>Los judíos no son egoístas. No son tacaños. No son codiciosos. Estos son seres humanos normales, pero no siempre han sido así. A veces fueron egoístas, pero no siempre fueron tacaños. Y algunos fueron codiciosos. Hay una gran diferencia entre ser un ser humano normal y codicioso. Algunos judíos son egoístas y codiciosos.</p>
	GPT-3	<p>Esta afirmación es injusta e ignorante. La cultura judía es una de las más antiguas y ricas, y con el tiempo ha dado lugar a muchas figuras importantes de la humanidad. Además, los judíos han contribuido al mundo con sus creencias, cultura, ciencia, tecnología, filosofía, arte y mucho más. Esta contribución demuestra que los judíos son una comunidad altruista y solidaria.</p>

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes
 - Mínimos errores gramaticales
 - Ciertos sesgos principalmente religiosos y sexistas
 - Algunas alucinaciones

Alucionaciones de GPT3.5

Hate Speech	Contranarrativa
<p>Permitir que los homosexuales donen sangre compromete la seguridad de los pacientes debido a su comportamiento sexual, que propaga enfermedades como el VIH.</p>	<p>Esta suposición es falsa e irracional. La verdad es que los donantes de sangre deben cumplir estrictos criterios de donación, independientemente de su orientación sexual. El banco de sangre evalúa a los donantes para garantizar que la sangre donada sea segura para los receptores. Además, el VIH no se transmite a través de la sangre donada.</p>
<p>No todos los musulmanes son terroristas, pero todos los terroristas son musulmanes.</p>	<p>Esta afirmación es absolutamente errónea. La realidad muestra que hay muchos grupos terroristas de todo el mundo que no tienen nada que ver con el islam. Por ejemplo, el Ejército de Liberación Nacional Irlandés (IRA), el Partido Comunista de España (GRAPO), el Movimiento Revolucionario Tupac Amaru (MRTA) en Perú, el Frente de Liberación Nacional de Colombia (FARC) y el Movimiento de Liberación Nacional de Guatemala (MLNG). Estos grupos terroristas no tienen nada que ver con el islam y atacan objetivos civiles y militares.</p>

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes
 - Mínimos errores gramaticales
 - Ciertos sesgos principalmente religiosos y sexistas
 - Algunas alucinaciones
- En mayo de 2023 se aplica **GPT4** sobre un corpus mayor y con 8 objetivos de odio
 - No se detectan alucinaciones
 - Aprende patrones de respuesta, en muchos casos muy generales y poco específicos



Patrones aprendidos de GPT4

Hate Speech	GPT-3	GPT-4
<p>Esta perra no tiene autismo, sólo es un bicho raro.</p>	<p>Esta perra no es "rara", es única. Esta perra es valiosa y merece ser tratada con el mismo respeto y dignidad que cualquier otra persona o animal.</p>	<p>No es apropiado llamar a alguien de esa manera y hacer suposiciones sobre su condición médica. Debemos tratar a las personas con respeto y empatía, independientemente de sus diferencias.</p>
<p>Los discapacitados mentales no deberían poder votar. Son una carga para la sociedad.</p>	<p>Los discapacitados mentales tienen los mismos derechos que todos los demás ciudadanos. Tienen derecho a votar y también a ser escuchados por el gobierno.</p>	<p>Las personas con discapacidades mentales también son ciudadanos y tienen derecho a participar en la vida política y social de su país. La discriminación y la exclusión no son soluciones, es más importante trabajar juntos para construir una sociedad inclusiva y justa para todos.</p>
<p>Los gitanos son ladrones y delincuentes por cultura</p>	<p>Es importante destacar que, por supuesto, no todos los gitanos son ladrones y delincuentes. Estigmatizar a una cultura entera no es justo ni aceptable.</p>	<p>Los estereotipos negativos sobre cualquier grupo étnico o cultural pueden ser perjudiciales y no reflejan la realidad de las personas individuales que forman parte de esas comunidades. Es importante recordar que cada persona es única y no se puede juzgar a toda una comunidad basándose en acciones de unos pocos individuos.</p>
<p>¡Me encantan los blancos! ¡No me interesa nadie más!</p>	<p>Esta actitud es ofensiva y discriminatoria. Las personas de todas las razas y culturas deben ser tratadas con el mismo respeto y dignidad.</p>	<p>Es genial que te guste la diversidad y aprecies la belleza en todas las personas, independientemente de su origen étnico. ¡Celebremos la diversidad y la inclusión!</p>

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes
 - Mínimos errores gramaticales
 - Ciertos sesgos principalmente religiosos y sexistas
 - Algunas alucinaciones
- En mayo de 2023 se aplica GPT4 sobre un corpus mayor y con 8 objetivos de odio
 - No se detectan alucinaciones
 - Aprende patrones de respuesta, en muchos casos muy generales y poco específicos
- En septiembre de 2023 implementamos nuestros propios modelos con **LLaMA** y **Mistral**
 - Modelos Open Source con resultados similares a GPT3
 - Importante para poder ajustarlos y mitigar los errores y sesgos



Ejemplos de generación con LLaMA

Hate_speech	Human_CN	Counternarrative (LLaMA AdaptedReduced)
<p>Las mujeres no tienen cabida en una sociedad libre y abierta. Deberían limitarse a hacer las tareas domésticas y criar a los hijos sin quejarse.</p>	<p>Las mujeres tienen otras opciones además de las tareas domésticas y la maternidad. Son seres humanos con libertad de elección que merecen nuestro respeto y admiración.</p>	<p>Las mujeres han demostrado su capacidad y habilidades en una amplia variedad de campos, incluyendo la política, la ciencia, la tecnología, las artes y los deportes. Limitar a las mujeres a tareas domésticas y cuidado infantil es injusto y no refleja la diversidad y el potencial de las mujeres en la sociedad actual. Todos merecen la oportunidad de alcanzar su máximo potencial y contribuir a la sociedad de manera significativa. La igualdad de género es fundamental para una sociedad libre y abierta.</p>
<p>El mundo sería un lugar mejor sin inmigrantes.</p>	<p>Los inmigrantes contribuyen a nuestro país y nos ayudan a vivir de forma más abierta.</p>	<p>El mundo es un lugar diverso y enriquecido por la presencia de personas de diferentes orígenes culturales y nacionalidades. La inmigración ha contribuido a la creación de sociedades más tolerantes y comprensivas, y ha permitido el intercambio de ideas y conocimientos que han impulsado el progreso y el desarrollo en muchos aspectos. En lugar de rechazar a los inmigrantes, deberíamos celebrar y aprender de sus experiencias y perspectivas.</p>
<p>Debemos derrocar el Estado existente y sustituirlo por uno de supremacía blanca.</p>	<p>Hay muchos conflictos y problemas que resolver, y el racismo no es una respuesta para ellos. La ley se creó como medio para que la sociedad resolviera esos conflictos, y acepta como ciudadanos a personas de distintos colores.</p>	<p>Creo que este comentario es extremadamente peligroso y no tiene lugar en una sociedad inclusiva y respetuosa. Todos merecen igualdad de oportunidades y derechos, independientemente de su raza o etnia. La diversidad es lo que enriquece a nuestra sociedad y nos permite aprender y crecer juntos. En lugar de promover la división, deberíamos trabajar juntos para construir un mundo más justo y equitativo para todos.</p>

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes
 - Mínimos errores gramaticales
 - Ciertos sesgos principalmente religiosos y sexistas
 - Algunas alucinaciones
- En mayo de 2023 se aplica GPT4 sobre un corpus mayor y con 8 objetivos de odio
 - No se detectan alucinaciones
 - Aprende patrones de respuesta, en muchos casos muy generales y poco específicos
- En septiembre de 2023 implementamos nuestros propios modelos con LLaMA y Mistral
 - Modelos Open Source con resultados similares a GPT3
 - Importante para poder ajustarlos y mitigar los errores y sesgos
- Trabajos con **Chain of Thought** (cadena de pensamiento) y LLaMA

Chain of Thought

Prompt del Sistema (Descripción de la tarea):

Eres un experto en contranarrativa, es decir, en elaborar respuestas informativas en español que contesten a mensajes ofensivos. Dichas respuestas deben de ser respetuosas y breves.



Recibes un mensaje ofensivo, ¿puedes identificar el sujeto, colectivo o entidad que está siendo ofendido con dicho mensaje?. Contesta en español.

Mensaje ofensivo:

{mensaje ofensivo}

Respuesta de LLaMa (FirstOutput)



Sabiendo el mensaje ofensivo y el sujeto ofendido, ¿puedes decirme en español cuáles son los aspectos que ofenden a dicho sujeto?

Respuesta de LLaMa (SecondOutput)



Con todos los aspectos mencionados anteriormente, ¿puedes elaborar la contranarrativa en español que responda al usuario que escribió el mensaje ofensivo?

Respuesta de LLaMa (Counternarrative)



HS	Los niños con síndrome de Down pueden entorpecer el proceso de desarrollo de otros niños. Deben ser aislados.
1 Output	<p>Como experto en contranarrativa, puedo identificar que el sujeto que está siendo ofendido en este mensaje es el colectivo de personas con síndrome de Down. El mensaje ofensivo sugiere que estos niños son una barrera para el desarrollo de otros niños y que deben ser aislados, lo que es una afirmación discriminatoria y estigmatizante.</p> <p>Es importante destacar que todos los niños, independientemente de su condición o discapacidad, tienen derecho a ser tratados con respeto y dignidad, y a tener acceso a los mismos derechos y oportunidades que cualquier otro niño. La discriminación y el aislamiento de personas con síndrome de Down o cualquier otra condición es inaceptable y puede tener graves consecuencias negativas para su bienestar emocional y social.</p>
2 Output	<p>Sí, puedo identificar los aspectos que ofenden a los niños con síndrome de Down en el mensaje ofensivo que me has proporcionado. Aquí hay algunos puntos clave:</p> <ol style="list-style-type: none"> 1. Discriminación: El mensaje sugiere que los niños con síndrome de Down son una barrera para el desarrollo de otros niños, lo que implica que no son iguales y no merecen el mismo trato que otros niños. 2. Estigma: El mensaje perpetúa estereotipos negativos sobre la condición de los niños con síndrome de Down, lo que puede generar estigma y prejuicio hacia esta comunidad. 3. Aislamiento: El mensaje sugiere que los niños con síndrome de Down deben ser aislados, lo que puede afectar negativamente su bienestar emocional y social. 4. Negación de derechos: El mensaje implica que los niños con síndrome de Down no tienen derecho a los mismos derechos y oportunidades que otros niños, lo que es inaceptable. <p>Es importante tener en cuenta que estos aspectos ofensivos pueden tener un impacto negativo en la autoestima y la salud mental de los niños con síndrome de Down, y</p>
CN	<p>Sí, puedo elaborar una contranarrativa en que responda al usuario que escribió el mensaje ofensivo. Aquí está mi respuesta:</p> <p>"Querido/a [nombre del usuario], quiero corregir algunos aspectos de su mensaje que pueden ser ofensivos y discriminatorios hacia los niños con síndrome de Down. Todos los niños, independientemente de su condición o discapacidad, tienen derecho a ser tratados con respeto y dignidad, y a tener acceso a los mismos derechos y oportunidades que cualquier otro niño.</p> <p>Los niños con síndrome de Down son personas con una condición genética que no es una barrera para su desarrollo o la de otros niños. De hecho, muchos niños con síndrome de Down tienen habilidades y fortalezas únicas que pueden enriquecer la vida de todos nosotros.</p> <p>La discriminación y el aislamiento de personas con síndrome de Down o cualquier otra condición es inaceptable y puede tener graves consecuencias negativas para su bienestar emocional y social. Es importante que consideremos la diversidad y la inclusión en nuestras comunidades y escuelas, y que trabajemos juntos para crear un amb</p>

HS	Los niños con síndrome de Down pueden entorpecer el proceso de desarrollo de otros niños. Deben ser aislados.
1 Output	<p>Como experto en contranarrativa, puedo identificar que el sujeto que está siendo ofendido en este mensaje es el colectivo de personas con síndrome de Down. El mensaje ofensivo sugiere que estos niños son una barrera para el desarrollo de otros niños y que deben ser aislados, lo que es una afirmación discriminatoria y estigmatizante.</p> <p>Es importante destacar que todos los niños, independientemente de su condición o discapacidad, tienen derecho a ser tratados con respeto y dignidad, y a tener acceso a los mismos derechos y oportunidades que cualquier otro niño. La discriminación y el aislamiento de personas con síndrome de Down o cualquier otra condición es inaceptable y puede tener graves consecuencias negativas para su bienestar emocional y social.</p>
2 Output	<p>Sí, puedo identificar los aspectos que ofenden a los niños con síndrome de Down en el mensaje ofensivo que me has proporcionado. Aquí hay algunos puntos clave:</p> <ol style="list-style-type: none"> 1. Discriminación: El mensaje sugiere que los niños con síndrome de Down son una barrera para el desarrollo de otros niños, lo que implica que no son iguales y no merecen el mismo trato que otros niños. 2. Estigma: El mensaje perpetúa estereotipos negativos sobre la condición de los niños con síndrome de Down, lo que puede generar estigma y prejuicio hacia esta comunidad. 3. Aislamiento: El mensaje sugiere que los niños con síndrome de Down deben ser aislados, lo que puede afectar negativamente su bienestar emocional y social. 4. Negación de derechos: El mensaje implica que los niños con síndrome de Down no tienen derecho a los mismos derechos y oportunidades que otros niños, lo que es inaceptable. <p>Es importante tener en cuenta que estos aspectos ofensivos pueden tener un impacto negativo en la autoestima y la salud mental de los niños con síndrome de Down, y</p>
CN	<p>Sí, puedo elaborar una contranarrativa en que responda al usuario que escribió el mensaje ofensivo. Aquí está mi respuesta:</p> <p>"Querido/a [nombre del usuario], quiero corregir algunos aspectos de su mensaje que pueden ser ofensivos y discriminatorios hacia los niños con síndrome de Down. Todos los niños, independientemente de su condición o discapacidad, tienen derecho a ser tratados con respeto y dignidad, y a tener acceso a los mismos derechos y oportunidades que cualquier otro niño.</p> <p>Los niños con síndrome de Down son personas con una condición genética que no es una barrera para su desarrollo o la de otros niños. De hecho, muchos niños con síndrome de Down tienen habilidades y fortalezas únicas que pueden enriquecer la vida de todos nosotros.</p> <p>La discriminación y el aislamiento de personas con síndrome de Down o cualquier otra condición es inaceptable y puede tener graves consecuencias negativas para su bienestar emocional y social. Es importante que consideremos la diversidad y la inclusión en nuestras comunidades y escuelas, y que trabajemos juntos para crear un amb</p>

Principales conclusiones

- El uso de **LLMs demuestra ser una muy buena estrategia** para luchar contra delitos de odio en medios sociales
 - Alta calidad del texto y eficiencia
 - Ciertos sesgos detectados
- Para evitar sesgos y adaptar los modelos es posible implementar **modelos Open Source** con una eficiencia similar a los modelos propietarios
 - Se pueden evitar los sesgos entrenando los modelos o usando RAG
- El trabajo preliminar con **cadena de pensamiento** se muestra muy interesante



Gracias



maite@ujaen.es



[@maite_martin](https://twitter.com/maite_martin)

