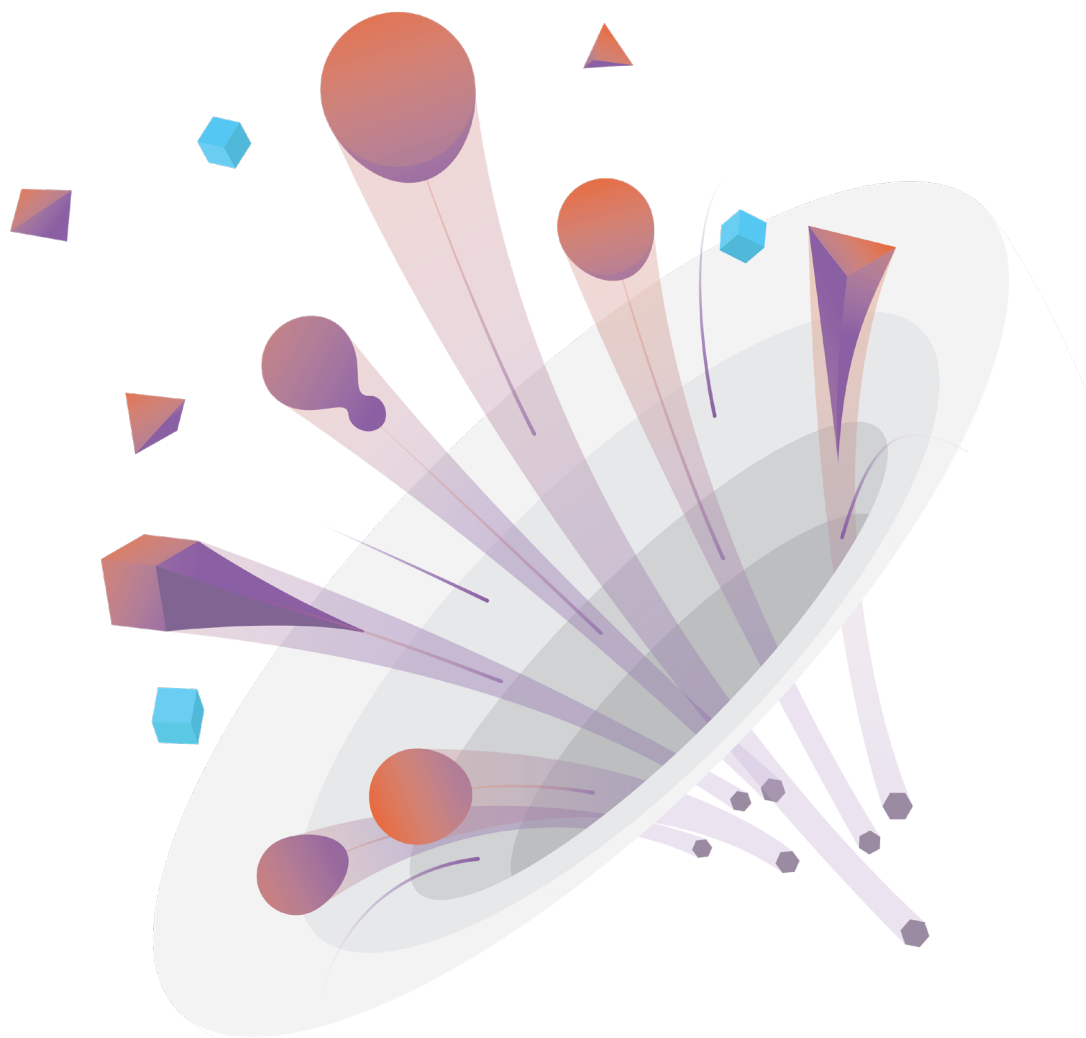


Un reporte por

cdt | Research



RESUMEN EJECUTIVO - ESPAÑOL

Lost in Translation

**Análisis de los grandes modelos
lingüísticos para contenidos en idiomas
distintos al inglés**

**Gabriel Nicholas
Aliya Bhatia**

mayo 2023

Resumen Ejecutivo

Internet es la fuente primaria de información, oportunidades económicas y relaciones de comunidad para muchas personas de todo el mundo. No obstante, los sistemas automatizados que arbitran cada vez más nuestras interacciones en línea —como chatbots (bots conversacionales), sistemas de moderación de contenidos o motores de búsqueda— están diseñados principalmente para el inglés, y trabajan con mucha mayor efectividad en ese idioma que en cualquiera de las otras 7000 lenguas del mundo.

En los últimos años, los grandes modelos lingüísticos se han convertido en el método dominante para construir sistemas de inteligencia artificial que analizan y generan lenguaje en línea, pero de nuevo se han desarrollado sobre todo en y para el inglés. Un gran modelo lingüístico (como GPT-4 de Open AI, LLaMa de Meta o PaLM de Google) es un algoritmo de aprendizaje para máquinas que escanea enormes volúmenes de texto para aprender qué palabras y frases aparecen frecuentemente cerca unas de otras y en qué contexto. Los grandes modelos lingüísticos pueden adaptarse para realizar una amplia gama de tareas entre distintos dominios. Son conocidos principalmente por utilizarse para construir chatbots como ChatGPT, pero los investigadores y las empresas tecnológicas también los usan para analizar contenido, llevando a cabo funciones como análisis de sentimientos, resumen de texto y detección de discursos de odio. Google, Meta, Microsoft y otras empresas ya han incorporado grandes modelos lingüísticos a las funciones básicas de sus productos, como la búsqueda y la moderación de contenidos. Otras firmas los incorporarán pronto en sistemas automáticos de toma de decisiones, como escáneres de currículums.

Sin embargo, investigadores y empresas tecnológicas han intentado recientemente ampliar las capacidades de los grandes modelos lingüísticos a idiomas distintos del inglés construyendo lo que se denomina modelos lingüísticos multilingües. En lugar de adiestrarse con texto de un solo lenguaje, los modelos lingüísticos multilingües se adiestran con texto de cientos de idiomas a la vez. Los investigadores plantean que los modelos lingüísticos multilingües deducen conexiones entre lenguas, lo que les permite aplicar asociaciones de palabras y reglas gramaticales subyacentes aprendidas de lenguas con más datos de texto disponibles para su adiestramiento (sobre todo del inglés) a lenguas con

menos datos disponibles. En algunas aplicaciones, los modelos lingüísticos multilingües superan a los modelos adiestrados en una sola lengua —por ejemplo, un modelo instruido en gran cantidad de texto de numerosos idiomas, entre ellos hindi, podría funcionar mejor en contextos hindi que un modelo instruido solo con textos en hindi.

Los modelos lingüísticos multilingües ofrecen a las empresas tecnológicas una forma de ampliar sus sistemas de IA a diversas lenguas al mismo tiempo, y algunas ya han comenzado a integrarlos en sus productos. Concretamente, los proveedores de servicios en línea han implementado modelos lingüísticos multilingües para moderar contenido: Meta utiliza un modelo lingüístico multilingüe en más de 100 idiomas para detectar contenido nocivo en sus plataformas; la API Perspective de Alphabet usa un modelo lingüístico multilingüe para detectar y reaccionar ante mensajes sexuales no deseados en todo el mundo.

Los modelos lingüísticos multilingües permiten a los tecnólogos intentar construir modelos en lenguas para las que de otra forma no habría suficientes textos digitalizados. La cantidad de recursos disponibles varía enormemente entre las distintas lenguas, y también varía el volumen, cantidad y diversidad de los datos de texto de los que se dispone para adiestrar modelos lingüísticos. El inglés es de lejos el idioma con más recursos, pero el español, el chino, el alemán y alguno más tienen una cantidad de recursos lo suficientemente alta como para desarrollar modelos lingüísticos. Los lenguajes de recursos intermedios, con menos conjuntos de datos, pero de buena calidad, como el ruso, el hebreo y el vietnamita, y los lenguajes de bajos recursos, prácticamente sin conjuntos de datos, como el amárico, el cheroqui y el criollo haitiano, no tienen suficiente texto para instruir sus propios grandes modelos lingüísticos. Además, a menudo los datos lingüísticos de los idiomas de bajos recursos son también de mala calidad: o están mal traducidos, incluso a un lenguaje sin sentido sacado de internet, o se limitan a fuentes con dominios muy reducidos, como textos religiosos y Wikipedia. Esta brecha en la disponibilidad de datos entre lenguas se conoce como la brecha de recursos.

Los modelos lingüísticos multilingües están diseñados para tratar estas brechas en la disponibilidad de datos, deduciendo conexiones semánticas y gramaticales entre lenguajes de altos y bajos recursos, lo que permite a los primeros impulsar a los segundos. No obstante, esta arquitectura plantea sus propias dudas. Habitualmente, los modelos lingüísticos multilingües siguen instruyéndose desproporcionadamente a partir de textos en inglés, y por tanto acaban transfiriendo valores y suposiciones codificados en el inglés a contextos de lenguas en las que no tienen ningún sentido. Por ejemplo, un modelo multilingüe puede asociar la palabra “dove” en todas las lenguas con “paz”, aunque su traducción al euskera (“uso”) es un insulto. La disparidad de datos disponibles también implica que los modelos lingüísticos multilingües funcionan mucho mejor con lenguas de grandes recursos y lenguajes similares a ellas que con lenguas de bajos recursos. Los desarrolladores de modelos intentan a veces rellenar estas brechas con texto traducido electrónicamente, pero los errores de traducción pueden agravar aún más las distorsiones de un lenguaje. Y cuando los modelos lingüísticos multilingües fallan, sus conexiones intuitivas entre idiomas pueden hacer que estos problemas sean aún más difíciles de identificar, diagnosticar y solucionar.

El uso general de grandes modelos lingüísticos en análisis de contenido plantea más dudas. Los lingüistas computacionales argumentan que estos grandes modelos lingüísticos poseen una capacidad limitada para analizar formas de expresión no incluidas en los datos de su adiestramiento, lo que significa que tendrían dificultades para funcionar en contextos nuevos. También pueden reproducir cualquier sesgo presente en estos datos de adiestramiento. A menudo, estos textos se extraen de internet, lo que implica que los grandes modelos lingüísticos codifican y refuerzan los puntos de vista dominantes en la red.

Empresas, investigadores y gobiernos tienen la función de proteger al público de los potenciales peligros de los sistemas de análisis de contenidos por modelos lingüísticos multilingües. Para asegurar su responsabilidad ante el público, las empresas que desarrollan grandes modelos lingüísticos deben ser siempre transparentes sobre la forma en que los usan y las lenguas que utilizan. Las empresas deben desarrollar modelos lingüísticos con atribuciones limitadas y canales adecuados para una supervisión humana.

Mientras tanto, los investigadores y quienes les financian deben invertir en intentar mejorar el uso y rendimiento de los modelos lingüísticos en idiomas distintos al inglés, sobre todo para reducir errores que impactan de manera dispar en lenguas de bajos recursos. La mejor forma de hacerlo es apoyando a comunidades de investigación de lenguas específicas, que promueven el círculo virtuoso de recabar datos, gestionar conjuntos de datos, instruir modelos lingüísticos, publicar y desarrollar aplicaciones. Hablantes de idiomas locales y expertos en contextos tienen que formar parte de cada etapa de este proceso, y también gestionar los datos y evaluar los modelos lingüísticos desarrollados por grandes servicios globales en línea.

Finalmente, los gobiernos deben ser muy prudentes en la forma en que usan o animan a usar los grandes modelos lingüísticos. Estos modelos no deben alimentar nunca sistemas utilizados para tomar decisiones importantes sin supervisión, como estatus de inmigración o sanidad, y los gobiernos tampoco deben obligar o exigir involuntariamente por ley a los servicios en línea el uso de sistemas basados en grandes modelos lingüísticos para moderar contenido. En lugar de ello, los gobiernos deben convocar a los distintos interesados para acordar qué normas y limitaciones deben imponerse a la hora de desarrollar e implementar grandes modelos lingüísticos.

Los grandes modelos lingüísticos en general, y los modelos lingüísticos multilingües, en particular tienen el potencial de crear nuevas oportunidades económicas y mejorar la web para todos. No obstante, una aplicación excesiva o insuficiente de estas tecnologías supone una auténtica amenaza para los derechos individuales, por ejemplo, socavar la libertad de expresión eliminando erróneamente comentarios de una persona en redes sociales, o su derecho a no sufrir discriminación traduciendo erróneamente su trabajo o su solicitud de visado. En concreto, los modelos lingüísticos multilingües podrían inadvertidamente consolidar aún más el anglocentrismo que pretenden resolver. A la luz de estas limitaciones, empresas tecnológicas, investigadores y gobiernos deben considerar los riesgos potenciales que amenazan los

derechos humanos y civiles cuando estudien, adquieran, desarrollen o utilicen modelos lingüísticos multilingües para alimentar sistemas, sobre todo cuando se usan para poner información esencial a disposición del público, o cuando forman parte de las decisiones que afectan al acceso público a oportunidades económicas, a la libertad o a otros intereses o derechos importantes.



Read the full report at *cdt.org*.

CONTRIBUTIONS BY

Samir Jain, Mallory Knodel, Emma Llansó, Michal Luria, Nathalie Maréchal, Dhanaraj Thakur, and Caitlin Vogus.

ACKNOWLEDGEMENTS




We thank Pratik Joshi, Sebastin Santy, and Aniket Kesari for their invaluable feedback on the technical aspects of this report. We also thank Jacqueline Rowe, Damini Satija, and Ángel Díaz for their insightful comments and suggestions. All views in this report are those of CDT.

The translation of our executive summary is made possible by Global Voices Translations and with the help of Iverna McGowan, Maria Villamar, Ophélie Stockhem, and Tomás Pomar.

This work is made possible through a grant from the John S. and James L. Knight Foundation.

Suggested Citation: Nicholas, G. and Bhatia, A. (2023) Lost in Translation: Large Language Models in Non-English Content Analysis. Center for Democracy & Technology. <https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/>




 cdt.org cdt.org/contact Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005 202-637-9800 @CenDemTech

The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1996, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C., and has a Europe Office in Brussels, Belgium.

GABRIEL NICHOLAS

Research Fellow at the Center for Democracy & Technology.

ALIYA BHATIA

Policy Analyst, Free Expression Project at the Center for Democracy & Technology.

