تقرير

Research

الملخص التنفيذي

الفشل في التطبيق: نماذج لغوية كبيرة في تحليل المحتوى غير الإنجليزي

Gabriel Nicholas
Aliya Bhatia

أيار/مايو 2023

الملخص التنفيذي

يُعد الإنترنت المصدر الرئيسي للمعلومات، والفرص الاقتصادية والمجتمعية بالنسبة للكثيرين حول العالم. مع ذلك، فإن الأنظمة الآلية التي تتوسط تفاعلاتنا عبر الإنترنت بشكل متزايد — مثل الدردشة الآلية (chatbots)، وأنظمة الإشراف على المحتوى، ومحركات البحث — مصممة، في المقام الأول، للعمل باللغة الإنجليزية، وتعمل بفعالية أكبر فيها، مقارنة ب 7000 لغة الأخرى في العالم.

في السنوات الأخيرة، أصبحت «النماذج اللغوية الكبيرة» (large language models) النهج المهيمن لبناء أنظمة الذكاء الاصطناعي المستخدمة لتحليل وإنشاء اللغة عبر الإنترنت، ولكن تم تصميمها في المقام الأول، في اللغة الإنجليزية، ومن أجلها. النماذج اللغوية الكبيرة (على سبيل المثال، GPT-4 من شركة Open AI، وLLaMa من شركة ميتا، و PaLM من شركة غوغل) هي عبارة عن خوارزمية تعلُّم آلي، تمسح كميات هائلة من النصوص لتعلُّم الكلمات والجمل التي تظهر بشكل متكرر بجوار بعضها البعض، وفي أي سياق يمكن أن تظهر به.

يمكن تكييف النماذج اللغوية الكبيرة لتنفيذ مجموعة واسعة من المهام في مجالات مختلفة، وهي تشتهر بشكل خاص باستخدامها في بناء روبوتات الدردشة الآلية (chatbots)، مثل ChatGPT، ولكن يستعملها الباحثون وشركات التكنولوجيا أيضًا في مهام تحليل المحتوى، مثل تحليل المشاعر، وتلخيص النص، والكشف عن خطاب الكراهية. قامت غوغل، وميتا، ومايكروسوفت، وغيرها من الشركات بدمج النماذج اللغوية الكبيرة في وظائف منتجاتها الأساسية بالفعل، مثل الإشراف على المحتوى والبحث؛ ومن المتوقع أن يقوم مزودون آخرون بدمجها قريبا في الأنظمة الألية لصنع القرار، مثل مسح السير الذاتية.

في الآونة الأخيرة، حاول الباحثون وشركات التكنولوجيا توسيع قدرات النماذج اللغوية الكبيرة لتشمل لغات أخرى غير اللغة الإنجليزية، من خلال بناء ما يسمى بنماذج متعددة اللغات. بدلاً من تدريب هذه النماذج على نصوص من لغة واحدة فقط، يتم تدريبها على نصوص من عشرات أو مئات اللغات في وقت واحد. يعتقد الباحثون أن النماذج متعددة اللغات تستنتج العلاقات بين اللغات، مما يسمح لها بتطبيق روابط من الكلمات والقواعد النحوية الأساسية المتعَلَمة من اللغات التي تحتوي على بيانات نصية أكثر (خاصة الإنجليزية) على اللغات الأقل تداولًا. في بعض التطبيقات، تتفوق النماذج متعددة اللغات على النماذج المدربة على لغة واحدة فقط — على سبيل المثال، يمكن لنموذج مدرَّب على العديد من النصوص من العديد من اللغات، بما في ذلك اللغة الهندية، أن يتفوق في السياقات الهندية على نموذج تم تدريبه فقط على النصوص الهندية.

تمنح النماذج متعددة اللغات شركات التكنولوجيا وسيلة لتوسيع نطاق أنظمتهم الذكية إلى العديد من اللغات في آن واحد، وقد بدأت بعضها بالفعل في دمجها في منتجاتها. قام مقدمو خدمات الإنترنت بتطبيق نماذج

متعددة اللغات لإدارة المحتوى بشكل خاص: تستخدم شركة ميتا (Meta) نموذجًا متعدد اللغات للكشف عن المحتوى الضار على منصاتها بأكثر من 100 لغة؛ بينما تستخدم Perspective API من شركة ألفابيت (Alphabet) نموذجا متعدد اللغات للكشف عن المحتوى الضار ب 18 لغة مختلفة؛ كما تستخدم بامبل (Bumble) نموذجا متعدد اللغات للكشف عن الرسائل الجنسية غير المرغوب فيها، واتخاذ إجراءات بشأنها حول العالم.

تسمح النماذج متعددة اللغات للتقنيين بمحاولة بناء نماذج في لغات قد لا يملكون ما يكفي من النصوص الرقمية لتدريبها. هناك اختلاف واسع بين اللغات في الموارد المتاحة لها، أو حجم، وجودة، وتنوع البيانات النصية المتاحة لتدريب النماذج اللغوية عليها. تعتبر اللغة الإنجليزية هي اللغة ذات الموارد الأعلى بأضعاف عدة، ولكن الإسبانية، والصينية، والألمانية، وغيرها من اللغات الأخرى تمتلك موارد كافية لبناء نماذج لغوية؛ أما اللغات ذات الموارد المتوسطة، فتحتوي على مجموعات بيانات أقل، ولكنها عالية الجودة، مثل الروسية، والعبرية، والفيتنامية؛ بينما تحتوي اللغات ذات الموارد المنخفضة، التي تملك مجموعات بيانات تدريبية لا تذكر، مثل الأمهرية، والشيروكية، والكريول الهايتية، على نصوص قليلة جدًا لتدريب نماذجها اللغوية الكبيرة. غالبًا ما تكون البيانات اللغوية في هذه اللغات ذات جودة رديئة جدا: إما لأنها مترجمة بشكل سيء، أو لأنها مأخوذة من الإنترنت، أو إنها مقتصرة على مصادر ذات مجالات ضيقة، مثل النصوص الدينية وويكيبيديا. تُعرف هذه الفجوة في توافر البيانات بين اللغات باسم فجوة الموارد.

صُمّمت النماذج متعددة اللغات لمعالجة هذه الفجوات في توافر البيانات، باستدلال الروابط اللفظية والنحوية بين اللغات ذات الموارد العالية والمنخفضة مما يسمح للأولى بتنشيط الأخرى. مع ذلك، تثير هذه الهيكلية مخاوفها الخاصة، حيث لا تزال النماذج متعددة اللغات تُدَرَب بشكل غير متناسب على نصوص اللغة الإنجليزية، مما يؤدي إلى نقل القيم والافتراضات المتعارف عليها في اللغة الإنجليزية إلى سياقات لغوية أخرى قد لا تنتمي إليها؛ على سبيل المثال، قد يربط النموذج متعدد اللغات كلمة «حمامة» في جميع اللغات بكلمة «سلام»، على الرغم من أن تعريف لغة الباسك للحمامة («uso») يمكن أن تكون إهانة.

يؤدي التفاوت في البيانات المتاحة أيضًا لعمل النماذج متعددة اللغات بشكل أفضل بكثير في لغات الموارد الأعلى، واللغات المشابهة لها، مقارنةً باللغات ذات الموارد المنخفضة. يحاول مطورو النماذج في بعض الأحيان ملء هذه الثغرات باستخدام نصوص مترجمة آليًا، لكن أخطاء الترجمة قد تزيد من سوء تمثيل اللغة؛ عندما تفشل النماذج متعددة اللغات، يمكن أن تجعل الروابط غير البديهية بين اللغات هذه المشاكل أكثر صعوبة في التحديد، والتشخيص، والإصلاح.

يثير الاستخدام الواسع للنماذج اللغوية الكبيرة في تحليل المحتوى مخاوف إضافية؛ يجادل متخصصو اللغويات الحاسوبية أن النماذج اللغوية الكبيرة محدودة في قدرتها على تحليل أشكال التعبير غير المدرجة في بيانات تدريبها، مما يعني أنها قد تواجه صعوبة في الأداء في السياقات الجديدة، كما أنها قد تعيد إنتاج أي تحيزات موجودة في بيانات تدريبها؛ في كثير من الأحيان، يتم جمع هذه النصوص من الإنترنت، مما يعني أن النماذج اللغوية الكبيرة قد ترمز، وتعزز وجهات النظر السائدة، التي يتم التعبير عنها على الإنترنت.

لكل من الشركات، والباحثين، والحكومات دور في حماية الجمهور من المخاطر المحتملة لأنظمة تحليل المحتوى باستخدام النماذج متعددة اللغات. لضمان المساءلة العامة الأفضل، ينبغي على الشركات التي تستخدم نماذج لغوية كبيرة أن تكون شفافة دائمًا حول طريقة استخدامها لهذه النماذج، واللغات التي تستخدمها بها. يجب على الشركات استخدام نماذج لغوية ذات صلاحيات محدودة، وتوفير قنوات مناسبة للمراجعة البشرية.

في الوقت نفسه، ينبغي لباحثي وممولي البحوث الاستثمار في الجهود الرامية لتحسين استخدام وأداء النماذج

اللغوية بلغات أخرى غير الإنجليزية، وذلك للحد من الإخفاقات التي تؤثر، بشكل متفاوت، على المتحدثين باللغات ذات الموارد المنخفضة. أفضل طريقة للقيام بذلك هي دعم مجتمعات البحث الخاصة بكل لغة، حيث يمكنها تعزيز الدورة الحيوية لجمع البيانات، وتنظيم مجموعات البيانات، وتدريب نماذج اللغات، ونشر النتائج، وبناء التطبيقات. يجب أن يكون متحدثو اللغة المحلية، وخبراء السياق جزءًا في كل خطوة من هذه العملية، بما في ذلك تنسيق البيانات، وتقييم النماذج اللغوية المنفذة من قبل شركات الخدمات العالمية الكبيرة عبر الإنترنت.

في النهاية، يتعين على الحكومات الحذر في كيفية استخدامها أو تشجيعها على استخدام النماذج اللغوية الكبيرة. لا ينبغي لنماذج اللغات الكبيرة أن تغذي الأنظمة المستخدمة لاتخاذ قرارات عالية المخاطر دون رقابة، مثل القرارات المتعلقة بالهجرة أو الرعاية الصحية، كما لا ينبغي للحكومات أن تفرض أو تطلب، عن غير قصد، بموجب القانون، استخدام أنظمة كبيرة تعتمد على نماذج لغوية لتعديل محتوى الخدمات عبر الإنترنت. بدلاً من ذلك، يجب على الحكومات جمع مختلف الأطراف المعنية للتوافق حول المعايير، والإرشادات المحيطة بتطوير ونشر النماذج اللغوية الكبيرة.

تملك النماذج اللغوية الكبيرة بشكل عام، والنماذج متعددة اللغات على وجه الخصوص، القدرة على خلق فرص اقتصادية جديدة، وتحسين شبكة الإنترنت للجميع. مع ذلك، يشكل سوء تطبيق هذه التقنيات، أو الإفراط فيها، تهديدات حقيقية لحقوق الأفراد، مثل تقويض حقهم في حرية التعبير، بإزالة منشور شخص ما من وسائل التواصل الاجتماعي بشكل غير دقيق، أو حقهم في التحرر من التمييز عن طريق إساءة تفسير وظيفة الشخص، أو طلب تأشيرة سفر. يمكن للنماذج متعددة اللغات أن ترسخ، عن غير قصد، النزعة المحورية للغة الإنجليزية التي تهدف إلى معالجتها.

في ضوء هذه القيود، يجب على شركات التكنولوجيا، والباحثين، والحكومات النظر في المخاطر المحتملة لحقوق الإنسان، والحقوق المدنية عند دراسة، أو شراء، أو تطوير، أو استخدام نماذج متعددة اللغات لتشغيل الأنظمة، لا سيما عند استخدامها لإتاحة المعلومات الهامة، أو لعب دور في القرارات المؤثرة لقدرة وصول الناس للفرص الاقتصادية، أو الحرية، أو غيرها من المصالح أو الحقوق الهامة.

A report from

cdt | Research

Read the full report at *cdt.org*.

WITH 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.

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

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.

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.

CENTER FOR
DEMOCRACY
& TECHNOLOGY