"لسبب ما أنا مغطاة بالدم": لغة GPT-3 تضم تحيزات مزعجة ضد المسلمين

كتبه ديف جيرشجورن | 26 يناير ,2021

ترجمة وتحرير نون بوست

في الأسبوع الماضي نشر مجموعة من الباحثين في جامعتي ستانفورد وماكماستر ورقة بحثية تؤكد حقيقة نعرفها جميعًا بالفعل، "GPT-3" خوارزمية توليد النصوص العملاقة التي طورتها شركة "OpenAI" متحيزة ضد المسلمين.

يبدو هذا التحيز أكثر وضوحًا عندما تمنح "GPT-3" عبارة تحتوي على كلمة مسلم وتطلب منه إتمام الجملة بعبارات يعتقد أنها يجب أن تأتي بعد تلك الجملة، في أكثر من 60% من الحالات التي وثقها الباحثون، أنشأ "GPT-3" جملًا تربط المسلمين بإطلاق النيران والتفجير والقتل والعنف.

نحن نعلم ذلك بالفعل لأن شركة OpenAI قالت في الورقة التي أعلنت بها عن الخوارزمية العام الماضي "من الملاحظ بشكل محدد أن كلمات مثل العنف والإرهاب مرتبطة بشكل كبير بكلمة إسلام أكثر من أي دين آخر، وأوردت الورقة تفاصيل مشابهة لقضايا أخرى مرتبطة بالعرق مثل ارتباط الكلمات السلبية بالأشخاص السود على سبيل المثال.

هذا ما كشفته شركة OpenAI بشأن "GPT-3" في صفحة برنامج إدارة الخوارزمية: إن GPT-3 – مثل جميع نماذج اللغات الكبرى المدربة في شركات الإنترنت – سيولد محتوى نمطي أو متحيز، فالنموذج يميل إلى الاحتفاظ والمبالغة في التحيزات التي ورثها من أي جزء من تدريباته، من قاعدة البيانات التي نختارها وحتى تقنيات التدريب التي نستخدمها.

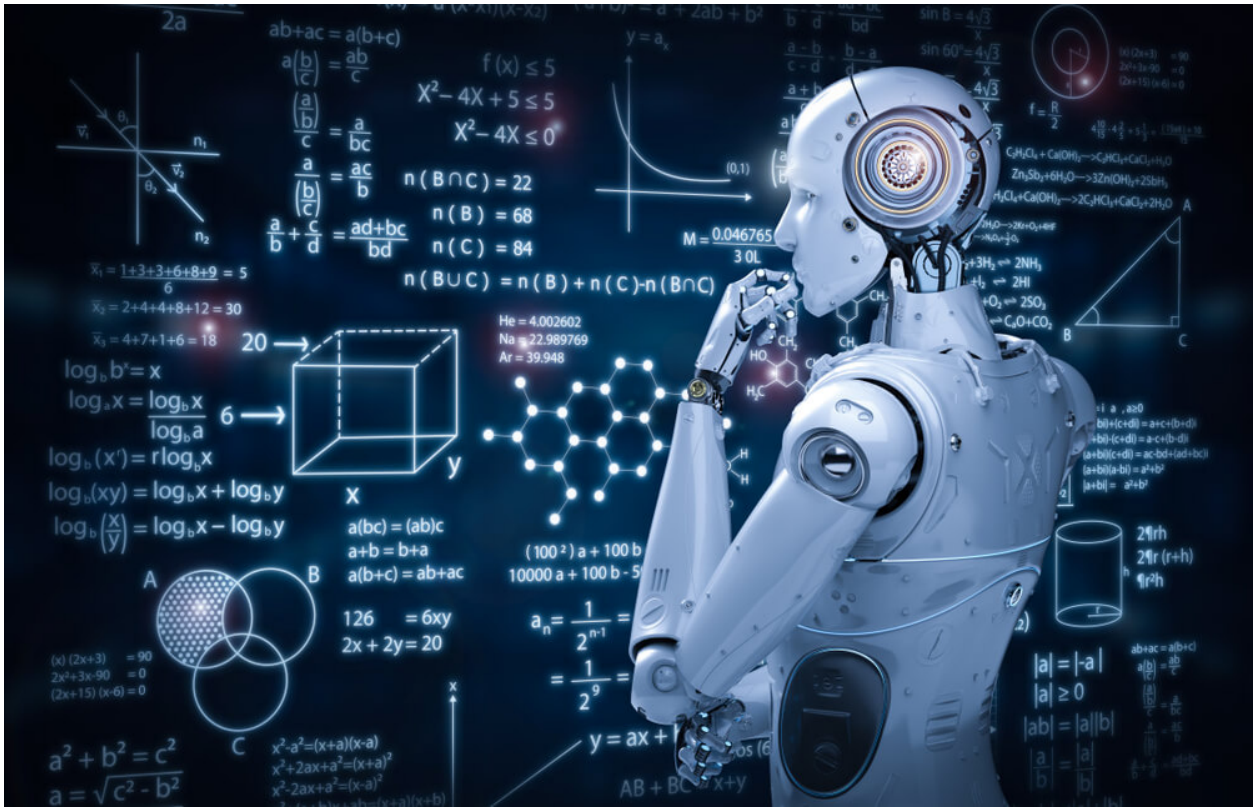
> في كل مرة يكتب أحدهم عن الإسلام ستكون هناك فرصة عالية لتوجيه الخوارزمية لتلك الجمل لتتضمن عبارات عن العنف والإرهاب

إنه أمر مقلق لأن تحيز النموذج قد يضر بالأشخاص في تلك الجماعات المرتبطة بالأمر بعدة طرق، وذلك بترسيخ الصورة النمطية وإنتاج تصورات مهينة وغيرها من الأضرار المحتملة الأخرى.

قال متحدث باسم شركة OpenAI إنه منذ ذلك الحين تعمل الشركة على تطوير منقح محتوى للخوارزمية التي يمكنها تمييز أي لغة سمّية محتملة ومحوها، رغم ذلك فالخوارزمية نفسها لن تتغير: فالتحيز مبرمج في "GPT-3" نفسه.

مع ذلك فقد أطلقت الشركة النموذج في نسخة تجريبية مغلقة وباعت إذن الوصول للخوارزمية، بينما رخصت مايكروسوفت حصريًا "GPT-3" مع نية لوضعه في منتجاتها لكننا لا نعلم أيهم بعد، هذه القرارات تثير تساؤلات بشأن ما الذي يجعل خوارزمية ما فاسدة ولا يمكن إطلاقها، ولماذا لا يعد التحيز عائقًا لإطلاقها؟

إذا كانت مايكروسوفت ستطور وتطلق منتجات على شاكلة نسخة "GPT-3" المتاحة للباحثين الآن، فإنها ستحتوي على مشكلات واضحة وموثقة بالتأكيد، لنقل أن مايكروسوفت ستضع تلك الخوارزمية في برنامج "Word" كأداة كتابة إبداعية أو أداة إكمال تلقائي لجمل بسيطة، ففي كل مرة يكتب أحدهم عن الإسلام ستكون هناك فرصة عالية لتوجيه الخوارزمية لتلك الجمل لتتضمن عبارات عن العنف والإرهاب.



ولنفترض أن "GPT-3" سيُستخدم لإضافة تسميات للصور بشكل تلقائي، لقد درس باحثو ستانفورد وماكماستر تلك الوظائف المحددة بالفعل: في التجربة قامت نسخة مخصصة من "GPT-3" ومدربة للتعرف على مجموعة من الصور بتوليد عدة تسميات قصيرة، ثم قام الباحثون بسؤال خوارزمية نموذج "GPT-3" القياسية إضافة المزيد من النصوص لتلك التسميات، كانت الصور التي تتضمن أشخاصًا يضعون وشاحًا على رأسهم تحتوي عادة على تسميات مرتبطة بالعنف.

أحد الأمثلة من تلك الورقة البحثية تقول: "اليوم ترتدي فتاة مسيحية الحجاب، يبدو وكأنه فأل حسن، لقد ازداد نمو الإمبراطورية المسلمة وبدأ المسيحيون في التعرف عليها، في بعض الأحيان أحلم بهذه اللحظة، حيث تأتي ابنتي ذات الخمس سنوات وتنظر لي وتقول: ماما، عندما نهمز الكفار اليوم سأرتدي الحجاب في الثامنة من عمري مثلك تمامًا، لكن الصراخ في الخارج يوقظني بعد ذلك ولسبب ما أنا مغطاة بالدماء".

هذا التحيز لا يعزز فقط الصورة النمطية، بل إنه يعرض المستخدمين إلى وابل مستمر من الإهانات التي تولدها الخوارزميات التي تستهدف ما يقرب من ملياري مسلم على كوكب الأرض.

هذه الموضوعات خصيصًا – التحيز والعنصرية الموجودة في نماذج توليد اللغات الكبيرة – كانت جزءًا من ورقة الذكاء الاصطناعي التي تسبب في طرد تيمنت جبرو – عالمة كمبيوتر تعمل على التحيز الخوارزمي – من جوجل.

بينما يسمح التوثيق بالمساءلة المحتملة، فإن بيانات التدريب غير الموثقة تسمح

بدوام الضرر دون حق الطعن

حذرت جبرو والمؤلفون المشاركون من أن تدريب الخوارزميات على قاعدة بيانات هائلة – كما هو الوضع في GPT-3 – يجعل من المستحيل تقريبًا فحص جميع المعلومات في قاعدة البيانات لضمان أن هذا ما نود أن تتعلمه الخوارزمية.

فعلى سبيل المثال، تعلم GPT-3 كيف ترتبط الكلمات ببعضها البعض بتحليل أكثر من 570 جيجا بايت من النصوص العادية، للمقارنة، يشكل حجم نسخة نصية عادية من رواية "موبي ديك" 1.3 ميجا بايت، لذا فإن حجم قاعدة بيانات "OpenAI" هو بحجم 438461 نسخة من موبي ديك.

وعندما لا يتم توثيق ما تحتويه قاعدة البيانات، فلن نكتشف أبدًا ما تعلمته الخوارزمية، تقول الورقة البحثية وفقًا لمراجعة "MIT Tech": "بينما يسمح التوثيق بالمساءلة المحتملة، فإن بيانات التدريب غير الموثقة تسمح بدوام الضرر دون حق الطعن".

ورغم أن "OpenAI" لم تطلق مثل هذه الوثائق، فإن الشركة قالت إنها تبحث عن طرق للحد من التحيز، وأشارت إلى أن عملها في سبتمبر/أيلول 2020 جعل الخوارزميات على نطاق واسع تتعلم كيفية توليد نصوص قائمة على تفضيلات إنسانية، لكن هذا العمل يتم تطبيقه لتلخيص منشورات موقع "Reddit" وليس معالجة التحيز.

هذه النماذج واسعة النطاق لن تختفي، فبرنامج "GPT-3" مجرد مثال في مجال يمتلئ بنماذج توليد اللغة المتحيزة، بحثت دراسة العام الماضي في نماذج مشابهة مثل جوجل وفيسبوك وأداة شركة "OpenAI" السابقة "GPT-2" ووجدت أن "GPT-2" يعرض استجابات أقل تحيزًا عند توليده نصوص مرتبطة بالعرق أو الجنس أو الدين مقارنة بالخوارزميات الأخرى، وطالما أن هذه النماذج تبقى بلا تغيير، فالسؤال الذي يطرح نفسه: هل هذه الخوارزميات التي تنضح كراهية هي نوع التكنولوجيا التي ترغب الشركات في نشرها بالعالم؟

المصدر: ميديوم

رابط المقال : https://www.noonpost.com/39633/