كيف يُساعدنا الذكاء الاصطناعي على كشف حملات التضليل؟

ترجمة وتحرير: نون بوست

نادرًا ما تُحدد الحقائق المجردة والقاسية اهتمامات الأفراد واعتقاداتهم. بل إن القوة والألفة التي تحملهما القصة المحكية جيدًا هي التي تسود. وسواءً أكانت حكاية مؤثرة، أو شهادة شخصية، أو "ميم" يُحاكي سرديات ثقافية مألوفة، فإن القصص تميل إلى أن تبقى في أذهاننا، وتُحرك مشاعرنا، وتُشكل معتقداتنا.

هذه السمة المميزة لسرد القصص هي تحديدًا ما يجعلها بالغة الخطورة عند استخدامها من قِبَل جهات غير مُختصة. وقد استخدم الخصوم الأجانب لعقود أساليب سردية في محاولاتهم للتلاعب بالرأي العام في الولايات المتحدة. وجلبت منصات التواصل الاجتماعي تعقيدًا وتحديًا جديدين لهذه الحملات؛ حيث حظيت هذه الظاهرة باهتمام عام واسع بعد ظهور أدلة على تأثير كيانات روسية على مواد متعلقة بالانتخابات على فيسبوك في الفترة التي سبقت انتخابات عام 2016.

وبينما يعد الذكاء الاصطناعي أحد الأسباب التي تُفاقم المشكلة، فإنه في الوقت نفسه يُصبح أحد أقوى أدوات الدفاع ضد مثل هذه التلاعبات. وقد استخدم الباحثون تقنيات التعلم الآلي لتحليل محتوى المعلومات المُضللة. نقوم في مختبر الإدراك والسردية والثقافة في جامعة فلوريدا الدولية ببناء أدوات ذكاء اصطناعي للمساعدة في الكشف عن حملات التضليل التي تستخدم أدوات الإقناع السردي حيث ندرب الذكاء الاصطناعي على تجاوز التحليل اللغوي على المستوى السطحي لفهم هياكل السرد، وتتبع الشخصيات والجداول الزمنية، وفكّ رموز المراجع الثقافية.

المعلومات المضللة مقابل المعلومات المغلوطة

في يوليو/ تموز 2024، أحبطت وزارة العدل عملية مدعومة من الكرملين استخدمت ما يقارب ألف حساب مزيف على وسائل التواصل الاجتماعي لنشر روايات كاذبة. لم تكن هذه حوادث معزولة، بل كانت جزءًا من حملة منظمة مدعومة جزئيًا بالذكاء الاصطناعي.

تختلف المعلومات المضللة بشكل حاسم عن المعلومات المغلوطة، ففي حين أن المعلومات المغلوطة هي ببساطة معلومات خاطئة أو غير دقيقة – أي أن المعلومات المضللة هي معلومات خاطئة – فإن المعلومات المضللة يتم تلفيقها ومشاركتها عمدًا لأجل التضليل والتلاعب.

ومن الأمثلة على ذلك ما حدث مؤخرًا في أكتوبر/ تشرين الأول 2024، عندما اجتاح مقطع فيديو منصات مثل "إكس" و"فيسبوك" يزعم أنه يُظهر عامل انتخابات في بنسلفانيا يمزق بطاقات الاقتراع التي تحمل علامة دونالد ترامب عبر البريد.

وفي غضون أيام، تمكن مكتب التحقيقات الفيدرالي من تعقب المقطع إلى جهة تأثير روسية، ولكنه كان بالفعل قد حقق ملايين المشاهدات. يوضح هذا المثال كيف تُصنّع حملات التأثير الأجنبي قصصًا ملفقة وتُضخّمها بشكل مصطنع للتلاعب بالسياسة الأمريكية وتأجيج الانقسامات بين الأمريكيين.

إن البشر مُبرمجون على استيعاب العالم من خلال القصص. فمنذ الطفولة، نكبر على سماع القصص وروايتها واستخدامها لفهم المعلومات المعقدة. والسرديات لا تُساعد الناس على التذكر فحسب، بل تُساعدهم على الشعور أيضًا؛ إنها تُعزز الروابط العاطفية وتُشكّل تفسيراتنا للأحداث الاجتماعية والسياسية.

هذا يجعلها أدوات فعّالة للإقناع، وبالتالي لنشر المعلومات المضللة، فالسردية الجذابة قادرة على تجاوز الشكوك وتوجيه الرأي العام بفعالية أكبر من سيل الإحصاءات. على سبيل المثال، غالبًا ما تُثير قصة إنقاذ سلحفاة بحرية، بعد أن علقت بها ماصة بلاستيكية، قلقًا أكبر بشأن التلوث البلاستيكي من كميات هائلة من البيانات البيئية.

أسماء المستخدمين والسياق الثقافي وزمن السرد

إن استخدام أدوات الذكاء الاصطناعي لربط صورة راوي القصة، والتسلسل الزمني لكيفية سردها، والتفاصيل الثقافية المتعلقة بمكان وقوعها، يُمكن أن يُساعد في تحديد متى تكون القصة غير مترابطة. ولا تقتصر السرديات على المحتوى الذي يُشاركه المستخدمون، بل تمتد أيضًا إلى الشخصيات التي يُكوّنها المستخدمون، حتى اسم المستخدم على وسائل التواصل الاجتماعي يُمكن أن يحمل إشارات مقنعة.

لقد طورنا نظامًا يُحلل أسماء المستخدمين لاستنتاج السمات الديموغرافية والهوية، مثل الاسم والجنس والموقع والمشاعر وحتى الشخصية، عندما تكون هذه الإشارات مُضمنة في اسم المستخدم. يسلط هذا العمل، الذي تم تقديمه في عام 2024 في المؤتمر الدولي للويب ووسائل التواصل الاجتماعي، الضوء على أن سلسلة قصيرة من الأحرف يمكن أن تشير إلى الكيفية التي يريد المستخدمون أن ينظر بها جمهورهم إليهم.

فعلى سبيل المثال، قد يختار المستخدم الذي يحاول الظهور كصحفي موثوق به اسم مستخدم مثل "@JimB_NYC". قد يشير مستخدم إلى كلاهما مثل عادي اسم من لابد "@JamesBurnsNYT" ذكر من نيويورك، لكن أحدهما يحمل ثقل المصداقية المؤسسية، وغالبًا ما تستغل حملات التضليل هذه التصورات من خلال صياغة عناوين تحاكي الأصوات أو الانتماءات الحقيقية.

وعلى الرغم من أن الاسم المستعار وحده لا يمكنه تأكيد ما إذا كان الحساب أصليًا، إلا أنه يلعب دورًا مهمًا في تقييم المصداقية بشكل عام. ومن خلال تفسير أسماء المستخدمين كجزء من السردية الأوسع التي يقدمها الحساب، يمكن لأنظمة الذكاء الاصطناعي تقييم ما إذا كانت الهوية مصطنعة لكسب الثقة أو الاندماج في مجتمع مستهدف أو تضخيم محتوى مقنع.

ويُسهم هذا النوع من التفسير الاستدلالي في اتباع نهج أكثر شمولية للكشف عن المعلومات المضللة –

نهج لا يأخذ في الاعتبار ما يُقال فحسب، بل هوية القائل وأسبابه.

ولا تتكشف القصص أيضًا حسب التسلسل الزمني في جميع الأحيان، فقد يُفتتح موضوع على وسائل التواصل الاجتماعي بحدث صادم، ثم يعود إلى لحظات سابقة ويتخطى التفاصيل الرئيسية بينهما. يتعامل البشر مع هذا الأمر بسهولة؛ فنحن معتادون على سرد القصص المجزأة. لكن بالنسبة للذكاء الاصطناعي، لا يزال تحديد تسلسل الأحداث بناءً على السرد القصصي أمرًا يمثل تحديًا كبيرًا.

يعمل مختبرنا أيضًا على تطوير أساليب لاستخراج التسلسل الزمني، مما يُعلّم الذكاء الاصطناعي كيفية تحديد الأحداث وفهم تسلسلها ورسم خريطة لكيفية ارتباطها ببعضها البعض، حتى عندما تُروى القصة بطريقة غير خطية. وغالبًا ما تحمل الأشياء والرموز معانيَ مختلفة في الثقافات المختلفة، وبدون الوعي الثقافي، تخاطر أنظمة الذكاء الاصطناعي بإساءة تفسير السرديات التي تحللها، ويمكن للخصوم الأجانب استغلال الفوارق الثقافية الدقيقة لصياغة رسائل يكون لها صدى أعمق لدى جمهور محدد، مما يعزز قوة الإقناع المعلومات المضللة.

تأمل الجملة التالية: "شعرت المرأة ذات الثوب الأبيض بفرحة غامرة". في السياق الغربي، تستحضر هذه العبارة صورة سعيدة، ولكن في أجزاء من آسيا، حيث يرمز اللون الأبيض إلى الحداد أو الموت، قد تبدو مقلقة أو حتى مسيئة.

لذلك، فإنه من المهم منح الذكاء الاصطناعي هذا النوع من المعرفة الثقافية من أجل استخدامه للكشف عن المعلومات المضللة التي يستغل الرموز والمشاعر والقصص داخل المجتمعات المستهدفة. وقد وجدنا في بحثنا أن تدريب الذكاء الاصطناعي على الروايات الثقافية المتنوعة يحسّن حساسيته لمثل هذه الفروق.

من المستفيد من الذكاء الاصطناعي المدرك للسرديات؟

يمكن لأدوات الذكاء الاصطناعي المدركة للسرديات أن تساعد محللي الاستخبارات على التعرف بسرعة على حملات التأثير المدبرة أو القصص المشحونة عاطفيًا التي تنتشر بسرعة غير عادية، وقد يستخدمون أدوات الذكاء الاصطناعي لمعالجة كميات كبيرة من منشورات وسائل التواصل الاجتماعي من أجل تحديد مسارات السردية المقنعة، وتحديد القصص شبه المتطابقة، وتحديد التوقيت المنسق لنشاط وسائل التواصل الاجتماعي، مما يمكّن أجهزة الاستخبارات بعد ذلك من اتخاذ تدابير آنية مضادة.

بالإضافة إلى ذلك، يمكن لوكالات الاستجابة للأزمات أن تحدد الروايات الضارة بسرعة، مثل ادعاءات الطوارئ الكاذبة أثناء الكوارث الطبيعية. ويمكن لمنصات التواصل الاجتماعي استخدام هذه الأدوات لتوجيه المحتوى عالي الخطورة بكفاءة للمراجعة البشرية دون رقابة غير ضرورية، كما يمكن للباحثين والمعلمين الاستفادة أيضًا من خلال تتبع كيفية تطور القصة عبر المجتمعات، مما يجعل التحليل السردي أكثر دقة وقابلية للمشاركة.

يمكن للمستخدمين العاديين أيضًا الاستفادة من هذه التقنيات؛ حيث يمكن لأدوات الذكاء الاصطناعي أن تشير بصورة آنية إلى منشورات وسائل التواصل الاجتماعي كمعلومات مضللة محتملة، مما يسمح للقراء بالتشكيك في القصص المشبوهة، وبالتالي مواجهة الأكاذيب قبل أن تترسخ.

ومع تزايد دور الذكاء الاصطناعي في رصد وتفسير المحتوى الإلكتروني، أصبحت قدرته على فهم سردية القصص، بما يتجاوز التحليل الدلالي التقليدي، أمرًا بالغ الأهمية. ولتحقيق هذه الغاية، نعمل على بناء أنظمة للكشف عن الأنماط الخفية، وفك رموز الإشارات الثقافية، وتتبع التسلسلات الزمنية للسردية بغرض الكشف عن كيفية انتشار المعلومات المضللة.

المصدر: ذا كونفيرسيشن

رابط المقال: https://www.noonpost.com/315593/