TAREA FINAL CIENCIA DE DATOS RESPONSABLES

Juan David Bejarano Taborda

Pontificia Universidad Católica de Chile

Magister en Ciencia de Datos

Ciencia de Datos Responsable

Septiembre 2023

Principios éticos del caso Facebook en Myanmar

Descripción del caso

En agosto de 2017, las fuerzas de seguridad de Myanmar emprendieron una campaña brutal de limpieza étnica contra la población musulmana rohinyá del norte del estado de Rajine. Una investigación de la ONU concluyó que el papel de Facebook en la violencia fue "considerable". Miles de homicidios de rohinyás, niños de corta edad incluidos; violaron a mujeres y niñas rohinyás, y perpetraron otros actos de violencia sexual contra ellas; incendiaron centenares de pueblos rohinyás. La violencia hizo que más de 700.000 rohinyás (más del 80% de la población rohinyá que vivía en el norte del estado de Rajine al comienzo de la crisis) tuvieran que vivir desplazamiento forzado hacia Bangladesh, donde la mayoría vive aun hoy en campos para refugiados.

En los meses y años previos a las atrocidades de 2017 contra los rohinyá en Myanmar, la plataforma de Facebook se convirtió en un caldo de cultivo para el discurso de odio y la propaganda antirohinyá. Actores ligados al ejército y a grupos nacionalistas budistas radicales utilizaron Facebook para difundir desinformación y propaganda sobre los rohinyá, con el objetivo de demonizarlos y justificar su persecución. Esta incitación al odio, a la violencia y a la discriminación llegó hasta la cúpula de los dirigentes militares de Myanmar.

Estas campañas de desinformación y propaganda tuvieron un impacto significativo en la violencia que se desató contra los rohinyá en 2017. La difusión masiva de mensajes que incitaban al odio y a la violencia contra los rohinyá alimentó el fuego de la discriminación histórica contra este grupo y aumentó el riesgo de un estallido de violencia masiva.

La Misión Internacional Independiente de Investigación de las Naciones Unidas generó un informe que se basó en una investigación exhaustiva sobre el papel de Meta (antes Facebook) en las violaciones graves de derechos humanos perpetradas contra la población rohinyá. Dicho informe revela que, en los meses y años que desembocaron en las atrocidades de 2017, la plataforma de Facebook se convirtió en una cámara de resonancia para el discurso de odio y la propaganda antirohinyá.

De acuerdo con Armistía internacional "Meta no fue una plataforma pasiva que simplemente permitió que se difundiera contenido de odio contra los rohinyá. Sus algoritmos, diseñados para maximizar la participación, amplificaron y promocionaron activamente este contenido". Esto ocurrió porque el modelo de negocio de Meta, basado en la publicidad selectiva, "está diseñado para alimentar la difusión de contenido que genere un alto consumo". Cuando más tiempo pasan los usuarios en Facebook, más ingresos por publicidad genera la empresa. Por lo tanto, los algoritmos de Meta priorizan "el contenido más incendiario, divisivo y dañino", ya que, esta demostrado, este tipo de contenido es el que tiene más probabilidades de mantener a los usuarios enganchados.

En los años previos a la limpieza étnica de los rohinyá en Myanmar, activistas de la sociedad civil advirtieron repetidamente a Meta sobre el riesgo de que su plataforma se utilizara para promover la violencia contra este grupo. Meta recibió estas advertencias, pero no hizo lo suficiente para abordarlas. La empresa no tenía suficientes moderadores de contenido en Myanmar para hacer frente al volumen de discurso de odio que se estaba difundiendo en

la plataforma. Además, no hizo cumplir sus propias normas sobre contenido en relación con la eliminación de "lenguaje que incita al odio".

Meta reconoció finalmente en 2018 que: “no estábamos haciendo lo suficiente para ayudar a evitar que se usara nuestra plataforma para fomentar la división e incitar a la violencia”. Una investigación interna de Meta reconoció claramente que “nuestros sistemas de recomendación aumentan el problema del extremismo”. Estos estudios internos pudieron y debieron haber hecho que Meta implementase medidas efectivas para mitigar los riesgos para los derechos humanos asociados a sus algoritmos, pero la empresa se abstuvo reiteradamente de actuar. Meta afronta actualmente al menos tres demandas judiciales activas en las que se reclama una reparación para la población rohinyá.

Principios éticos

En este caso, podemos encontrar un claro ejemplo de sesgo por confirmación. El sesgo por confirmación se trata de elegir información que corrobore nuestras líneas de pensamiento ya establecidas. Los algoritmos de Meta están diseñados para adaptar y personalizar información para el consumo, pero, además, pueden llevar a las personas a volverse extremas en cuanto a sus creencias, como en este caso. Podríamos afirmar que, en este caso, los algoritmos de Meta exacerbaron la polarización dentro de una población, lo que desembocó que individuos reforzaran sus creencias y terminara en una masacre.

Si analizamos este caso desde el principio ético de la justicia (fairness) y sus 3 dimensiones, podríamos afirmar que Meta omitió, en el mejor de los casos, las dimensiones: ética (no analizó cómo sus algoritmos podían afectar a las personas); legal (no se garantizó que se cumplieran los lineamientos judiciales de evitar el “lenguaje que incita al odio”) y la dimensión técnica (no se evitó la discriminación estadística contra una población). No se observa que se haya buscado mantener la equidad, la inclusividad en el impacto social y en el diseño. Tampoco se evidencia que se haya garantizado la justicia grupal.

Ahora, si analizamos desde el principio ético de la transparencia, donde se busca fomentar la comprensión general de los algoritmos de inteligencia artificial, para que las partes interesadas sean conscientes de sus interacciones, podríamos decir que Facebook no fue transparente sobre la cantidad de moderadores de contenido en lengua birmana cuando ocurrieron las atrocidades. A mediados de 2014, personal de Meta reconoció que únicamente tenían a una sola persona de lengua birmana que moderaba contenido dedicada a Myanmar en ese momento, que trabajaba desde su oficina en Dublín. Meta nunca ha revelado el número exacto de moderadores de contenido en lengua birmana que tenía empleados durante las atrocidades de 2017, pero la empresa afirmó haber contratado a “decenas” más a mediados de 2018. Meta no ha dado visibilidad del funcionamiento del sistema, aun cuando la plataforma se encuentra insertado dentro de un contexto social.

Desde el principio ético de la explicabilidad, para que un modelo sea explicable, debe contar con un adecuado nivel de transparencia. Se debe proveer explicaciones técnicas sobre el funcionamiento del modelo. La transparencia nos permite proporcionar información clara y relevante, mientras que la explicabilidad nos ayuda a que información sea comunicada de manera comprensible y significativa. Al parecer, ninguno de estos principios se cumplió en este caso.

El principio de la interpretabilidad responde la pregunta: ¿por qué funciona el modelo de esta manera? La interpretabilidad se logra cuando es posible discernir el por qué el

algoritmo ha determinado una salida. La interpretabilidad también requiere transparencia, explicabilidad y conocimiento del contexto. Nuevamente, vemos que se presentan omisiones en el principio ético de la interpretabilidad.

Análisis

El uso de la tecnología puede tener tanto beneficios como riesgos. En el caso de la masacre étnica de los rohinyá en Myanmar, la tecnología se utilizó para difundir desinformación y propaganda, acosar y discriminar a los rohinyá, y vulnerar su privacidad y seguridad. Estos riesgos ponen de manifiesto la importancia de regular el uso de la tecnología y de desarrollar medidas para mitigar sus efectos negativos.

Sin embargo. el uso de la tecnología puede tener una serie de beneficios sociales y/o individuales, entre los que se incluyen conectividad y acceso a la información, promoción de la diversidad y la inclusión y mejora de la eficiencia y la productividad. En el caso descrito, la tecnología podría haberse utilizado para: promover la tolerancia y el entendimiento, documentar las atrocidades y ofrecer asistencia a las víctimas. Lo que nos demuestra esto es la tecnología, y en este caso, una plataforma como Facebook no es buena ni mala per se: depende del uso que se le dé. A partir de las medidas que se tomen, se puede ayudar a garantizar que la tecnología se utilice para el beneficio de la sociedad y no para la opresión y la violencia.

Como lección aprendida de este caso, es muy importante garantizar como científicos de datos que, incluyamos en el diseño de nuestros algoritmos, estrategias para abordar este tipo de impactos éticos y sociales. En primer lugar, es muy importante prevenir la discriminación algorítmica, partiendo desde etapas reflexivas, pero llegando incluso a mitigaciones procedimentales y de estándares. La mitigación de sesgos es fundamental y se debe implementar en distintas etapas del proceso de desarrollo (desde la recolección de los datos hasta el postprocesamiento).

El diseño ético de algoritmos debe ser eje central del trabajo de cualquier científico de datos. Se deben adoptar medidas deliberadas y transparentes para garantizar que los prejuicios y estereotipos se aborden de forma adecuada y anticipada. Pensar en el contexto social, en los actores involucrados, puede mejorar la inclusión social y sensibilidad de los algoritmos para distintos grupos. Debemos incluso revisar si los modelos de negocio de nuestras empresas no están funcionando a expensas de dañar o afectar a la población.

Los científicos de datos debemos asegurarnos de llevar a cabo evaluaciones de impacto en los derechos humanos al diseñar e implementar sistemas de IA, incluso al expandir sistemas existentes a nuevos mercados internacionales. Esto implica la realización de consultas y la participación de la sociedad civil, defensores de los derechos humanos y representantes de comunidades marginadas o subrepresentadas antes de finalizar o lanzar un producto o servicio.

También es muy importante que los Estados y distintos mecanismos internacionales verifiquen el cumplimiento de los requisitos legales que obligan a las empresas, incluyendo aquellas en el ámbito de las redes sociales, a implementar medidas de debida diligencia en relación con los derechos humanos en todas sus operaciones, productos y servicios, así como en sus asociaciones comerciales.

BIBLIOGRAFÍA

- Armistía internacional (2022) - Los sistemas de facebook promovieron la violencia contra la población rohinyá. Meta debe una reparación. Tomado de: https://www.amnesty.org/es/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

- Armistía International (2022) - La atrocidad social meta y el derecho a un recurso para la comunidad rohinyá.

- Broussard, M. (2018). People problems. Artificial Unintelligence: How Computers Misunderstand the World (pp. 67-85). The MIT Press.

- Diakopoulos, N. (2020). Transparency. En Dubber, M. D., Pasquale, F., & Das, S. (Eds), The Oxford Handbook of Ethics of AI (pp. 197-213). Oxford University Press.

- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (SSRN Scholarly Paper No. 3518482). Social Science Research Network. https://doi.org/10.2139/ssrn.3518482

- Leavy, S., O'Sullivan, B., & Siapera, E. (2020). Data, Power and Bias in Artificial Intelligence. ArXiv, abs/2008.07341.

- Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–36.

- Varsha, P.S. (2023). How can we manage biases in artificial intelligence systems – A systematic literature review. International Journal of Information Management Data Insights, 3(1), 100165.