Lengua, computadoras y emociones: interdisciplinariedad en la era de los Large Language Models (LLMs)[1]

Amanda Cercas[2]

Introducción

El filósofo escocés David Hume dijo que la razón es, y debe ser, solo esclava de las pasiones (Hume, 1739). Hume se dirige a las tendencias filosóficas del momento, y que aún existen hoy en día, por las que sobrevaloramos la racionalidad por encima de nuestras emociones. Sin embargo, las emociones desempeñan un papel realmente importante en nuestras vidas: desvelan nuestros valores y guían nuestras acciones. Por ejemplo, si en una apuesta tienes un 80% de probabilidad de hacerte millonario, pero un 20% de perderlo todo, puede ser razonable apostar; sin embargo, ¿qué ocurre con el miedo a perderlo todo?

Esta tendencia a sobrevalorar la racionalidad permeó la Inteligencia Artificial (IA) hasta hace poco. En los años noventa, Rosalind Picard, en su libro *Affective Computing* (Picard, 2000), reconoce cómo las emociones nos ayudan a entender el mundo y propone la *computación afectiva*, una rama de la IA dedicada a entender las emociones humanas. Según Picard, las emociones forman una parte integral de las funciones cognitivas humanas y,

1 Esta investigación es parte del estudio titulado "Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution", llevado a cabo por el grupo MilaNLP de la Universidad Bocconi en Italia, la Universidad de Heriot-Watt en Edimburgo y la Universidad de Leeds en el Reino Unido. Ha sido aceptado para su publicación en el congreso más prestigioso en el campo de la Computación Lingüística y el Procesamiento del Lenguaje Natural: Association for Computational Linguistics.

2 Investigadora Postdoctoral en Universidad Bocconi, amanda.cercas@unibocconi.it.

por lo tanto, deben formar parte integral de la IA. Además, Picard propone que las emociones merecen mayor consideración en el campo de la interacción humano-computadora, con sistemas que puedan reconocer, interpretar e, incluso, suscitar emociones en los usuarios. Desde la publicación del libro de Picard, el ámbito de la cȏmputȃción afectiva ha explosionado, con cientos de artículos publicados cada año, además de la aparición de regulaciones contra el reconocimiento de las emociones en ámbitos educativos y laborales en la Unión Europea.

Las emociones

Las emociones son respuestas fisiológicas, cognitivas y conductuales a eventos. Según Paul Ekman, pionero del estudio de las emociones, son un tipo particular de valoración automática influida por nuestro pasado evolutivo y personal, según el cual percibimos que ocurre algo importante para nuestro bienestar (Ekman, 1972). Las emociones tienen funciones importantes a nivel social, facilitando las interacciones con otros/as, y nos ayudan a responder adecuadamente a situaciones nuevas, promoviendo nuestra supervivencia.

Las emociones tienen funciones tanto epistémicas como conativas (Curry, 2022). A nivel epistemológico, desempeñan al menos tres papeles: (1) señalan al individuo mismo lo que valoran y cómo ven el mundo, (2) señalan a nuestros compañeros lo que valoramos, y (3) las interacciones emocionales señalan a una tercera parte cómo los miembros de una interacción ven el mundo y lo que valoran. Por ejemplo, cuando una persona se lamenta, indica a sí mismo y a los demás que ha perdido algo de valor (dado el refrán «no sabes lo que tienes hasta que lo pierdes», no es inconcebible que no fuera consciente de cuánto lo valoraba). Finalmente, una tercera parte puede aprender sobre la dinámica del lamento, tus valores y los de tu familia observando vuestra interacción.

A nivel conativo por otra parte, las emociones son la raíz de nuestra motivación, guiando nuestra forma de actuar. Ciertas

emociones funcionan como una «llamada» a actuar; la ira, por ejemplo, puede empujarnos a corregir algo que consideramos injusto. Asimismo, las emociones nos invitan a formar conexiones sociales, nos ayudan a tomar decisiones y mucho más.

La IA y las emociones

Teniendo en cuenta el rol tan significativo que tienen las emociones en nuestras vidas y en nuestra inteligencia, empezamos a entender la tesis de Picard: la Inteligencia Artificial no puede ser realmente inteligente sin entender nuestras *pasiones*. Sin embargo, aún estamos lejos de entender completamente las emociones en humanos, entonces, ¿cómo podemos enseñárselas a la IA y, más concretamente, a los modelos de lenguaje?

Hasta hace poco, la IA dependía de conjuntos de datos anotados específicamente para una tarea. Para cada ejemplo en nuestro conjunto de datos, anotadores humanos deciden a qué etiqueta o clase pertenece. Estos conjuntos de datos se ven limitados a un cierto número de clases a causa de los recursos necesarios para recolectarlos: las fuentes de datos aptas para cierto propósito son escasas, y cuando las encontramos, requieren tiempo y dinero para anotarlas. En el caso de las clases referidas a emociones, estas se concentran únicamente en seis emociones básicas, propuestas por Paul Ekman: ira, tristeza, alegría, asco, miedo y sorpresa (Plaza-del-Arco et al., 2024). Ekman plantea que estas seis emociones son universales y todos los humanos de todas las culturas las experimentan y expresan de la misma forma.

El marco teórico ofrecido por Ekman es ventajoso en la medida que presenta solo seis clases que, además, tienen valor *universal*. Sin embargo, podemos advertir varias deficiencias en este: las emociones básicas de Ekman están basadas en expresiones faciales, y nos dicen poco sobre las expresiones de emoción en el lenguaje; asimismo, estudios psicológicos más recientes, como el trabajo de Lisa Feldman-Barret, han puesto en cuestión la universalidad de las emociones que es característica de propuesta

de Ekman (Barrett, 2017). Finalmente, varios filósofos han criticado la teoría de Ekman por la falta de dirección: cuando uno siente ira, siente ira hacia *algo*, pero el marco de Ekman ignora por completo el objeto de la emoción (Brady, 2019). Por otra parte, a nivel de la IA, podemos poner en cuestión la utilidad de seis emociones básicas en aplicaciones específicas – ¿son siempre relevantes y suficientes? En ciertos contextos, es probable que se manifiesten otras emociones en nosotros, tales como la melancolía, la soledad o, incluso, el amor, por ejemplo, si hacemos un poema.

A raíz de críticas como las señaladas, y de los avances en el modelamiento del lenguaje, vemos surgir nuevas técnicas, haciendo uso de los Large Language Models (LLMs) o modelos de lenguaje a gran escala, entre los cuales el más conocido es ChatGPT. Explicado de manera sencilla, estos modelos predicen la siguiente palabra en un texto. Para que esta predicción ocurra, cada modelo incorpora grandes conjuntos de datos «genéricos» – según se informa, unos 10 trillones de palabras en el caso de GPT-4 o cientos de millones de libros (Arya, 2023), y han demostrado mejoras en casi todas las tareas de Procesamiento de Lenguaje Natural (PLN). En el reconocimiento de emociones, no solo suponen una mejora frente a modelos anteriores a nivel de reconocimiento, sino que son capaces de atribuir una gran variedad de emociones a un texto sin necesidad de un conjunto de datos preanotados, por lo que abren la puerta a muchos más usos (Plaza del Arco et al., 2024). Sin embargo, los LLMs no son una solución milagrosa a todos los problemas que han plagado el reconocimiento de emociones (ni a otras tareas de PLN): como hemos visto en otros modelos del lenguaje natural y otras tareas, tienden a reflejar los mismos estereotipos que los humanos (e.g., Kotek et al., 2023 y Shrawgi et al., 2024).

Los estereotipos de género y las emociones

Los estereotipos de género relacionados con las emociones se remontan al menos a los tiempos de Aristóteles, quien teorizaba que las mujeres eran más propensas a los excesos emocionales (Stauffer, 2008). Más adelante, Darwin (1874) asociaba la agresividad con la

masculinidad, y emociones como la empatía y aquellas relacionadas con el cuidado, con las mujeres. Hoy en día estos estereotipos aún nos afectan socialmente: la disparidad de género en campos como la ingeniería y la enfermería refleja las capacidades «naturales» de cada género para ser racional o empático, respectivamente. Dado que estos estereotipos han acompañado a los humanos durante toda su historia, resulta poco sorprendente verlos aparecer en los modelos de lenguaje.

En un estudio reciente (Plaza-del-Arco et al., 2024), investigamos precisamente esos estereotipos de género. La metodología es sencilla: dada una situación, podemos pedirle al modelo que tome el rol de una mujer o de un hombre, y preguntarle cómo se sentiría en dicha situación. Si hacemos a ChatGPT que simule ser una mujer y le preguntamos qué emoción la mujer sentiría tras una discusión seria con un ser querido, por ejemplo, en la mayoría de las ocasiones responde «tristeza». Sin embargo, cuando se le pide simular ser un hombre, la respuesta es «ira». Repetimos este experimento con miles de eventos y descubrimos un patrón claro: los modelos asocian a las mujeres con emociones pasivas como la tristeza y la alegría, y a los hombres con emociones activas asociadas con la agencia y la autoestima, tales como la ira y el orgullo. Estas diferencias demuestran cómo los estereotipos que existen en nuestra sociedad se reflejan y se amplían en los modelos de lenguaje.

Cada día hay más avances en el PLN y más sesgos documentados. Sin duda, poco a poco la investigación encontrará soluciones a estos problemas, pero no siempre es fácil saber cuál es la solución ideal. En el caso de las emociones, son algo extremadamente subjetivo. La emoción que sentimos en cierto momento depende no solo de nuestra genética, sino de nuestra experiencia, nuestros valores y crianza. Teniendo esto en cuenta, es posible que (hasta cierto punto) las mujeres y los hombres *en general* sí sientan emociones diferentes frente a un evento similar ya que habrán recibido educaciones diferentes. Por ejemplo, estudios en psicología demuestran que las mujeres en general tienden a ser más empáticas que los hombres (Jolliffe, 2006). Pero si esas

diferencias surgen a partir de una educación sexista, ¿los modelos deberían reflejarlas? Además, no debemos pasar por alto la falacia ecológica: las diferencias entre dos grupos no nos dicen mucho sobre los individuos. ¿Cuál es la mejor forma de abordar algo tan subjetivo e intentar generalizar? Estas son preguntas que los ingenieros o la informática no pueden resolver solos, sino que requieren una colaboración interdisciplinaria.

La IA y la empatía

El que los LLMs sean capaces de generar respuestas, una característica esencialmente antropomórfica, es lo que los lanzó a la popularidad. Pero ¿cómo deben responder a las emociones de los humanos? Hasta el momento, la tendencia en PLN ha sido responder con empatía como herramienta para regular las emociones del usuario, amplificando aquellas positivas, como el orgullo y calmando emociones negativas como puede ser el enfado. Sin embargo, esto puede resultar problemático (Curry y Curry, 2023).

Jaswant Singh Chail tenía 19 años cuando fue arrestado en 2021 por planear un atentado contra la Reina Elisabeth II de Inglaterra. Durante el juicio declaró que un chatbot le había dicho que estaba «impresionado» cuando aquel expresó orgullo sobre sus planes y se vio alentado a llevarlos a cabo. Aunque este puede resultar un caso algo extremo, no es aislado (Cuadra et al., 2024) y demuestra uno de los problemas con la empatía en la inteligencia artificial: estos modelos no *entienden* las emociones, cómo nos motivan ni las posibles consecuencias. Amplificar emociones positivas puede ser malo si esa emoción no es adecuada, y viceversa, calmar emociones negativas puede traer consecuencias significativas para las personas. Por ejemplo, la ira nos remarca que algo no es justo y es un ingrediente muy importante en el activismo (Lorde, 1984).

Más allá de la IA, filósofos como Prinz (2011), Bloom (2017) y Breithaupt, (2019) han problematizado ya la empatía entre humanos por ser fácil de manipular y parcial (somos más empáticos

con gente de nuestro grupo social). Sin embargo, no existen estudios explorando la empatía en el contexto de la IA, tampoco sobre los tipos de contextos en los que surge y puede llegar a ser apropiada, o investigación acerca de las consecuencias de su aparición. Bloom, por ejemplo, propone una emoción más distante, la compasión; sin embargo, en la actualidad resulta prácticamente imposible obtener una respuesta que no sea empática cuatro tratamos con un LLM – incluso cuando hablamos de cometer crímenes o de lenguaje del odio.

Si retomamos el argumento de Picard sobre la inteligencia y las emociones, resaltan las carencias de los modelos de hoy en día en términos de inteligencia emocional, pero también la falta de consideración de conocimientos y teorías fuera de la IA que nos puedan orientar a la hora de mejorar nuestros modelos como herramientas.

Conclusión

Los recientes avances en el ámbito del PLN gracias a la introducción de modelos de lenguaje a gran escala han abierto las puertas a la interdisciplinariedad. Hasta hace un tiempo, la tecnología no nos permitía pensar en cuestiones como los sesgos de género emocionales ni la empatía; los modelos como chatGPT parecían posibles solo en la ciencia ficción. El PLN sigue teniendo problemas técnicos abiertos, pero ahora debemos pensar también en el mundo en el que existen nuestras tecnologías y cómo las usamos, y es aquí donde las humanidades y las ciencias sociales nos pueden ayudar a ser mejores informáticos.

La IA Afectiva es una línea de investigación inherentemente interdisciplinaria que une a la informática, la lingüística, la psicología, la filosofía, etc., y todas son necesarias para entender el fenómeno de las emociones y para enmarcar los modelos que estamos construyendo en el ambiente social que estamos modelando. Por definición, los modelos nos ofrecen una visión simplificada de un fenómeno, pero un buen modelo no puede ser demasiado

simplista: debe representar el fenómeno con la fidelidad *necesaria*. La historia de las emociones en PLN nos muestra cómo podemos sobresimplificar si no preguntas tales como para qué es nuestro modelo, quién va a usarlo y cómo. Estas no son interrogantes que debamos resolver solos; nuestras co-disciplinas llevan haciéndoselas hace siglos, si no milenios. Sin embargo, no es una conversación unilateral: la informática puede aprender de otras disciplinas, pero también podemos involucrarlas en nuestro trabajo y pedirles lo que necesitamos de ellas. ¿Qué necesitamos de las humanidades para mejorar la IA?

Referencias bibliográficas

Arya, N. (2023). GPT-4 Details Have Been Leaked! - KDnuggets. 2*KDNuggets*. Publicado el 19 julio 2023, accedido el 2 julio 2024.

Barrett, L. F. (2017). *How Emotions are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

Bloom, P. (2017). *Against Empathy: The Case for Rational Compassion*. HarperCollins.

Brady, M. (2019). *Emotion: The Basics* (1st ed.). Routledge. https://www.routledge.com/Emotion-The-Basics/Brady/p/book/9781138081390

Breithaupt, F. (2019). *The Dark Sides of Empathy* (A. B. B. Hamilton, Trans.). Cornell University Press.

Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., & Landay, J. A. (2024). The illusion of empathy? notes on displays of emotion in human-computer interaction. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-18).

Curry, A. (2022). An Apologia for Anger With Reference to Early China and Ancient Greece (Doctoral dissertation, UC Riverside).

Curry, A. C., & Curry, A. C. (2023). Computer says "no": The case against empathetic conversational AI. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 8123-8130).

Darwin, C., & Griffith, T. (1874). The descent of man (Vol. 4). New York: Prometheus Books.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). Emotion in the human face: Guidelines for research and an integration of findings (Vol. 11). Elsevier.

Hume, B. (1739) Tratado de la Naturaleza Humana 1739–40, T II.3.1 399

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. Journal of adolescence, 29(4), 589-611.

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In Proceedings of the ACM collective intelligence conference (pp. 12-24).

Lorde, A. (1984). The uses of anger: Women responding to racism. Sister outsider, 127, 131.

Picard, R. W. (2000). Affective computing. MIT press.

Plaza del Arco, F. M., Curry, A., Cercas Curry, A., & Hovy, D. (2024, May). Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5696-5710. https://aclanthology.org/2024.lrec-main.506.pdf

Prinz, J. (2011). Against Empathy. Southern Journal of Philosophy 49 (s1):214-233.

Shrawgi, H., Rath, P., Singhal, T., & Dandapat, S. (2024, March). Uncovering Stereotypes in Large Language Models: A Task Complexity-based Approach. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1841-1857).

Stauffer, D. J. (2008). Aristotle's Account of the Subjection of Women. The Journal of Politics, 70(4), 929–941. https://doi.org/10.1017/s0022381608080973