Revisión bibliográfica

Procesamiento de datos textuales EthicApp con algoritmos de procesamiento de lenguaje natural

Camilo Carvajal Reyes

18 de abril de 2023

Introducción

El procesamiento del lenguaje natural (NLP) ha ganado la atención de los medios y la sociedad en general en los últimos meses. Los últimos modelos conversacionales como chatGPT han generado mucha atención del público. A pesar de las capacidades que han demostrado, este tipo de modelos no son capaces de procesar ideas de la misma manera que lo hacen los humanos. Este es especialmente el caso de las decisiones que implican juicios morales.

En este contexto, responder éticamente a una pregunta se ha estudiado en el contexto de grandes modelos de lenguaje para mejorar su capacidad de ayudar a los seres humanos. Además, cuando entrenamos modelos para predecir respuestas similares a las humanas a dilemas éticos y otras tareas generales de NLP, a veces somos capaces de detectar patrones y estructuras que podrían explicar cómo las diferentes culturas enfrentan problemas morales.

En este informe analizaremos diferentes artículos que involucran tanto la moralidad como la semántica computacional. Estos trabajos van desde modelos de entrenamiento para la predicción, análisis de juicios morales en redes sociales, corpus y conjuntos de datos que contienen respuestas éticas y recursos lingüísticos que pueden ayudar a procesar grandes cantidades de texto.

Modelos de lenguaje aprendiendo valores humanos

Los modelos de lenguaje se han convertido en herramientas útiles a la hora de enfrentar diversas tareas de procesamiento de lenguaje, en especial aquellos que han sido entrenados en grandes corpuses de texto. Su capacidad de codificar de antemano elementos del lenguaje y naturaleza humana en su etapa de entrenamiento ha sido puesta a prueba también para predecir juicios morales humanos. Jentzsch et al. testean un modelo de estas características para predecir juicios usando simplemente similitudes de representaciones, mostrando que decisiones naturalmente humanas aparecen codificadas naturalmente [1].

En esta dirección, han habido iniciativas de verificar la capacidad que esta familia de modelos tendrían para imitar el razonamiento moral humano. Por un lado Jin et al. entrenan un modelo en un dataset de preguntas y respuestas morales [2]. El método usado combina ciencias de la cognición con semántica computacional al intentar construir una cadena de pensamientos morales. Los

relatados sugieren que es necesario incorporar elementos de razonamiento humano, y se sugieren lineas de investigación posteriores al respecto.

Otro ejemplo importante es *Delphi* propuesto por Jiang et al., un modelo cuyo objetivo principal es imitar el proceso de pensamiento ético de los humanos [3]. Este modelo ha logrado consistencia en casos donde modelos de lenguaje suelen fallar, especialmente en dilemas nunca antes observados por el modelo. Se propone su uso paralelo a modelos de lenguaje y tiene un prototipo de investigación disponible en linea [1]. Sus capacidades han sido puestas a prueba, argumentándose que sus capacidades efectivamente imitan aquellas que son propias de los grupos culturales que aportaron a las anotaciones de entrenamiento [4]. Naturalmente, existen críticas al proceso de entrenar modelos para cuestiones morales. Albrecht et al. argumentan que si bien ciertos modelos suelen tener métricas de desempeño similares a las de los seres humanos, esta no refleja un razonamiento o entendimiento de la problemática en cuestión [5]. Una muestra de esto es pedirle justificación a modelos conversacionales, en donde se desprenden justificaciones que resultan ser poco éticas.

Datasets que contienen decisiones o juzgamientos morales

En el contexto de los modelos previamente mencionados y, más generalmente, para ampliar la capacidad de exploración en el área, se han propuesto varios datasets que exploran las respuestas humanas a cuestiones morales en distintos formatos.

- MoralExceptQA: un dataset que consiste en preguntas y respuestas de excepciones morales [2].
- ETHICS dataset: un benchmark que incluye conceptos de justicia, bienestar, deberes, virtudes y moralidad de sentido común [6].
- Datset basado en el foro “AmITheAsshole” (AITA) del sitio web Reddit [7].
- Moral Foundations Twitter Corpus: un colección de 35108 tweets procesado para siete categorias de discurso y anotado a mano por anotadores entrenados para 10 categorías de sentimiento moral [8], siguiendo la teoría de fundamentos morales (ver sección).
- SCRUPLES: un dataset con 625000 juzgamientos éticos de más de 32000 anécdotas de la vida real [9].
- The Moral Machine: dataset con decisiones morales de personas de todas partes del mundo, en el contexto específico de decisiones morales que debe seguir la conducción automática de vehículos [10].

Recursos psico-linguisticos

Más allá del aprendizaje automático, la extracción de elementos que den muestra del razonamiento moral detrás de texto escrito por humano lleva siendo investigado por un tiempo. Uno de los marcos teóricos que ha tenido más influencia en este contexto ha sido la **teoría de fundamentos morales** (TFM), propuesta por Haidt y Graham [11, 12]. Esta teoría de psicología social

[1] https://delphi.allenai.org/

propone la existencia de fundamentos morales innatos y universales que afectan los juicios éticos de las personas. Estos incluyen: cuidado/daño, equidad/trampa, lealtad/fraude, autoridad/subversión, santidad/degradación y libertad/opresión. Se plantea la presencia de estos elementos en distintas culturas, pese a la evolución de estos. Las diferencias en comportamientos y creencias éticas (y como consecuencia distintos compartimientos grupales e ideologías) son resultado del enfasis que se le da a algunos fundamentos por sobre otros.

En este contexto, se han desarrollado recursos semánticos como el diccionario de fundamentos morales (DFM por sus siglas en inglés), que mide el grado en el que ciertas palabras reflejan los seis fundamentos morales planteados en la TFM [13]. En este se incluyen palabras reacionadas a los fundamentos y se les asigna un puntaje de acuerdo la frecuencia de uso y grado de lenguaje moral empleado. El diccionario se ha usado en varios estudios para examinar el contenido moral de varios textos, tales como discursos, artículos y mensajes de redes sociales.

El TFD ha sido extendido tanto usando métodos semi-automáticos [14], usando anotadores para registrar las probabilidades de pertenencia a los fundamentos [15], como también usando otras herramientas semánticas como *wordnet*[2] [16]. Estos recursos pueden usarse tanto para detectar directamente palabras en texto que corresponda a las distintas dimensiones morales, como para complementar el trabajo de clasificadores automáticos en distintas tareas que pueden ser favorecidas por tener conocimiento directo de palabras con valor moral.

Modelos que buscan predecir comportamiento humano

Varios son los trabajos que buscan la predicción/identificación de elementos morales en texto natural. En esta sección abordaremos algunos de ellos. Primeramente, Garten et al. 2016 busca la detección automática de retóricas morales [17]. Para esto usan el diccionario de fundamentos morales [13], combinado con representaciones de palabras distribuidas. Estas últimas, más conocidas por su traducción en inglés *word embeddings*, corresponden a asignar un vector fijo a cada palabra. Estos vectores suelen ser obtenidos usando redes neuronales, siguiendo la hipótesis distribucional, que nos dice que palabras con significado similar aparecerán en contextos similares (y por ende deben tener una representación vectorial similar). Otro trabajo que usa categorías provenientes de la TFM es el de Xie et al. que evalua modelos en la clasificación de dilemas morales en los distintos fundamentos, gracias a lo cual concluyen que modelos de lenguaje tienen ventaja sobre modelos como las representaciones distribucionales [18].

Por otro lado, Kennedy et al. buscan la predicción de preocupaciones morales propias a un individuo usando evidencias de lenguaje moral escritas por este [19]. Los datos usados consisten en estados de Facebook de usuarios que hayan contestado el cuestionario de fundamentos morales, basados en la teoría del mismo nombre mencionada anteriormente. Se utilizaron distintas técnicas de procesamiento de lenguaje para predecir los puntajes obtenidos por los usuarios, para cada una de las dimensiones morales planteadas en la TFM. Se destacan la variedad de métodos para vectorizar texto testeados, incluyendo *latent dirichlet allocation* (LDA) [20], *word embeddings*, conteo de ocurrencias del DFM [13] y BERT [21], que corresponde a un modelo de lenguaje profundo. Es este último que obtiene mejores resultados. Finalmente, tanto conteos de diccionario como LDA se usaron para interpretar que elementos linguísticos específicos explicaban cada fundamento por separado.

[2] Wordnet es una base de datos lexical para el idoma inglés: https://wordnet.princeton.edu/

Igualmente en la linea de la utilización de redes sociales, trabajos recientes hen hecho uso del dataset AITA del sitio *Reddit* [7]. En la comunidad en cuestión, un usuario expone su problemática, para que luego los usuarios juzguen si es que es culposo o no. Alhassan et al. por un lado intentan predecir el resultado final (veredicto con más votos) usando el texto del usuario original [22]. Por otro lado, Botzer et al. utilizan los comentarios para entrenar un clasificador que dirime entre una valoración positiva o negativa del actuar del usuario en la situación [23]. Efstathiadis et al. además de clasificar tanto posts como comentarios, exploran patrones que emergen desde el texto y desde los clasificadores mismos [24]. Estos trabajos hacen uso de modelos de lenguaje pre-entrenados como BERT [21].

1. Conclusión

Los trabajos mencionados en este reporte muestran la alta variedad de formatos en los cuales se ha evaluado la presencia y grado de categorías morales, así como también la capacidad de distintos modelos de procesamiento de lenguaje natural para modelarlos. No obstante, ninguno de los artículos estudiados enfrenta un desafío tan específico como el nuestro. Los datos que poseemos tienen la ventaja de poder verse desde distintos ángulos, lo cual plantea dificultades pero también abre puertas a que las conclusiones que se puedan tomar sean reflejo de nuevos descubrimientos en el área. Para finalizar, muchos de los artículos nos confirman la pertinencia de los modelos a usar, en el caso de modelos de lenguaje, y nos sugieren algunos tipos de modelos más simples que tengan capacidad interpretativa.

Referencias

[1] Jentzsch, S., Schramowski, P., Rothkopf, C., Kersting, K. (2019). Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 37–44. https://doi.org/10.1145/3306618.3314267

[2] Jin, Z., Levine, S., Gonzalez Adauto, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., Schölkopf, B. (2022). When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment. Advances in Neural Information Processing Systems, 35, 28458–28473. https://openreview.net/forum?id=uP9RiC4uVcR

[3] Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., Choi, Y. (2022). Can Machines Learn Morality? The Delphi Experiment (arXiv:2110.07574). arXiv. https://doi.org/10.48550/arXiv.2110.07574

[4] Fraser, K. C., Kiritchenko, S., Balkir, E. (2022). Does Moral Code have a Moral Code? Probing Delphi's Moral Philosophy. Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), 26–42. https://doi.org/10.18653/v1/2022.trustnlp-1.3

[5] Albrecht, J., Kitanidis, E., Fetterman, A. J. (2022). Despite 'super-human' performance, current LLMs are unsuited for decisions about ethics and safety (arXiv:2212.06295). arXiv. https://doi.org/10.48550/arXiv.2212.06295

[6] Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., Steinhardt, J. (2023, April 2). Aligning AI With Shared Human Values. International Conference on Learning Representations. https://openreview.net/forum?id=dNy_RKzJacY

[7] O'Brien, E. (2020, February 17). AITA for making this? A public dataset of Reddit posts about moral dilemmas (BLOG post). Developer Tools for Machine Learning | Iterative. https://iterative.ai/blog/a-public-reddit-dataset/

[8] Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., ... Dehghani, M. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. Social Psychological and Personality Science, 11(8), 1057-1071. https://doi.org/10.1177/1948550619876629

[9] Lourie, N., Le Bras, R., Choi, Y. (2021, May). Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 15, pp. 13470-13479).

[10] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I. (2018). The Moral Machine experiment. Nature, 563(7729), Article 7729. https://doi.org/10.1038/s41586-018-0637-6

[11] Haidt, J. (2007). The New Synthesis in Moral Psychology. Science, 316(5827), 998–1002. https://doi.org/10.1126/science.1137651

[12] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., Ditto, P. H. (2013). Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In P. Devine A. Plant (Eds.), Advances in Experimental Social Psychology (Vol. 47, pp. 55–130). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

[13] Graham, J., Haidt, J., Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology, 96, 1029–1046. https://doi.org/10.1037/a0015141

[14] Rezapour, R., Shah, S. H., Diesner, J. (2019). Enhancing the Measurement of Social Effects by Capturing Morality. Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 35–45. https://doi.org/10.18653/v1/W19-1305

[15] Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., Weber, R. (2021). The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behavior Research Methods, 53(1), 232–246. https://doi.org/10.3758/s13428-020-01433-0

[16] Araque, O., Gatti, L., Kalimeri, K. (2020). MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. Knowledge-Based Systems, 191(C). https://doi.org/10.1016/j.knosys.2019.105184

[17] Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., Dehghani, M. (2016). Morality Between the Lines: Detecting Moral Sentiment In Text. Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes.

[18] Xie, J. Y., Hirst, G., Xu, Y. (2020). Contextualized moral inference (arXiv:2008.10762). arXiv. https://doi.org/10.48550/arXiv.2008.10762

[19] Kennedy, B., Atari, M., Mostafazadeh Davani, A., Hoover, J., Omrani, A., Graham, J., Dehghani, M. (2021). Moral concerns are differentially observable in language. Cognition, 212, 104696. https://doi.org/10.1016/j.cognition.2021.104696

[20] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3(null), 993–1022.

[21] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

[22] Alhassan, A., Zhang, J., Schlegel, V. (2022). 'Am I the Bad One'? Predicting the Moral Judgement of the Crowd Using Pre–trained Language Models. Proceedings of the Thirteenth Language Resources and Evaluation Conference, 267–276. https://aclanthology.org/2022.lrec-1.28

[23] Botzer, N., Gu, S., Weninger, T. (2022). Analysis of Moral Judgment on Reddit. IEEE Transactions on Computational Social Systems, 1–11. https://doi.org/10.1109/TCSS.2022.3160677

[24] Efstathiadis, I. S., Paulino-Passos, G., Toni, F. (2022). Explainable patterns for distinction and prediction of moral judgement on reddit. https://arxiv.org/pdf/2201.11155.pdf