Dinámicas Sociales y Moralidad Colectiva en Reddit: Un Estudio del Subreddit r/AmITheA**hole utilizando técnicas de Machine Learning

Carlos Andres Espitia Alfonso*, Michael Santiago Moreno Bravo*,
Juan David Salguero Medina†, Juan Diego Yepes Parra‡
*Doctorado en Ingeniería
†Pregrado en Ingeniería de Sistemas y Computación,
‡Maestría en Ingeniería de Sistemas y Computación,
Facultad de Ingeniería, Universidad de los Andes, Bogotá, Colombia

***Resumen*—**
***Index Terms*—Palabras clave; separadas; por punto y coma.**

I. Introducción

Las redes sociales han revolucionado la manera en que las personas interactúan, comparten opiniones y construyen comunidades. Reddit es de las redes sociales más particulares. Especialmente conocida por su estructura y sus foros especializados en los que los usuarios pueden debatir sobre cualquier tema, de forma anónima. Un aspecto interesante de Reddit son sus ”subreddits”, foros o comunidades centradas en un tema donde los usuarios pueden crear y compartir publicaciones, comentarlas y votar en ellas.

Dentro de estas comunidades, existe r/AmITheA**hole (AITA), un espacio donde los usuarios comparten historias personales sobre situaciones complejas y piden a la comunidad que los juzgue o aconseje, respondiendo a la pregunta: ”¿Soy el a**hole (el malo de la historia)?”. Este tipo de interacción genera una fuente de datos o dataset único que puede servir de insumo para uno o varios modelos de aprendizaje automático o machine learning; los cuales pueden ayudar con el objetivo de analizar cómo se desarrollan las dinámicas sociales en línea, cómo se toma una decisión colectiva y si es posible automatizarla, y qué factores influyen en la aceptación o rechazo de una historia.

Esta comunidad ha ganado una popularidad considerable; las publicaciones varían en complejidad, desde simples malentendidos hasta situaciones más serias que involucran relaciones familiares, parejas, amigos o compañeros de trabajo. Los participantes, al presentar sus historias, esperan obtener un veredicto, en la cual se tienen en cuenta los comentarios de la publicación, donde los participantes de la comunidad opinan utilizando un comentario con el voto específico, el cual puede ser YTA you’re the a**hole (eres es el malo), NTA not the a**hole (no eres el malo), ESH everyone sucks here (todos son malos) entre otros votos posibles, al igual que expresan sus opiniones respecto a la situación y los actores. Sin embargo, los votos a favor o en contra (upvotes o downvotes), no están ligados al veredicto, sino a la popularidad y qué tan interesante es el post.

Existen varios problemas que afectan la calidad del contenido y la carga de trabajo de los moderadores de la comunidad. A medida que la comunidad crece, los moderadores pueden verse sobrecargados con la gestión de las publicaciones, lo cual consume mucho tiempo, especialmente si se alcanza una cantidad significativa de publicaciones diarias. Esta saturación no solo afecta la efectividad de los moderadores, sino que también puede influir en la calidad de las interacciones y en la precisión de los veredictos que se emiten.

Esto abre la posibilidad para considerar una herramienta automatizada que apoye a los moderadores a poner el veredicto en la publicación. La inclusión de dicha herramienta sería en complemento a su labor actual, donde el análisis de publicaciones se puede realizar de manera más eficiente, permitiendo que los moderadores se enfoquen en aspectos más subjetivos y delicados, mientras que el sistema se encargaría de dar una sugerencia de acuerdo al contenido de la publicación.

Otro desafío significativo es la gestión de la interacción entre los moderadores y los usuarios. Si bien los moderadores deben ser imparciales, es posible que existan casos donde los prejuicios personales influyan en el veredicto de una publicación, lo cual podría generar conflictos dentro de la comunidad. Además, las reglas del subreddit pueden no ser siempre claras para todos los usuarios, lo que genera cierta ambigüedad sobre qué publicaciones son aceptables y cuáles no. La existencia de subcategorías, como WIBTA (Would I Be The A**hole), que plantea situaciones hipotéticas sobre el futuro, requiere una segmentación más clara entre los distintos tipos de publicaciones para asegurar una correcta categorización y evaluación.

Es por ello que se considera de interés investigar los temas que más frecuentemente influencian los veredictos, como las relaciones familiares o las disputas laborales, para poder identificar patrones en la toma de decisiones y en las dinámicas que subyacen a los votos. ¿Por qué algunas historias reciben más popularidad que otras? ¿Es el contenido de las publicaciones

lo que determina su aceptación o el perfil del usuario que las comparte? ¿Qué tan relevantes son la longitud del texto y otros factores en la popularidad de un post? Estas preguntas requieren una investigación para entender cómo el algoritmo de Reddit y la interacción humana juegan un papel crucial en los resultados finales.

r/AmITheA**hole ofrece una oportunidad para comprender mejor las dinámicas sociales en línea, así como para investigar cómo se toman decisiones colectivas en comunidades virtuales. Este conjunto de datos es lo suficientemente grande y diverso como para permitir un análisis profundo de los factores que influyen en la popularidad de las publicaciones y en los veredictos que emiten los usuarios. La variedad de temas tratados en las publicaciones hace que este análisis sea aún más relevante, ya que abarca una gama amplia de situaciones personales y sociales.

Además, este análisis es un ejercicio académico interesante, ya que permite aplicar técnicas de procesamiento de lenguaje natural (NLP) y modelos de machine learning para examinar cómo los usuarios interactúan en este tipo de plataformas. Combinar ambos enfoques ofrece una oportunidad para probar la efectividad de los modelos de manera práctica, obteniendo resultados que puedan contribuir al aprendizaje sobre el funcionamiento de estos sistemas. Finalmente, este ejercicio también puede llegar a ser interesante en la evaluación de modelos extensos de lenguaje (LLMs), utilizando como prompt el post y preguntándole cuál veredicto daría, con el fin de probar programáticamente qué tan acertados son.

Este estudio tiene el potencial de mejorar la comprensión de los mecanismos que rigen las decisiones de los usuarios en Reddit y en otras plataformas de redes sociales, lo que puede abrir nuevas posibilidades para la automatización de tareas y la creación de herramientas más eficientes para la moderación de contenido. Al mismo tiempo, este tipo de investigación puede ser útil para desarrollar algoritmos que no solo analicen el contenido de las publicaciones, sino que también reconozcan los matices humanos que influyen en las decisiones de las personas.

I-A. Objetivo General

I-B. Objetivos Específicos

1. I

II. Estado del arte

El análisis automatizado de decisiones morales en comunidades en línea, especialmente en Reddit, ha captado la atención académica por su potencial para mejorar la comprensión de las dinámicas sociales en estas plataformas. En particular, investigaciones recientes se han enfocado en estudiar cómo las comunidades virtuales realizan juicios éticos sobre situaciones personales complejas, buscando identificar factores que influyen en el veredicto colectivo, tales como características del contenido, perfil del usuario o señales sociales implícitas.

Botzer *et al.* [1] abordaron este tema mediante un análisis detallado del subreddit `r/AmITheAsshole` (AITA), examinando la influencia del género, la edad y el tipo de juicio moral expresado sobre la popularidad y recepción social de las publicaciones. Sus hallazgos mostraron que las historias consideradas moralmente positivas suelen tener mayor aceptación y popularidad, reflejada en una mayor cantidad de votos positivos y comentarios generados por la comunidad. Este estudio utilizó modelos avanzados basados en BERT (Judge-BERT) para clasificar publicaciones según el juicio moral colectivo, destacando la importancia de considerar tanto factores lingüísticos como sociales en la dinámica de aceptación del contenido. Aunque su investigación contribuye significativamente a identificar qué factores influyen en la popularidad de las publicaciones, no considera directamente la automatización de veredictos como herramienta práctica para apoyar a los moderadores, aspecto central en nuestra propuesta.

Por otro lado, Alhassan *et al.* [2] presentaron una investigación centrada exclusivamente en la predicción del juicio moral en AITA utilizando modelos avanzados de procesamiento de lenguaje natural (RoBERTa y Longformer). Sus resultados demostraron una alta efectividad (hasta 87 % de precisión), confirmando que estos modelos pueden replicar de forma acertada las decisiones colectivas tomadas por la comunidad. Esta investigación es relevante para nuestro trabajo, ya que valida la aplicación de modelos avanzados para clasificar juicios morales en textos de interacciones cotidianas en línea. Sin embargo, no profundizan en otros aspectos claves que abordaremos, como la influencia de la longitud del texto y otros factores contextuales en la aceptación general de las publicaciones.

Osama y Bsher [3] introdujeron una perspectiva diferente al investigar la generación automática de comentarios con razonamiento moral explícito mediante transformers como BART y T5. Este enfoque mostró que es posible producir contenido automatizado moralmente fundamentado con resultados convincentes desde el punto de vista humano. Aunque nuestro proyecto no contempla la generación de comentarios, este trabajo resalta la complejidad de captar matices éticos mediante técnicas avanzadas de NLP, enfatizando la necesidad de entender profundamente las señales implícitas y explícitas presentes en el contenido generado por los usuarios.

Strathern *et al.* [4] aportan una perspectiva complementaria al estudiar cómo surgen las .[ex]plosiones de indignación moral"(firestorms) en Twitter. Su estudio analizó 21 casos de indignación colectiva, desarrollando métodos para detectar sistemáticamente estos cambios mediante señales lingüísticas. Los autores encontraron patrones significativos, como la disminución en el uso del pronombre "yo" y el incremento de negatividad durante estos episodios. Utilizando técnicas de detección de puntos de cambio, lograron identificar el inicio de estas explosiones aproximadamente media hora antes de su manifestación evidente, demostrando que los cambios en el uso del lenguaje en comentarios textuales pueden proporcionar información sobre el comportamiento cambiante y la perspectiva cambiante.

Preniqi *et al.* [5] desarrollaron MoralBERT, un modelo de lenguaje especializado para capturar valores morales en

discursos sociales basado en la Teoría de los Fundamentos Morales. Utilizando datasets heterogéneos de Twitter, Reddit y Facebook, implementaron tanto entrenamiento agregado como adversarial de dominio para mejorar la generalización entre plataformas. Sus resultados mostraron que el marco propuesto logra una puntuación F1 promedio entre 11 % y 32 % más alta que enfoques tradicionales, y particularmente relevante para nuestro trabajo es su hallazgo de que modelos especializados de tamaño moderado pueden competir con LLMs gigantes para detectar valores morales, sugiriendo que un enfoque similar podría ser efectivo para la clasificación automática de veredictos en AITA sin requerir recursos computacionales excesivos.

III. Materiales

IV. Métodos

V. Resultados

VI. Discusión y Conclusión

Referencias

[1] N. Botzer, S. Gu, and T. Weninger, "Analysis of moral judgment on reddit," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 3, pp. 947–957, 2023.

[2] A. Alhassan, J. Zhang, and V. Schlegel, "'am i the bad one'? predicting the moral judgement of the crowd using pre–trained language models," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 267–276.

[3] M. Osama and O. Bsher, "Aita generating moral judgements of the crowd with reasoning," *arXiv preprint arXiv:2305.17466*, 2023.

[4] W. Strathern, M. Schoenfeld, R. Ghawi, and J. Pfeffer, "Against the others! detecting moral outrage in social media networks," *arXiv preprint arXiv:2010.07237v1*, 2020.

[5] V. Preniqi, I. Ghinassi, J. Ive, C. Saitis, and K. Kalimeri, "Moralbert: A fine-tuned language model for capturing moral values in social discussions," in *International Conference on Information Technology for Social Good (GoodIT '24)*. ACM, 2024.