Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data

Maryam Al Emadi, Wajdi Zaghouani
Hamad Bin Khalifa University, Qatar
{mmalemadi, wzaghouani}@hbku.edu.qa

Abstract

Freedom of speech on online social media platforms, often comes with the cost of hate speech production. Hate speech can be very harmful to the peace and development of societies as they bring about conflict and encourage crime. To regulate the hate speech content, moderators and annotators are employed. In our research, we look at the effects of prolonged exposure to hate speech on the mental and physical health of these annotators, as well as researchers with work revolving around the topic of hate speech. Through the methodology of analyzing literature, we found that prolonged exposure to hate speech does mentally and physically impact annotators and researchers in this field. We also propose solutions to reduce these negative impacts such as providing mental health services, fair labor practices, psychological assessments and interventions, as well as developing AI to assist in the process of hate speech detection.

Keywords : Hate Speech, Violent content, Moderators, Annotators, Data Annotation, Mental Health consequences

1. Introduction

“Warning! Today’s presentation contains harmful and toxic materials that are offensive.” This is what appeared on the screen of a researcher when he wanted to present his research about hate speech detection on social media platforms. Hate speech is any verbal attack against a certain group of people with a specific characteristic such as gender, race, ethnic group, religion, or political preference (Dreißigacker et al., 2024).

With the rise in the use of social media platforms, the generation and creation of abusive and hateful content such as texts, pictures, videos, or memes is evident as content creation on these platforms has gained great freedom (Roberts, 2016). These contents are prolonged as they stay forever on the platforms (Oksanen et al., 2021) which increases their consequences on individuals and causes societal implications. Consequently, researchers aim to mitigate the amount of toxicity in these mediums and create a safe place to share different beliefs and ideas. As a result of some flexible policies and the failure of machine learning to mitigate them, harmful content is being posted on a day-to-day basis by users. To reduce the amount of harmful content, tech companies need to rely on human decisions rather than completely depending on machine learning. In response to the spread of hate speech, tech companies have relied on employing content moderators from low-wage countries (Gillespie, 2018) to continually screen the user-generated content (UGC) posted on social media and decide whether they comply with the platforms’ policies and rules or not (Roberts, 2016).

Content moderators play an important role in maintaining digital civility. (Gilliespie, 2018) They get exposed to hate and violent content for long hours daily to identify the harmful content and decide whether specific content complies with the platform’s policy or not, or whether it is acceptable or not (Roberts, 2016). However, the process of moderation itself which exposes the moderators who work as the “gatekeepers of digital civility” to a barrage of disturbing material leads to psychological and emotional consequences. (Newton, 2019) Just like individuals who are victims of hate speech on social media, or even more, moderators face psychological and emotional consequences that can both affect their mental and physical well-being.

Content moderators confront a huge number of challenges such as the prolonged exposure to hate speech, violent content, and other forms of harmful content. This exposure can take a toll on their mental health and physical well-being. Psychological damage, post-traumatic stress disorder (PTSD), anxiety, depression, and insomnia are all effects of long-term exposure to harmful and abusive UGC (Das et al., 2020).

Acknowledging these consequences, researchers like (Das, Dang, & Lease, 2020) suggested a way of getting accurate decisions from moderators in mitigating harmful content in social media and complying EU Service Digital Act and considering the risks on moderators. They emphasize the importance of blurring the contents to reduce the negative effects on moderators’ welfare. However, this approach cannot ensure accurate decisions all the time. Therefore, other solutions tended to test interventions such as gray scaling to achieve the same goal of minimizing emotional impact (D'cruz, Noronha, 2020), which was effective to a certain limit only. Other solutions were imposed such as limiting the time of the exposure, frequent breaks, rotation of duties, on-cite psychological support, and interdisciplinary collaboration between government, tech companies, and mental health experts to set rules that mitigate harm.

Although many research studies were established to discuss the implications of abusive content on individuals, society, and moderators, less is known about the experiences of annotators and the implications of hate speech content on them. Researchers in the field of hate speech detection studies mostly think about ways to detect hate speech, and how to get accurate, transparent, and unbiased results to build ethical datasets to mitigate hate speech on social media. They also discuss the issue of the way annotators perceive toxicity on social media and how their different characteristics influence their decisions (Sap et al., 2022; Waseem, 2016). However, they do not see or acknowledge the emotional consequences on the annotators or themselves as researchers in the field of hate speech detection.

Annotators are a group of individuals who spend their days labeling and tagging hateful content such as videos, photos, memes, and texts for research purposes and for “good” (Kudan, 2022). Unlike moderators, the annotators’ job is more difficult as they must read and see each content carefully to be able to label them with different labels, to train machine learning algorithms to detect hate speech and other abusive content (Kudan, 2022). Therefore, the nature of their work leads to more harm to their mental and physical health. The task of annotation, particularly when it involves labeling harmful and offensive content, carries with it a profound psychological toll that merits closer examination.

Consequently, annotators are mostly expected to suffer from vicarious trauma. This phenomenon occurs when individuals are indirectly exposed to traumatic material through their work, leading to symptoms that mirror those experienced by direct trauma survivors (Pearlman & Saakvitne, 1995). For data annotators, the daily confrontation with content depicting graphic violence, hate speech, sexual abuse, and other forms of human cruelty can lead to a host of distressing symptoms, including intrusive thoughts, hyperarousal, and avoidance behaviors, which are hallmark indicators of Post-Traumatic Stress Disorder (PTSD) (Craig & Sprang, 2010).

In the development and annotation of datasets aimed at detecting hate speech, the authors of this paper have faced the significant challenge of being exposed to hate speech content. This exposure was an essential yet challenging part of their work while curating datasets for various studies. For instance, in their work on the "UPV at the Arabic Hate Speech 2022 Shared Task" (De Paula et al., 2022), they analyzed offensive language and hate speech using transformers and ensemble models. Their subsequent research on hate speech detection in Arabic languages further underscores the complexity of this issue (Magnossão de Paula et al., 2023). Additionally, the creation of a multi-label hate speech annotated Arabic dataset highlighted the nuanced aspects of hate speech across different contexts (Zaghouani et al., 2024). Their collaborative efforts extended to developing the MARASTA corpus, focusing on multi-dialectal Arabic cross-domain stance (Charfi et al., 2024), and analyzing Facebook comments to gather insights on stance, sentiment, and emotion in response to Tunisia's July 25 measures (Laabar & Zaghouani, 2024).

Moreover, constant exposure to these types of content compounds the risk, creating an environment where adequate psychological protection seems virtually impossible (D'cruz & Noronha, 2020). The resulting emotional numbing, a defense mechanism against overwhelming distress, further complicates the annotators' ability to disengage from the trauma of their work, impacting their personal lives and relationships.

Furthermore, the stigma associated with discussing mental health issues, particularly in professional contexts, can deter annotators from seeking the help they need. This silence perpetuates a cycle of suffering, with many feelings isolated in their experiences and uncertain of where to turn for support (Rosenbaum et al., 2018).

Highlighting these challenges faced by annotators in addressing hate speech on social media platforms, the need for greater awareness for those employees who are constantly engaged in such work is needed. As well as implementing effective strategies to minimize the psychological and physiological harm impacts imposed on annotators.

Despite the serious consequences and implications of data annotation on the annotators’ well-being, research addressing this issue is not found. This calls us to establish our comprehensive study titled “Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data”.

In this study, we will not only focus on the impact of hate speech on annotators, but rather on researchers as well whose lives are evolving around hate speech related topics as they constantly read about them.

2. Methodology

Our research begins by thoroughly and comprehensively exploring the essential responsibilities shouldered by annotators in the process of hate speech annotation. This investigation necessitates an in-depth review of existing literature pertaining to the roles and obligations of annotators engaged in tasks related to hate speech detection. Through an exhaustive examination of available scholarly works, our objective is to grasp the guidelines and protocols that are instituted for annotators prior to the commencement of the annotation process.

More specifically, our focus lies in elucidating the guidelines imparted to annotators to ensure the accurate labeling and categorization of hate speech content. This endeavor involves a thorough analysis of the training methodologies employed to instill adherence to these guidelines among annotators, as well as the mechanisms for assigning suitable labels to the annotated data. Additionally, we explore the impediments faced by annotators in adhering to these guidelines, particularly within the confines of the time constraints imposed for completing annotation tasks.

By amalgamating insights gleaned from the literature, we discern pivotal tasks undertaken by annotators, the nature of their training, and the repercussions of time constraints on the quality of annotated data. This methodological approach facilitates a comprehensive understanding of the factors that influence annotators' performance and the emotional toll incurred in the process of hate speech annotation.

Annotators often operate within demanding and stressful environments, compelled to annotate vast quantities of hate speech content within specified time frames. Each piece of content necessitates meticulous review and labeling, exerting significant cognitive effort. Consequently, we investigate the physical and digital work environments of annotators, including their work schedules, breaks, technological tools utilized, and the support systems provided by their employers to mitigate associated challenges.

Furthermore, our study evaluates the impact of imposed deadlines and productivity targets on annotators' well-being, seeking to strike a balance between the quality and quantity of their work. Moreover, we explore the effects of exposure to harmful content on the mental health of both researchers and annotators involved in such endeavors. Through the administration of a survey targeting both cohorts, we aim to quantify the impact of prolonged exposure to harmful content on their mental well-being and elucidate the diverse ramifications that impede their daily activities.

The paper also raises the ethical questions about the psychological toll of data annotation and the regulatory complexities that are continuously evolving. It discusses the moral imperative to protect the well-being of these workers, necessitating the implementation of comprehensive mental health support systems, regular psychological assessments, and accessible interventions designed to address the unique challenges faced by annotators (Roberts, 2016). It aims to explore the partnership with mental health organizations to provide support for annotators with issues related to their constant exposure to such harmful content.

Furthermore, fostering a workplace culture that prioritizes mental health, encourages open discussions about emotional well-being, and actively destigmatizes mental health issues is crucial. Such measures not only support annotators in managing the psychological impacts of their work but also contribute to a more compassionate and ethical approach to data annotation (Armstrong et al., 2018).

In this paper, we provide policies that can actively destigmatize mental health issues among employees as well as addressing the ethical considerations surrounding the work of annotators which aims to protect their well-being. To achieve that, the study engages in a broader discussion about the responsibility of tech companies, policymaker, and the global community in recognizing the psychological and emotional consequences of such work on annotators and how to support their mental health.

3. Discussion

In suggesting a solution to mitigate harm to annotators, the study recommends that tech companies provide fair labor practices and on-site mental health support, transparency, and accountability about the nature of data annotation, and developing ethical AI technologies to assist annotators in data annotation. Tech companies have an inherent ethical responsibility to ensure that the working conditions of data annotators meet high standards of fairness and respect for human dignity. Given the psychologically taxing nature of annotating harmful and offensive content, companies must go beyond traditional labor practices to implement comprehensive mental health support systems. These systems should include access to psychological counseling, mental health days, and programs designed to mitigate the impact of vicarious trauma (Armstrong et al., 2018; Craig & Sprang, 2010).

Moreover, the ethical obligation extends to providing a work environment that fosters open communication about mental health challenges without fear of stigma or reprisal. Implementing regular mental health assessments and training for managers and supervisors on recognizing and addressing signs of psychological distress among their teams can create a supportive atmosphere conducive to employee well-being (D'cruz & Noronha, 2020).

Transparency about the nature of data annotation work and the potential psychological risks associated with it is another critical ethical responsibility. Tech companies must ensure that annotators are fully informed about the content they will encounter and understand the available support mechanisms. This transparency should also extend to the public and regulatory bodies, with companies openly disclosing their practices and the measures they take to safeguard annotator well-being (Roberts, 2016).

Accountability mechanisms, such as independent audits of working conditions and mental health support provisions, can further ensure that companies adhere to their ethical obligations. These measures not only protect annotators but also build trust among stakeholders, including users, regulators, and the broader public (Gillespie, 2018).

The development of AI systems for content moderation raises profound ethical questions about the reliance on human-annotated data. Tech companies must grapple with the dual imperatives of advancing technological innovation and ensuring that this progress does not come at the expense of human well-being. Ethical AI development practices require a commitment to minimizing the reliance on human annotation of harmful content wherever possible, exploring alternative methods that reduce exposure to such content, and investing in research aimed at improving AI's ability to understand context and nuance without extensive human input (Gorwa, Binns, & Katzenbach, 2020).

The ethical obligations of tech companies in the realm of data annotation are multifaceted and complex. As the digital world continues to evolve, the need for responsible, ethical practices in the development and maintenance of AI systems becomes increasingly paramount. By prioritizing the health and well-being of data annotators, fostering transparency and accountability, and pursuing ethical AI development, tech companies can navigate the challenges of data annotation while upholding their moral responsibilities to their employees and society at large.

The paper also touches on legal and regulatory considerations for tech companies in data annotation. It suggests evolving legal frameworks necessitate a proactive approach to compliance. This involves not only implementing robust content moderation systems but also ensuring that the processes of data annotation — a critical component in the development of these systems — align with legal standards regarding worker rights and data privacy.

Furthermore, due to the regular harm that is imposed on annotators, the study believes that legal considerations extend beyond compliance to encompass the ethical implications of data annotation work, particularly regarding the protection of annotators from harm.

In addition to adhering to legal requirements, there is a growing call for tech companies to engage in self-regulation and the development of industry standards for data annotation. This involves creating transparent, accountable practices that ensure the ethical treatment of annotators and the responsible development of content moderation technologies. Industry standards could include guidelines for annotator well-being, data privacy, and the accuracy and fairness of annotated datasets used to train AI systems (Roberts, 2016).

4. Conclusion

The task of data annotation is a difficult task that has prolonged consequences, which emerges as a poignant emblem of the hidden costs associated with building safer online environments. It underscores a significant yet unappreciated human cost in the effort to create a safer online environment with less harm to its users. Recognizing the long-lasting effects on annotators mental health is paramount to developing sustainable and ethical practices in the field of data annotation. The exploration of this critical yet often overlooked aspect of digital infrastructure reveals profound ethical, psychological, and regulatory challenges that demand our immediate attention and action.

The psychological toll on data annotators, highlighted through the lens of vicarious trauma and the elevated risks of mental health issues such as PTSD, anxiety, and depression, underscores a pressing moral imperative (Craig & Sprang, 2010). These individuals, who serve as the first line of defense against the proliferation of harmful content, endure significant emotional and psychological strain, necessitating a robust framework of support (D'cruz & Noronha, 2020). The ethical obligations of tech companies in this context extend beyond mere compliance with legal standards to encompass a duty of care that honors the humanity and dignity of each annotator (Roberts, 2016). Which includes providing comprehensive mental health support, fostering a workplace culture that prioritizes their well-being and implementing fair labor practices.

Furthermore, the evolving legal and regulatory landscape presents both challenges and opportunities for safeguarding the well-being of data annotators. Legislation such as the Digital Services Act in Europe represents a critical step towards holding tech companies accountable for the content on their platforms and, by extension, for the conditions under which data annotators work. However, these regulations must be carefully crafted to ensure they do not inadvertently exacerbate the pressures on annotators, instead fostering an environment that prioritizes their mental health and well-being (Keller, 2020). Therefore, an interdisciplinary collaboration needs to be established between tech companies, policymakers, and mental health experts to come up with regulations that can effectively protect the public users and the annotators.

Looking ahead, the future of data annotation and content moderation lies in the delicate balance between leveraging technological advancements and preserving the essential human element. The potential of artificial intelligence and machine learning offer promising avenues for reducing the burden on human annotators by automating aspects of content moderation. However, these technologies are not a panacea. The nuances of human communication and the contextual understanding necessary for evaluating content underscore the irreplaceable value of human judgment (Gorwa, Binns, & Katzenbach, 2020). Therefore, those innovations should aim to support not replace the critical work of annotators, ensuring that technologies enhance rather than diminish human well-being.

In conclusion, the discourse surrounding the data annotation of harmful and offensive content invites us to reflect on the broader implications of our digital age. It compels us to consider not only the technological and economic dimensions but also the human cost of creating and maintaining digital spaces. As we navigate this complex terrain, we must forge a path that respects the contributions of data annotators, addresses the ethical challenges inherent in their work, and envisions a future where technology serves to enhance human well-being. The integrity and safety of our digital spaces depend on our collective ability to recognize, support, and protect those who labor in the shadows to keep them clean.

5. Limitations and Future Work

Our study was significantly constrained by both temporal limitations and ethical considerations. The sensitive nature of our research topic, which involved examining the potential impacts of sensitive data annotation on the mental and physical well-being of annotators, necessitated a careful approach to participant engagement. However, the lack of Institutional Review Board (IRB) approval emerged as a major impediment, limiting our ability to conduct in-depth interviews and, consequently, restricting the scope of our investigation. The unavailability of ethical clearance precluded the collection of direct testimonies from annotators, thus curtailing our understanding of the effects of their work.

The primary reason for the absence of IRB clearance was the stringent time constraints under which our study operated. The time-sensitive nature of the research process did not allow for the completion of the extensive and rigorous IRB approval procedures, thereby hindering our capacity to engage directly with annotators through interviews or surveys.

This limitation not only highlights the ethical complexities associated with research on mental health topics but also stresses the necessity for future studies to meticulously navigate the ethical review process. The experience underscores the critical importance of obtaining IRB approval to ensure a comprehensive exploration of the research subject.

To address the aforementioned constraints and augment the methodological rigor of our subsequent inquiries, securing Institutional Review Board (IRB) clearance will be our foremost priority. Achieving this will facilitate the execution of in-depth, qualitative interviews with a carefully selected cohort of data annotators. This strategic approach is intended to yield a more nuanced understanding of the intricacies and repercussions inherent in data annotation processes.

We posit that conducting individualized, qualitative interviews will be essential for eliciting profound insights into the annotators' personal experiences, their strategies for managing work-related stress, and their perceptions of support from their employers. Such an investigative framework will enable the collection of detailed personal narratives, thereby illuminating the experiences of those involved in sensitive data annotation and the subsequent effects on their mental and physical well-being. This methodological enhancement is expected to significantly contribute to the body of knowledge concerning the occupational health aspects of data annotation work.

6. Acknowledgments

This publication was made possible through the financial support of NPRP grant NPRP13S-0206-200281 from the Qatar National Research Fund and the Qatar Research Development and Innovation Council (QRDI). The findings and opinions expressed herein are solely the responsibility of the authors.

7. Bibliographical References

Armstrong, C., Davis, J., Holder, K., Knowles, K., & Patel, K. (2018). The Trauma Floor: The Secret Lives of Facebook Moderators in America. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Armstrong, G., Blashki, G., Jorm, A. F., Kitchener, B. A., & Crisp, D. A. (2018). An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. Asian Journal of Psychiatry, 37, 25-31. https://doi.org/10.1016/j.ajp.2018.07.017

Charfi, A., Ben-Sghaier, M., Atalla, A., Akasheh, R., Al-Emadi, S., & Zaghouani, W. (2024). MARASTA: A multi-dialectal Arabic cross-domain stance corpus. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

De Paula, A. F. M., Rosso, P., Bensalem, I., & Zaghouani, W. (2022). UPV at the Arabic hate speech 2022 shared task: Offensive language and hate speech detection using transformers and ensemble models. In Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection (pp. 181-185).

Craig, C. D., & Sprang, G. (2010). Compassion satisfaction, compassion fatigue, and burnout in a national sample of trauma treatment therapists. Anxiety, Stress & Coping, 23(3), 319–339. https://doi.org/10.1080/10615800903085818

Das, A., Dang, V., & Lease, M. (2020). Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. In Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019). https://www.ischool.utexas.edu/~ml/papers/das_hcomp20.pdf

D'cruz, P., & Noronha, E. (2020). Navigating the extended reaches of online harm: The experience of online content moderators. Work, Employment and Society, 34(3), 456-475. https://doi.org/10.1177/0950017020914877

D'cruz, R., & Noronha, E. (2020). Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 8, 128–137. https://ojs.aaai.org/index.php/HCOMP/article/view/5270

Dreißigacker, A., Müller, P., Isenhardt, A., & Schemmel, J. (2024, February 10). Online hate speech victimization: consequences for victims' feelings of insecurity. Crime Science. https://doi.org/10.1186/s40163-024-00204-y

European Commission. (2020). Digital Services Act: Ensuring a safe and accountable online environment. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Ghosh, D., & Guha, R. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 112–117). Association for Computational Linguistics. https://aclanthology.org/W16-5618.pdf

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape social media. Yale University Press.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951720919770

Keller, D. (2020). Internet platforms: Observations on speech, danger, and money. Hoover Institution, Aegis Series Paper No. 1907.

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. Harvard Law Review, 131, 1598-1670.

Koops, B. J. (2014). The trouble with European data protection law. International Data Privacy Law, 4(4), 250-261. https://doi.org/10.1093/idpl/ipu023

Kudan, G. (2022). Unpublished manuscript.

Kudan. (2022, December 8). The role of a data annotator in machine learning. Toloka. Retrieved from https://toloka.ai/blog/what-does-a-data-annotator-do/

Laabar, S., & Zaghouani, W. (2024). Multi-dimensional insights: Annotated dataset of stance, sentiment, and emotion in Facebook comments on Tunisia's July 25 measures. In Proceedings of the Second Workshop on Natural Language Processing for Political Sciences co-located with the 2024 International Conference on Computational Linguistics, Language Resources and Evaluation.

Magnossão de Paula, A. F., Bensalem, I., Rosso, P., & Zaghouani, W. (2023). Transformers and ensemble methods: A solution for hate speech detection in Arabic languages. arXiv preprint arXiv:2303.

Newell, J. M., MacNeil, G. A. (2010). Professional burnout, vicarious trauma, secondary traumatic stress, and compassion fatigue. Best Practices in Mental Health, 6(2), 57-68.

Newton, C. (2019). The toll of the online content moderator. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Newton, C. (2019). The Trauma Floor: The secret lives of Facebook moderators in America. The Verge. Retrieved from https://www.theverge.com/2019/2/25/18229714/co gnizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Newton, C. (2020). Inside the traumatic life of a Facebook moderator. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Newton, C. (2020, January 28). What tech companies should do about their content moderators' PTSD. The Verge. https://www.theverge.com/interface/2020/1/28/21082642/content-moderator-ptsd-facebook-youtube-accenture-solutions

Oksanen, A., Celuch, M., Latikka, R., Oksa, R., & Savela, N. (2021, November 23). Hate and harassment in academia: the rising concern of the online environment. Higher Education. https://doi.org/10.1007/s10734-021-00787-4

Pearlman, L. A., & Saakvitne, K. W. (1995). Trauma and the therapist: Countertransference and vicarious traumatization in psychotherapy with incest survivors. W. W. Norton.

Pearlman, L. A., & Saakvitne, K. W. (1995). Trauma and the Therapist: Countertransference and Vicarious Traumatization in Psychotherapy with Incest Survivors. W.W. Norton & Company.

Roberts, S. T. (2016). Commercial Content Moderation: Digital Laborers' Dirty Work. In The Intersectional Internet: Race, Sex, Class, and Culture Online (pp. 147-159). Peter Lang.

Roberts, S. T. (2016). Content Moderation. Social Media + Society, 2(2), 2056305116644493. https://www.academia.edu/31637827/Content_Moderation

Robertson, A. (2019, June 19). BODIES IN SEATS At Facebook's worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives. The Verge. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-ta

Rosenbaum, L., et al. (2018). Caring too much: Compassion fatigue and mental Smith, J. K. (Year). Title of the article. Journal Name, Volume(Issue), Page range. Retrieved from URL

Rosenbaum, L., Shieber, S., & McNally, R. J. (2018). The National Stressful Events Survey PTSD Short Scale (NSESSS). Psychological Assessment, 30(3), 405–415. https://doi.org/10.1037/pas0000562

Sap, Swayamdipta, Vianna, Zhou, Choi, & Smith, A. (2022). Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. https://aclanthology.org/2022.naacl-main.431.pdf

Waseem. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. https://aclanthology.org/W16-5618.pdf European Commission. (2020). Code of Practice on Disinformation. https://ec.europa.eu/digital-single-market/en/code-practice-disinformation

Zaghouani, W., Mubarak, H., & Biswas, M. R. (2024). So hateful! Building a multi-label hate speech annotated Arabic dataset. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).