Investigating the Hurtfulness of Misogynistic Tweets Across Professions

Alessio Cascione[1,*], Aldo Cerulli[2,*], Marta Marchiori Manerba[1] and Lucia C. Passaro[1]

[1]Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, Pisa, 56127, Italy
[2]Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Via Santa Maria 36, Pisa, 56126, Italy

Abstract

With the increasing popularity of social media platforms, the dissemination of misogynistic content has become more prevalent and challenging to address. In this work, we investigate the phenomenon of online misogyny through the lens of hurtfulness, qualifying its different manifestations with respect to the profession of offended women. By combining manual and automatic annotation, we find that specific types of misogynistic attacks are more intensely directed toward professional figures: derailing discourse mainly targets authors and cultural figures, while dominance-oriented speech and sexual harassment primarily target politicians and athletes. Additionally, hurtfulness and emotive lexica are leveraged for assigning hurtfulness scores to social media posts. Our analysis shows these scores align with the profession-based distribution of misogynistic speech, highlighting the targeted nature of the attacks.

Keywords

Abusive Language, Automatic Misogyny Detection, NLP

1. Introduction

Misogyny is a radical manifestation of sexism directed primarily toward the female gender, which persists in various forms in our society, especially on social media platforms [1, 2, 3, 4]. Historically, women have faced numerous barriers that limited their access to certain professions and subjected them to offenses related to their work [5]. Perpetuating inequality serves as a breeding ground for misogyny. In our work, we focus on automated misogyny detection, investigating whether different professional roles trigger varying nuances of hurtfulness across social media posts. We aim to fill a gap in a field that has not yet addressed fine-grained forms of online misogyny [6].

While various works have contributed to misogyny detection through datasets and evaluation tasks [7, 8, 9, 10, 11, 12, 13] and to the qualitative study of misogyny targeting specific individuals [14, 15, 16, 17, 18], to the best of our knowledge there are no works that simultaneously explore from a data-driven perspective the instantiation of misogyny addressed to women engaged in particular professions.

2. Data Expansion and Labeling Workflow

We take as a starting point the EVALITA 2018 AMI dataset [7], which encompasses ground-truth information on five categories of misogyny: *derailing*, *discredit*, *dominance*, *sexual harassment*, and *stereotype*. We enrich the subsection of AMI for which it was possible to infer the victims' professions (i.e., 380 tweets) with the manual annotation of professions grouped into four classes, namely 'artist', 'author', 'athlete', and 'politician'. We exploit Wikidata as taxonomy.

Discovery Science - Late Breaking Contributions 2024

*Corresponding authors. These authors contributed equally.
alessio.cascione@phd.unipi.it (A. Cascione); a.cerulli1@studenti.unipi.it (A. Cerulli); marta.marchiori@phd.unipi.it (M. Marchiori Manerba); lucia.passaro@unipi.it (L. C. Passaro)
https://martamarchiori.github.io/ (M. Marchiori Manerba); https://luciacpassaro.github.io/ (L. C. Passaro)
0009-0003-5043-5942 (A. Cascione); 0000-0003-2251-1824 (M. Marchiori Manerba); 0000-0003-4934-5344 (L. C. Passaro)

Moreover, we expand the dataset by crawling new tweets directed to famous women with a known profession. This crawling process results in 760 tweets with ground-truth information on professions, which we refer to as the PRF dataset. Since the PRF dataset lacks information on the type of misogyny, we use a BERTweet [19] model fine-tuned on AMI (Weighted Avg. F1 of .704 on the test set) to classify the category of misogyny automatically. Overall, we conduct our study on 1140 tweets with both misogyny and professional information.

3. Findings and Discussion

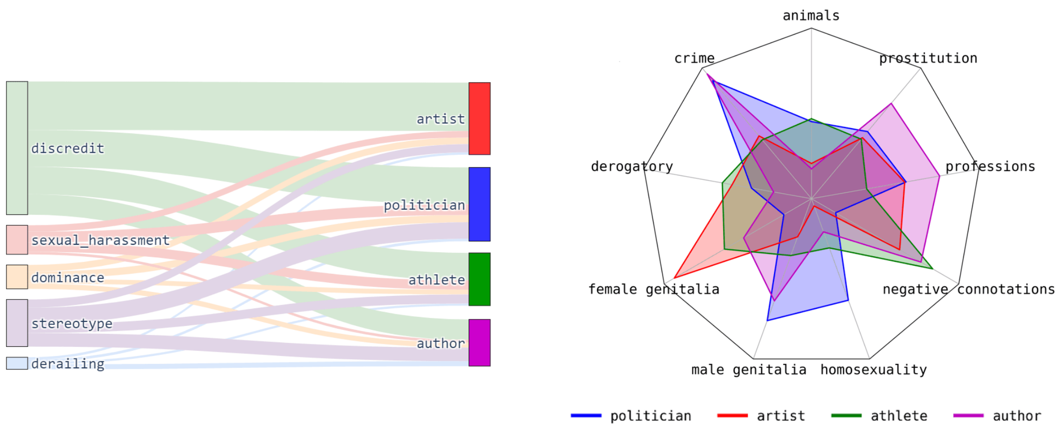

Figure 1: Left: Misogyny types and professions. Right: Hurtfulness scores of tweets.

Our findings show distinct patterns in the distribution of types of misogynistic speech concerning professions (Fig. 1 Left). To further analyze the lexicon of misogynistic content, we leverage a hurtfulness lexicon based on ItEM [20] using 9 categories from HurtLex [21] as seed words (Fig. 1 Right).

Overall, we find that *derailing* misogyny primarily targets authors and intellectuals, *dominance* and *stereotype/objectification* predominantly attack politicians, while *sexual harassment/threats of violence* is directed to politicians and athletes. As for the average hurtfulness scores, we notice that politicians are mainly targeted with insults related to crime, homosexuality, and male genitalia, consistently with *sexual harassment/threats of violence*. Artists present a peak in abusive language referring to female genitalia, while for athletes we notice a more balanced misogyny type. Authors seem to be mainly targeted with hate speech addressing crime and professions as main topics, consistent with the fact that the types of misogyny mostly faced by this profession are *derailing* and *stereotype*

Acknowledgments

This research was funded by PNRR-PE00000013 "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" under NextGeneration EU, ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the eXplanation of AI decision making* under Horizon 2020, and PRIN 2022 PIANO (Personalized Interventions Against Online Toxicity) project, CUP B53D23013290006.

References

[1] E. A. Jane, 'Back to the kitchen, cunt': Speaking the unspeakable about online misogyny, Continuum 28 (2014) 558–570.

[2] D. Ging, E. Siapera, Special issue on online misogyny, Feminist media studies 18 (2018) 515–524.

[3] M. E. David, Reclaiming feminism: Challenging everyday misogyny, Policy Press, 2016.

[4] C. Tileagă, Communicating misogyny: An interdisciplinary research agenda for social psychology, Social and Personality Psychology Compass 13 (2019) e12491.
[5] J. Marques, Exploring gender at work, Springer, 2021.
[6] L. Fontanella, B. Chulvi, E. Ignazzi, A. Sarra, A. Tontodimamma, How do we study misogyny in the digital age? A systematic literature review using a computational linguistic approach, Humanities and Social Sciences Communications 11 (2024) 1–15.
[7] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (AMI), in: Tommaso Caselli and Nicole Novielli and Viviana Patti and Paolo Rosso (Ed.), Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2263/paper009.pdf.
[8] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.
[9] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1425–1447. URL: https://aclanthology.org/2020.semeval-1.188. doi:10.18653/v1/2020.semeval-1.188.
[10] D. Felmlee, P. Inara Rodis, A. Zhang, Sexist slurs: Reinforcing feminine stereotypes online, Sex Roles 83 (2020) 16–28.
[11] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, V. Varma, Multi-label categorization of accounts of sexism using a neural framework, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1642–1652. URL: https://aclanthology.org/D19-1174. doi:10.18653/v1/D19-1174.
[12] P. Chiril, F. Benamara, V. Moriceau, "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification?, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2833–2844. URL: https://aclanthology.org/2021.findings-emnlp.242. doi:10.18653/v1/2021.findings-emnlp.242.
[13] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples, in: C. Budak, M. Cha, D. Quercia, L. Xie (Eds.), Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021, AAAI Press, 2021, pp. 573–584.
[14] D. Silva-Paredes, D. Ibarra Herrera, Resisting anti-democratic values with misogynistic abuse against a chilean right-wing politician on twitter: The# camilapeluche incident, Discourse & Communication 16 (2022) 426–444.
[15] E. B. Phipps, F. Montgomery, "Only YOU Can Prevent This Nightmare, America": Nancy Pelosi As the Monstrous-Feminine in Donald Trump's YouTube Attacks, Women's Studies in Communication 45 (2022) 316–337.
[16] J. Ritchie, Creating a monster: Online media constructions of Hillary Clinton during the democratic primary campaign, 2007–8, Feminist Media Studies 13 (2013) 102–119.
[17] N. Saluja, N. Thilaka, Women leaders and digital communication: Gender stereotyping of female politicians on twitter, Journal of Content, Community & Communication 7 (2021) 227–241.
[18] S. Ghaffari, Discourses of celebrities on instagram: digital femininity, self-representation and hate speech, in: Social Media Critical Discourse Studies, Routledge, 2023, pp. 43–60.
[19] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, in:

Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 9–14.

[20] L. C. Passaro, A. Lenci, Evaluating context selection strategies to build emotive vector space models, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016, European Language Resources Association (ELRA), 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/637.html.

[21] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018, volume 2253 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2253/paper49.pdf.