# Women's Professions and Targeted Misogyny Online

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 paper, we investigate the phenomenon of online misogyny on Twitter through the lens of hurtfulness, qualifying its different manifestation in English tweets considering the profession of the targets of misogynistic attacks. By leveraging manual annotation and a BERTweet model trained for fine-grained misogyny identification, we find that specific types of misogynistic speech are more intensely directed towards particular professions. For example, derailing discourse predominantly targets authors and cultural figures, while dominance-oriented speech and sexual harassment are mainly directed at politicians and athletes. Additionally, we use the HurtLex lexicon and ItEM to assign hurtfulness scores to tweets based on different hate speech categories. Our analysis reveals that these scores align with the profession-based distribution of misogynistic speech, highlighting the targeted nature of such attacks.

**Keywords**

Abusive Language, Online Misogyny, Hurtfulness

## 1. Introduction

Misogyny is a radical manifestation of sexism directed toward the female gender, which becomes subject of hatred. Its effects are widespread and systematic, bearing severe both social and individual consequences, such verbal and physical violence, rape and femicide. Indeed, misogyny, prejudice, and contempt towards women continue to persist in various forms in our society. While overt acts of discrimination and sexism have received attention, it is crucial to acknowledge that misogyny often manifests in subtle and nuanced ways [1, 2]. Moreover, with the increasing popularity of social media platforms, the dissemination of misogynistic content has become more prevalent and challenging to address [3, 4].

From a socio-historical perspective, women have faced numerous barriers that limited their access to certain professions, hindered their career progression, and subjected them to belittlement and offense related to their work [5]. These gendered biases not only perpetuate inequality but also serve as breeding grounds for misogyny.

In this paper, we focus on automated misogyny detection, specifically investigating whether different professional roles trigger varying degrees of hurtfulness across social media posts. By examining the correlation between the profession of offended women and the prevalence of misogynistic attitudes, we aim to shed light on the extent to which misogyny is perpetuated within specific professional domains.

Fontanella et al. [6] highlight how research focusing on automatic detection of misogyny tends to show weak connections with other conceptual areas addressing different aspects of the phenomenon. The finding suggests that current research has not yet adequately addressed the fine-grained manifestations of online misogynistic attacks. Our contribution conducts novel analyses to uncover and measure misogynistic attitudes within different professional fields. Specifically, we examine how different types of misogyny are distributed across various women's professions and how the language used in misogynistic posts varies across them. To explore this relationship, we expand the English misogyny identification dataset introduced by Fersini et al. [7], known as AMI, by incorporating the professions of the women targeted. By adding professional categories to AMI, we enable novel analyses on how misogynistic attacks against women differ based on their profession. Our research is driven by the following research questions:

- RQ1 **How does misogyny distribute across professions?** We analyze women's profession according to the type of misogyny directed towards them.
- RQ2 **How does the language used in misogynistic tweets vary across different professions?** We investigate how specific hurtful expressions are directed at specific professions more frequently than others.

To address our RQs, we proceed following the work-

*CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy*

*Corresponding authors. These authors contributed equally.
a.cascione@studenti.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); 0000-0002-0877-7063 (M. Marchiori Manerba); 0000-0003-4934-534 (L. C. Passaro)

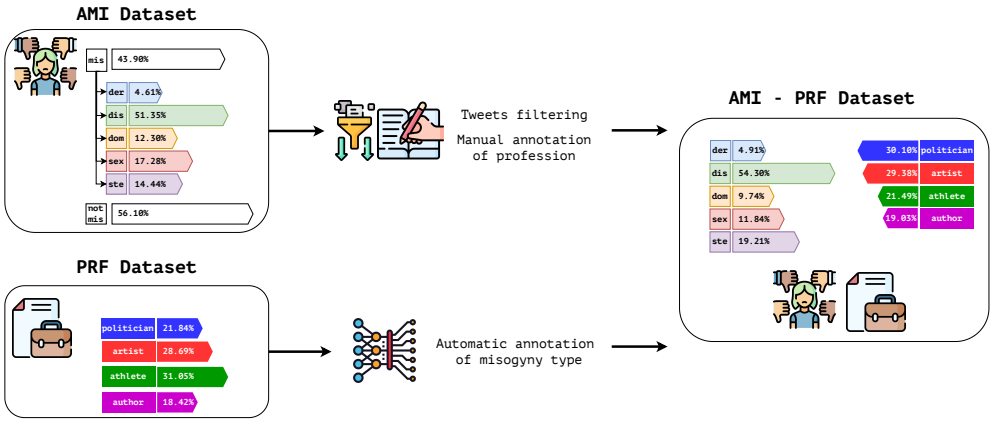


**Figure 1:** A subset of the AMI dataset, containing ground-truth misogyny annotations, is manually labeled with the professions of victims of misogynistic attacks, as detailed in Section 3. The PRF dataset, featuring professions by-design, is extracted and automatically annotated with misogyny types using a BERTweet model trained on the AMI dataset. The manually annotated AMI subset and the automatically annotated PRF dataset are then combined to form the AMI-PRF dataset. Labels distributions of each dataset are displayed in the workflow.

flow depicted in Figure 1. We begin by utilizing a subset of the AMI dataset, which contains ground-truth annotations for misogyny. This subset is manually labeled with the professions of the victims of misogynistic attacks, as detailed in Section 3.2. We then employ a misogyny classifier to automatically annotate with various types of misogyny a novel collection, the Profession (PRF) dataset, which comprises 760 tweets labeled with professions. The final step involves combining the manually annotated AMI subset with the automatically annotated PRF dataset, resulting in the AMI-PRF dataset[1]. This enriched dataset provides a resource that enables a thorough investigation of the phenomenon.

The remainder of this paper is organized as follows. Section 2 discusses previous works that closely related to ours, while Section 3 details the enrichment of the AMI dataset with professional categories. Section 4 reports the experiments conducted to answer our RQs, whereas Section 5 outlines conclusions, limitations, and future directions of the work.

## 2. Related Work

In recent years, the field of NLP has witnessed a growing interest in detecting misogyny and sexist content on social media platforms. Various works have significantly contributed to this area by publicly introducing diverse datasets and evaluation tasks tailored for misogyny detection [7, 8, 9]. Indeed, it is a pressing need to develop systems for detecting emotive [10, 11] and offensive word lexicons for harassment research [12], as highlighted by Rezvan et al. [13]. Contributing to the field of sexism categorization, Parikh et al. [14] provide a large dataset for multi-label classification of sexism. Chiril et al. [15] explore the detection of sexist hate speech, examining the relationship between gender stereotype detection and sexism classification. Similarly, Felmlee et al. [16] investigate online aggression towards women on social media platforms, focusing on the strategic nature of sexist tweets and the reinforcement of stereotypes.

Emphasizing the interaction and co-influence of social dimensions, like gender and profession, can assist in capturing complex social dynamics and informing the development of norms that promote equity and justice, as outlined by Hancock [17] and Dhamoon [18]. Specifically, previous social science research has examined hate discourse directed at specific groups of women, such as politicians and celebrities. For example, Silva-Paredes and Ibarra Herrera [19] offer a corpus-based analysis of gender-based aggression towards a Chilean right-wing female politician, while Phipps and Montgomery [20] and Ritchie [21] focus on forms of hate speech in media campaigns against Nancy Pelosi and Hillary Clinton, respectively. Specifically for tweets, Saluja and Thilaka [22] employ the Feminist Critical Discourse Theory to perform gender-specific inferences w.r.t. Twitter discourse concerning Indian political leaders. On the other hand, Ghaffari [23] analyzes 2000 user-generated posts focusing on American celebrity Lena Dunham, examining manifestations of hate and stereotypes. To the best of our knowledge, this is the first data-driven work that

[1]The dataset is accessible for research purposes by requesting it by email from the authors. To protect the identities of the affected women, we chose to omit explicit references to profiles and original tweet IDs from the dataset.

examines the relationship between women professional categories and types of misogynistic attacks on online platforms.

## 3. Data Exploration and Enrichment

In this section, we detail the construction of our novel AMI-PRF dataset.

### 3.1. AMI Dataset

We address the lack of misogynous data annotated w.r.t. victims' professions by enriching the AMI dataset[2] [7]. The dataset includes a coarse-grained distinction between *mis*ogynistic and *not-mis*ogynistic tweets, as well as a fine-grained labeling for misogynistic tweets, categorizing them into five different types of misogynistic hate speech: *der*ailing (to justify women abuse), *dis*credit (general slurring), *dom*inance (to assert men superiority), *sex*ual harassment (sexual advances and violence) and *ste*reotype (oversimplification and objectification).

We enrich AMI by adding information about the professions of the victims. This enrichment is performed through retrieving from Wikidata[3] professional figures that are subclasses of the *person* class.

Our annotation of professions include four categories, namely 'artist', 'author', 'athlete', 'politician (and activist)'. We focus on these professions as they are represented in the AMI dataset, based on the popular women referenced. Although the first two are both subclasses of *creator*, which is an immediate subclass of *person*, we keep them separate due to their different natures: the former encompasses visual and performing arts, the latter intellectual activities. On the other hand, we choose to group politicians and activists together to highlight their shared involvement in public social activities, even though they are not directly related according to Wikidata taxonomy.

As shown by Fig. 4 (Appendix A), each macro-profession initiates a potentially large set of nested sub-professions based on Wikidata *subclass of* relationship.

We leverage these professions to manually label AMI misogynistic tweets that actually refer to women. In order to produce a consistent labeling, we establish the following conventions: if the tweet refers to a famous woman, we choose the first (or unique) occupation among those appearing on her Wikidata page, tracing it back to the appropriate macro-category. This approach mitigates annotation inconsistencies by leveraging an established external resource for labeling. When such information is unavailable, we determine the professional category by examining relevant job details in the tweet content or on the profile page of the victim, if mentioned. For such cases, a collaborative approach was taken during group meetings to share general insights, ensuring that any disagreements were addressed through discussions and ultimately resolved through consensus. In absence of clues regarding the profession, the tweet is simply labeled as 'generic'.

Finally, we point out that not all tweets in the AMI dataset have women as victims. In several cases, misogynist language is used to insult men, companies or political parties. Out of 5000 AMI tweets, we initially filtered out those that were not directed at women. Among the remaining tweets, 2187 were labelled as misogynistic. However, we were able to obtain professional categories for only a subset of 380 of these tweets, highlighting the need for additional data collection.

**Table 1**
BERTweet multi-classification results on AMI test set.

| | support% | Precision | Recall | F1-score |
|---|---|---|---|---|
| *der* | 2.391% | 0.250 | 0.273 | 0.261 |
| *dis* | 30.65% | 0.626 | 0.794 | 0.700 |
| *dom* | 26.95% | 0.811 | 0.484 | 0.606 |
| *sex* | 9.565% | 0.500 | 0.773 | 0.607 |
| *ste* | 30.43% | 0.906 | 0.821 | 0.861 |
| Macro Avg, | - | 0.618 | 0.629 | 0.607 |
| Wtd. Avg. | - | 0.740 | 0.704 | **0.704** |
| Accuracy | - | - | - | 0.704 |

### 3.2. PRF Dataset

To address the issue of having only a small number of tweets annotated for both misogyny and profession, we crawl additional tweets. From the most common expressions in the misogynistic tweets of AMI, we derive a list of misogynistic keywords. For each of our target professions, we choose five representative popular women, collecting tweets containing a reference to them in the form of a hashtag, mention and/or explicit name and surname. As a result, we extract 760 tweets labeled with professions, which have been posted before the beginning of February 2023: we refer to this collection as the Profession (PRF) dataset. Since these tweets are filtered using specific keywords and are directed at popular women, we consider them inherently misogynistic, as a woman is the primary target of hate speech.

To identify the type of misogyny in PRF, we leverage BERTweet[4], a transformer-based [24] model trained on the AMI multi-classification dataset. We opt for this model since it is pre-trained on Twitter, and it achieves

[2] https://live.european-language-grid.eu/catalogue/corpus/7272
[3] https://www.wikidata.org/wiki/Wikidata:Main_Page
[4] https://github.com/VinAIResearch/BERTweet

state-of-the-art performance in Twitter sentiment analysis tasks [25]. Before training, the AMI tweets are preprocessed with a TweetNormalizer function[5] which maps emojis into text strings and substitutes user mentions and web/url links with @USER and HTTPURL placeholders. For model selection, we perform a stratified cross-validation with $k = 5$. We search for the best weight decay and learning rate in [1e-2,1e-5] and [1e-5,3e-5], respectively. For each configuration, we set 10 epochs, 500 warm up steps and a train/validation batch of 16/8. The optimal performance is achieved with a learning rate of 3e-5 and a weight decay of 1e-2. Tab. 1 shows BERTweet performances for the multi-class misogyny detection task on AMI test set, comprising 1000 tweets (460 misogynistic). For the multi-classification task, we focus only on misogynistic tweets. The evaluation metrics include Accuracy, as well as weighted and unweighted average Precision, Recall, and F1-score. We adopt this model to label our PRF dataset with types of misogyny.

**AMI-PRF Dataset** By combining the 380 tweets from AMI, having ground-truth information regarding the type of misogyny, and the PRF dataset, labeled with our trained model, we obtain 1140 tweets featuring both misogyny type and professions. Such dataset, named AMI-PRF, is leveraged to investigate the relation between misogyny and professions.

# 4. Experiments and Data Analyses

## 4.1. Misogyny Type by Profession (RQ1)

To address RQ1, we examine how different types of misogynistic speech are distributed across various professions in AMI-PRF. For each type of misogyny, we find how many tweets belonging to such class are directed towards a specific profession and qualitatively compare the results in Fig. 2.

**Discussion** We observe distinct patterns in the usage of misogynistic speech across professions: derailing discourse, which focuses on justifying women abuse and rejecting male responsibility, tends to primarily target authors compared to the other professions. This aligns with the nature of derailing speech, which seeks to rationalize mistreatment of women and deflect male accountability. Therefore, this kind of discourse can be expected to be commonly directed at public intellectuals or cultural figures. In contrast, dominance-oriented misogynistic discourse, aimed at asserting male superiority along with stereotypical negative speech, is predominantly directed at powerful figures such as politicians. This prevalence could be explained as an attempt to undermine the legitimacy and value of women holding relevant public roles. Sexual harassment is notably prevalent towards politicians and athletes, as expressions of intent to assert power over women through threats of violence.

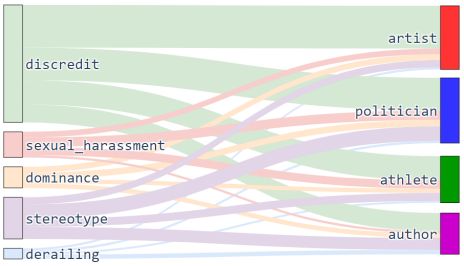


**Figure 2:** Alluvial plot depicting the relationship between misogyny types and professions. Thicker streams indicate a higher number of tweets corresponding to the misogyny type originating from the respective block.

## 4.2. Hurtfulness by Profession (RQ2)

To address RQ2 – whether specific hurtful expressions target women in certain professions – we define a quantitative lexicon-based measure for assessing the hurtfulness of tweets.

**Hurtfulness Evaluation** To define a hurtfulness measure for tweets, we leverage the HurtLex lexicon, which compiles offensive words and stereotyped expressions aimed at insulting and degrading marginalized individuals and groups [26]. HurtLex organizes words into 17 fine-grained categories, each identifying a specific target or form of offense.

Inspired by the work of Nozza et al. [12], where a harmful sentence completions indicator is defined for generative language models, we employ a subset of 9 HurtLex categories for our purposes: animals, prostitution, professions, negative connotations, homosexuality, male genitalia, female genitalia, derogatory terms, and crime[6]. The hurtfulness score for a tweet w.r.t. one of the 9 categories could be computed as the ratio of HurtLex lemmas[7] from that category to the total HurtLex lemmas from any category present in the tweet. However, an approach relying solely on the HurtLex lexicon would not provide a sufficiently comprehensive analysis, as HurtLex has low coverage of the vocabulary in the AMI-PRF dataset, with only 15.42% of the lemmas in a tweet occurring in HurtLex on average.

[5] https://github.com/VinAIResearch/BERTweet/blob/master/TweetNormalizer.py

[6] For detailed descriptions of each category, we refer to Bassignana et al. [26].

[7] We retain only conservative-level lemmas.

**Table 2**
Average cosine similarity between HurtLex lemmas and ItEM centroids using Word2vec Twitter embeddings.

| HurtLex Category | Centroid similarity |
|---|---|
| animals | 0.57 |
| prostitution | 0.60 |
| professions | 0.60 |
| negative connotations | 0.55 |
| homosexuality | 0.59 |
| male genitalia | 0.52 |
| female genitalia | 0.56 |
| derogatory | 0.56 |
| crime | 0.57 |

To enhance our reference vocabulary, we leverage ItEM[8], a methodology proposed by Passaro and Lenci [10]. For each lemma in the HurtLex subset, we obtain its vectorial representation using ItEM and the Word2vec Twitter embeddings[9], following Godin [27]. For each category, we compute a centroid embedding by averaging the vectors associated with each lemma in that category. This allows us to represent each category through a unique embedding. Tab. 2 reports the average cosine similarity between lemmas of a specific category and the respective centroid. Finally, we compute the cosine similarity between each word embedding in the Word2vec Twitter vocabulary and each centroid, thus creating a new lexicon featuring a coverage of $76.51\%$ w.r.t. the AMI-PRF dataset.

We leverage the similarity scores to define a hurtful emotive score for each tweet as follows: let $\mathbf{t}$ be a lemmatized tweet, $w$ a lemma in $\mathbf{t}$, $k$ one of the 9 HurtLex categories, $\tilde{k}$ the centroid of category $k$, $s$ the cosine similarity function and $V$ the set of vocabulary items, i.e. the words for which we have a Twitter emmbedding. For each $w \in V$, we define the $ItEM$ function as:

$$ItEM(w, \tilde{k}, thr) = \begin{cases} s(w, \tilde{k}) & \text{if } s(w, \tilde{k}) \geq thr \\ 0 & \text{if } s(w, \tilde{k}) < thr \end{cases} \quad (1)$$

where $thr$ designates a threshold in $[0, 1]$ range. In other words, the $ItEM$ function outputs the cosine similarity value between $w$ and $k$'s centroid if such value is greater or equal then $thr$, while it outputs 0 if it is lower than $thr$. Additionally, if $w$ is not found in the vocabulary, its $ItEM$ value is also considered 0.

The Emotive score for a tweet $\mathbf{t}$ w.r.t. a category $k$ and a threshold $thr$ is then computed as:

$$\text{Emotive}(\mathbf{t}, k) = \frac{\sum_{w \in \mathbf{t}} ItEM(w, k, thr)}{q} \quad (2)$$

[8]https://github.com/Unipisa/ItEM/
[9]https://github.com/FredericGodin/TwitterEmbeddings

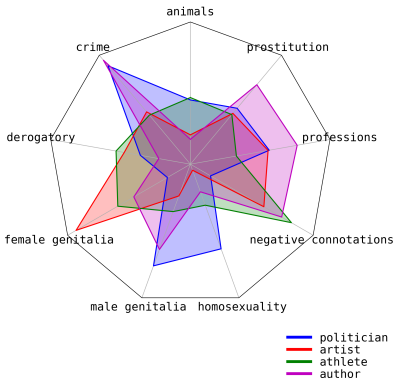


**Figure 3:** Emotive z-scores for HurtLex categories with respect to professions.

where $q$ is the number of lemmas in $\mathbf{t}$ which occur in $V$. This allows us to obtain, for each tweet-category pair, a score between $[0, 1]$, indicating the tweet hurtfulness tendency.

**Discussion** Fig. 3 provides a visual analysis of the results. The Emotive score is computed category-wise as the average of the scores for each tweet, after having standardized the values with a z-score approach. We keep a $thr$ of 0.2 in terms of cosine similarity to filter out excessively noisy category associations, while still allowing low values to contribute to the average score. This provides a general overview on the hurtful language across different professions. According to the Emotive analysis, politicians are mainly targeted with insults related to crime, homosexuality and male genitalia. This is consistent with what has been observed in Fig. 2, where forms of sexual harassment discourse were mainly directed toward political figures. For artists, we notice a peak w.r.t. female genitalia, while for athletes we register a more balanced trend, except for a peak in negative connotation. On the other hand, authors seem to be mainly targeted with crime and profession-related topics, consistent with the fact that the type of misogyny mostly inflicted towards this profession consists of derailing and stereotypes.

## 5. Conclusion

In this paper, we investigated the phenomenon of misogyny on Twitter through the lens of hurtfulness, qualifying its different manifestation considering the profession of the targets of the misogynistic attacks.

Specifically, we examined how different types of misogyny are distributed across various professions, unveiling how derailing discourse is mostly used to attack authors,

while dominance and sexual harassment speech targets especially politicians.

Additionally, we studied through a hurtfulness score measure how the language used in misogynistic tweets varies across different professions: politicians tend to be targeted with hate speech revolving around sexuality (female/male genitalia, homosexuality) and crime, while artists seem to be insulted mainly through general derogatory terms. On the other hand, less heterogeneous results were obtained for athletes and authors, except for peaks in hurtful topics regarding crimes and professions.

We acknowledge two potential limitations of our contribution: the incomplete coverage of our dataset's vocabulary by the Hurtlex-based ItEM lexicon, and our decision to focus on just four professions, which, as motivated, was guided by the representation of those professions in the AMI dataset. We therefore plan to extend the approach adopting a richer vocabulary w.r.t. datasets as well as expanding the set of professions. Indeed, as further future investigations, it could be assessed how hurtfulness dimensions change using different lexicons or automatic approaches. We also intend to investigate the distribution of misogynistic language both textual and multi-modal, as well as the broader expression of emotions in posts associated with different professions.

## Acknowledgments

Research partially 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] M. E. David, Reclaiming feminism: Challenging everyday misogyny, Policy Press, 2016.

[2] C. Tileagă, Communicating misogyny: An interdisciplinary research agenda for social psychology, Social and Personality Psychology Compass 13 (2019) e12491.

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

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

[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] 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.

[11] A. Bondielli, L. C. Passaro, Leveraging CLIP for image emotion recognition, in: E. Cabrio, D. Croce, L. C. Passaro, R. Sprugnoli (Eds.), Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2021), Online event, November 29, 2021, volume 3015 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-3015/paper172.pdf.

[12] D. Nozza, F. Bianchi, D. Hovy, HONEST: measuring

hurtful sentence completion in language models, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 2398–2406.

[13] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, A. P. Sheth, A quality type-aware annotated corpus and lexicon for harassment research, in: H. Akkermans, K. Fontaine, I. E. Vermeulen, G. Houben, M. S. Weber (Eds.), Proceedings of the 10th ACM Conference on Web Science, WebSci 2018, Amsterdam, The Netherlands, May 27-30, 2018, ACM, 2018, pp. 33–36. URL: https://doi.org/10.1145/3201064.3201103. doi:10.1145/3201064.3201103.

[14] 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.

[15] 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.

[16] D. Felmlee, P. Inara Rodis, A. Zhang, Sexist slurs: Reinforcing feminine stereotypes online, Sex Roles 83 (2020) 16–28.

[17] A.-M. Hancock, When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm, Perspectives on politics 5 (2007) 63–79.

[18] R. K. Dhamoon, Considerations on mainstreaming intersectionality, Political research quarterly 64 (2011) 230–243.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[25] S. Barreto, R. Moura, J. Carvalho, A. Paes, A. Plastino, Sentiment analysis in tweets: an assessment study from classical to modern word representation models, Data Min. Knowl. Discov. 37 (2023) 318–380. URL: https://doi.org/10.1007/s10618-022-00853-0. doi:10.1007/S10618-022-00853-0.

[26] 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.

[27] F. Godin, Improving and interpreting neural networks for word-level prediction tasks in natural language processing, Ghent University, Belgium (2019).

# A. Supplementary Material

In Figure 4, we display the tree of nested professions based on the Wikidata taxonomy concerning the popular women selected to collect the PRF dataset (§3.2). Branches identify Wikidata *subclass of* relationships, while dashed marks the connections between women and the first (or unique) occupation appearing on their Wikidata pages.We avoid reporting women's names to maintain anonymity.

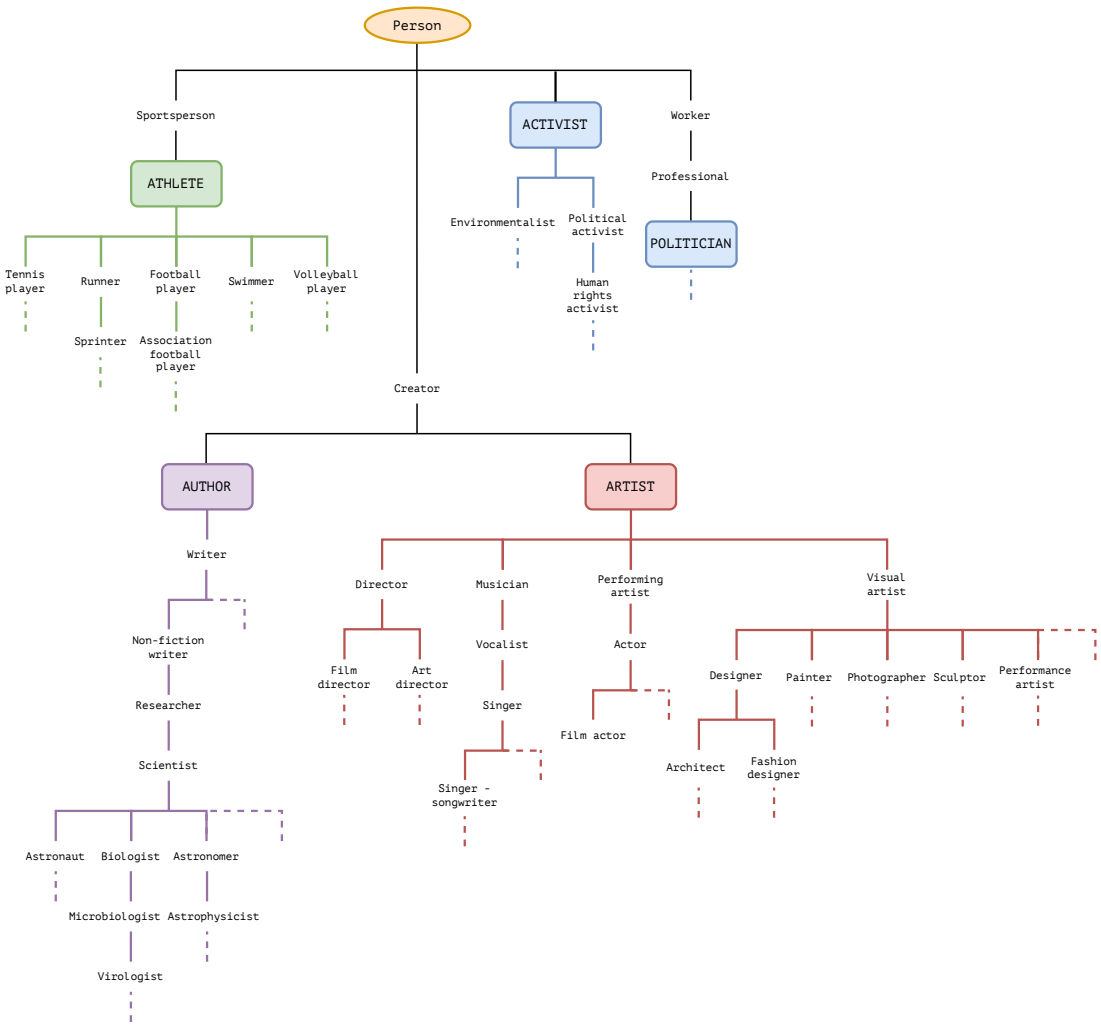


**Figure 4:** Tree of professions held by the group of popular women selected to collect the PRF dataset.