First Ask Then Act (FATA): A Community-Centered Inclusive Approach for Hate Speech Detection

Chiara Ferrando[1,*,†], Lia Draetta[1,*,†], Andrea Marra[2,*,†], Angela Zottola[3], Cristina Bosco[1] and Viviana Patti[1]

[1]Department of Computer Science, University of Turin, Italy
[2]Department of Foreign Languages, Literature and Modern Cultures, University of Turin, Italy
[3]Department of Cultures, Politics and Society, University of Turin, Italy

Abstract

In the NLP field, the significant attention paid to Hate Speech (HS) detection has highlighted how difficult it is to define HS with clear boundaries, revealing its being a context-dependent phenomenon. A recent challenge in the field of HS detection is to overcome the risks of both over-moderation and under-moderation, emphasizing the need to better understand what is perceived as hateful and what is not by communities affected by abusive language. Additionally, an interest in developing more inclusive approaches that actively involve target groups has recently emerged. This shift includes an increasing focus on underrepresented languages and communities, encouraging researchers to more actively consider ethical issues. Against this backdrop, we present a position paper with a twofold aim: firstly, we propose a review of some interdisciplinary approaches adopted so far in the field of NLP related to HS and abusive language detection; secondly, we present First Ask Then Act (FATA), a collaborative approach based on the direct involvement of individuals and target communities to collect fair and informed data. FATA proposes a multidisciplinary methodology, which integrates methods from sociolinguistics, such as surveys and focus group interviews, into the NLP data gathering workflow for HS detection.

Keywords

Hate Speech, Community-based approach, Natural Language Processing, Sociolinguistics, Multidisciplinarity

1. Introduction

In recent years, language technologies have increasingly focused on understanding and categorizing the granular nuances of language. In this context, Hate Speech (HS) and abusive language detection have received significant attention in Natural Language Processing (NLP) field [1, 2, 3, 4], emerging as a fundamental tool for moderating online content and limiting the diffusion of harmful language. The adaptability of Large Language Models (LLMs) [5] led various scholars to attempt at exploring the different nuances that HS could assume depending on diverse contexts, topical focuses and targets [6]. This has encouraged the development of increasingly precise models capable of capturing the specific shapes that HS assumes depending on the affected target, such as misogyny [7, 8, 9, 10, 11], sexism [12, 13], homophobic and transphobic discourses [14, 15]. Even though this research field is now widespread and state-of-the-art models achieve impressive results [16, 17, 18], it remains challenging to provide a univocal definition of what constitutes hate speech and to determine the extent to which certain terms should be considered harmful. In fact, HS is a context-dependent phenomenon [19, 20, 21], and it is often simplistic to classify it using clear-cut boundaries [22, 23], as the meaning of certain terms depends on the speaker's background and communicative intent[24, 25, 26]. Recent studies [27, 28, 26] highlighted that state-of-the-art HS models are at risk of both over-moderation (i.e., classifying non-hateful content as hateful) and under-moderation (i.e., failing to detect and classify hateful content), potentially leading to the removal of not abusive speech and, paradoxically, contributing to the marginalization of vulnerable groups. This can be also related to the fact that models still struggle

AEQUITAS 2025: Workshop on Fairness and Bias in AI | co-located with ECAI 2025, Bologna, Italy
*Corresponding author.
†These authors contributed equally.
chiara.ferrando@unito.it (C. Ferrando); lia.draetta@unito.it (L. Draetta); andrea.marra@unito.it (A. Marra); angela.zottola@unito.it (A. Zottola); cristina.bosco@unito.it (C. Bosco); viviana.patti@unito.it (V. Patti)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-01-07

CEUR-WS.org/Vol-4147/paper7.pdf

in distinguishing between abusive and not-abusive swearing contexts [25], disregarding the multifaceted nature of swear words, which are often used in casual contexts, also with positive social functions [29].

In line with this, a still open challenge is the identification of reclamatory uses of slurs, instances in which speakers, usually members of target groups, repurpose terms historically used to derogate their group, to express belonging and identity, manifesting solidarity and subverting structures of discrimination [30, 31, 32, 33]. Primarily explored in philosophy of language, this phenomenon offers valuable insights for developing socially relevant HS detection models, able to recognize context-specific usages. In the NLP field, the reclamation of slurs is mostly overlooked [1, 26, 23]; therefore, a current challenge is to collect new fair data suitable for models fine-tuning, to avoid the risk of mistakenly censoring not abusive speech in AI-supported content moderation, and to find a balance between detecting HS and preserving the free dissemination of ideas and opinions. In promoting inclusive and fair language several ethical questions arise, particularly concerning how certain linguistic uses are perceived by the target communities. Understanding these perceptions is crucial for achieving more accurate representations of language and its associated cultural diversity in NLP models. Moreover, it is essential to prevent linguistic discrimination, ensure social acceptance by the communities involved, and assess the broader impact of language technologies on individuals and groups.

With this in mind, in this position paper we propose to go beyond the conventional methodology used in the field of NLP to collect data, by promoting a new collaborative approach based on fairness and representativeness. Reviewing the approaches that so far involved communities in HS and abusive language detection (Section 2) and embracing the perspectives of Queer Linguistics and Intersectionality as theoretical and analytical frameworks (described in the Section 3), we present the First Ask Then Act (FATA) proposal (Section 3). The main contribution of FATA is to be a collaborative data-gathering approach, in which communities are directly engaged both in the definition and modeling processes through methodological procedures from social sciences.

2. Related Works

In recent years, several studies have addressed some of the current limits of NLP in understanding the nuances of language. Some studies have focused on how NLP can take advantage of sociolinguistics [34, 35], relying on the hypothesis that different facets of language variation and the related sociocultural meanings must be taken into account when building NLP models [36, 37, 38]. In the field of HS detection, state-of-the-art approaches have increasingly demonstrated that incorporating a sociolinguistic perspective, such as adapting language models to target dialects [39], or accounting for social context and variables [40, 41], can significantly enhance overall system performance. Despite advances in NLP, social meaning remains often largely overlooked [42], and cultural differences are poorly represented. On the same line, Nguyen [35] highlights several limitations of current NLP models, including the under-representation of language variation in training and fine-tuning datasets [38], as well as the tendency to impose dominant language ideologies while dismissing others as noise. Furthermore, Hofmann and colleagues [43] demonstrate that LLMs often produce biased and stereotypical representations associated with specific language varieties.

Recent searches have emphasized the problem of human label variation, which affects all stages of the Machine Learning pipeline from data modeling to evaluation [44, 45], and have questioned the reliability of common labeling techniques to develop NLP corpora, advocating for a rethinking of the traditional annotation practices that assume a single ground truth.

A theoretical framework known as perspectivism[1] [46] aims to enhance Machine Learning (ML), leveraging data annotated by individuals belonging to different groups and identities. Recent studies have adopted a perspectivist approach in the context of HS detection [47, 48, 49], revealing that annotators from different demographic groups often disagree on what constitutes HS, and that cultural factors significantly influence perception and annotation agreement. Madeddu and colleagues [50] introduce a disaggregated hate speech dataset that includes sociodemographic information about

[1] https://pdai.info/

annotators, offering valuable insights into how different population groups perceive social phenomena and how this knowledge can be leveraged to enhance models performance. In this direction, Kurrek and colleagues [51] propose the Inclusivity by Design approach, which promotes opinion diversity by developing a novel data collection and annotation method that pairs annotators with differing demographic backgrounds.

In several areas of ML, the Participatory Design (PD) approach aimed to amplify the voices of people represented by technology and its development [52]. Birhane and colleagues [53] emphasize that participation is an important tool for the responsible development of AI, since it improves the human-like performances of algorithms. In the context of AI for social good, participatory activities are evoked as a means to improve AI systems that affect communities where, ideally, impacted groups take part as stakeholders through participatory design and implementation. According to researchers [53, 54], a key objective of participatory methods is to disseminate knowledge about technical systems and their impacts by involving experts and non-experts stakeholder. In the domain of HS detection, several studies support the need of collaborative approaches. For instance, Parker and Ruths [22] aware of the risks of marginalization, actively involves people in the evaluation of algorithmic results to ensure that human perspectives define what constitutes reliable and fair results.

However, PD is not an end-all solution [52]. Sloane and colleagues [55] provide a discussion of how PD can result in "participation-washing" and how such design must be context-specific, long-term, and genuine. The authors present different modalities of participation, able to place marginalized groups at the center of collaborative and creative design processes. In a recent work, PD practices are presented as a mutual learning process among participants and researchers [56].

Another valuable perspective is intersectionality, that emphasizes the complexity of social categorizations, demonstrating how different axes of identity - such as gender, ethnicity, sexuality, class, and ability - interact and build various humans identities [57, 58, 59]. In the context of HS detection, some studies have adopted an intersectional lens that takes on different nuances according to specific traits that members of a target group have [60]. Researches that take intersectionality into account mostly focus on bias [61, 62] and stereotypes [63], investigating how different degrees of discrimination are possible due to the intersection of various axes [64], generating a more subtle and complex hatred [65, 66]. In this context, a recent study [60], focused on Gender-Based Violence, proposes an innovative annotation schema for a fine-grained detection of misogynistic content related to intersectional traits of people involved.

3. The First Ask Then Act (FATA) Approach

As a theoretical framework, the Queer Linguistics stance has been fundamental in embracing a change of perspective in which sociolinguistic techniques inform NLP studies. It consists of an interdisciplinary field of study that could be defined as a theoretical background, emerging from feminist and LGBTQ+ studies, that critiques essentialist views of identity and highlights how language both constructs and deconstructs sexual and gender identities [67]. By disrupting binary thinking and essentialist assumptions, this field not only provides insights into the linguistic practices of a variety of individuals and identities, but also critiques the sociopolitical power structures embedded in language.

Along the same line, intersectionality engages directly with multifaceted reality [68], as it confronts intricate and overlapping structures of oppression without resorting to oversimplification. This perspective has proven particularly useful in feminist and queer discourse, as it enables a nuanced understanding of how power operates within different contexts and across multiple social dimensions [69]. In linguistic studies, intersectionality provides a valuable lens for examining how discourse constructs and reinforces social hierarchies. The intersectional approach aligns with the broader movement towards discursive analysis in linguistics, which critiques fixed and universalist notions of meaning, instead highlighting context-dependent nature of identity construction [70].

This is relevant in the methodology proposed in this contribution: by integrating such a perspective, it is possible to better analyze the ways in which language, language uses and language users reflect

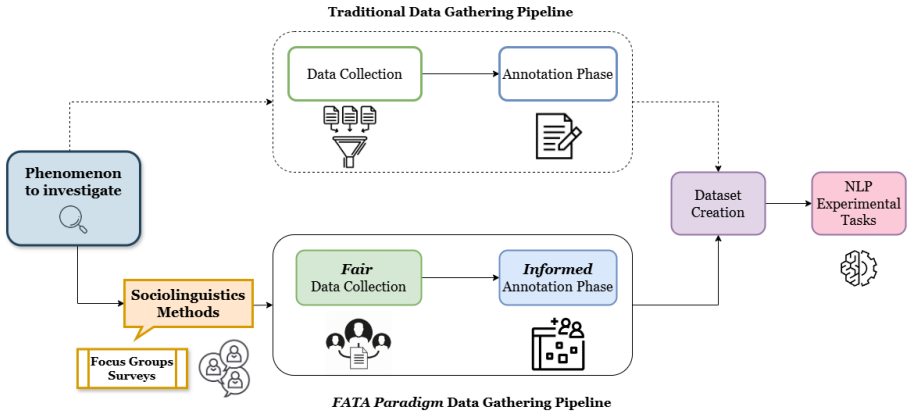

Figure 1: The traditional data gathering pipeline and the First Ask Then Act (FATA) pipeline for HS detection.

and perpetuate intersecting systems of oppression and privilege, thus offering critical insights into the discursive mechanisms that sustain social inequalities [71]. With regard to the risk of censoring in automatic HS detection systems, a theoretical framework able to include different perspectives strengthens the collaboration with communities at the margins, that are traditionally under-represented in NLP studies [52], as well as broader groups, also allowing researchers to go a step further and explore phenomena in depth.

3.1. The FATA Data Gathering Pipeline

In this section, we show how the First Ask Then Act (FATA) approach can be integrated into the traditional data gathering process by proposing the introduction of well established methods in sociolinguistics, such as focus groups and surveys. We advocate the importance of including additional steps in the conventional workflow of data collection, in order to collect fair and community-informed data. In this sense, we propose to concretely involve the communities to gather opinions and points of view from the groups affected by a societal phenomenon as the first stage of the FATA proposal.

Considering the pervasiveness of linguistics phenomena such as HS and online abusive language, involving individuals and target communities who experience this type of hatred on a daily basis, could lead to an in-depth understanding of all the nuances that characterized these forms of hate. In Figure 1, the traditional data gathering pipeline and the FATA approach pipeline are reported. As shown, both pipelines move from a phenomenon under investigation, and end with the creation of a dataset. The main novelty of this proposal is the introduction of sociolinguistics methods in the data gathering pipeline, which are fundamental to collect opinions on which the data collection and annotation phases in the FATA proposal are based. This innovative and multidisciplinary approach leads to the design of an accurate representation of the studied phenomenon, reducing the risk of bias and unfair propagation.

In the following subsections, all the steps that compose the FATA approach are described in details.

3.2. Sociolinguistic Methods

As delineated above, the FATA proposal claims the importance of introducing sociolinguistic methodologies reliable in involving individuals and target groups related to the phenomenon on study, such as focus group, interviews for fine-grained investigations and surveys for large-scale analysis.

Focus Group Interviews A focus group is a semi-structured group interview, which functions as a data collection procedure that brings together people to answer input questions under the guidance of a research team, usually consisting of a moderator and some observers [72, 73]. This methodology provides direct access to language and concepts, facilitates the collective construction of meaning and gives researchers a better understanding of how a community structures and organizes its social world.

It allows the direct involvement of people and communities leading to record linguistic strategies and to gather opinions on the facets of the topic under investigation. In this initial stage, small groups from the target community are involved in a safe space in which personal experiences and opinions are shared. From the group interactions, a series of insights, thoughts or doubts may emerge that can be explored on a larger scale through a survey.

Surveys A sociolinguistic survey is a structured investigation that involves a representative sample of people. The distribution of an online questionnaire enables to reach a substantial number of people and gather diverse perspectives to trigger a reflection on the sociolinguistic variability of the phenomenon [74, 75]. In our proposal, it is a useful technique to activate the metalinguistic competence of the person to whom the questionnaire is proposed, and obtain both sociodemographic and linguistic-perceptual insights, without neglecting self-identification information. In addition, a survey can provide confirmation of certain linguistic usages, collect data on contexts and reasons for use, as well as explicit opinions on the their acceptability. Finally, the implementation of a survey could serve as a valuable tool for informing subsequent stages of the study, particularly in the context of data collection. By identifying variations in opinions across different demographic groups, it facilitates the selection of the most suitable group of annotators, thereby enhancing the reliability and relevance of the collected data.

3.3. Community-based Data Gathering

As shown in Figure 1, sociolinguistic techniques offer useful insights into the strategies that compose the FATA proposal, promoting fairness and representativeness in both data collection and annotation for hate speech detection.

Fair Data Collection Data collection is a critical phase in the NLP pipeline, often raising concerns about how data are gathered and its representativeness [76, 38]. It is now well recognized in the NLP community that imbalanced or non-representative data can reinforce existing biases and contribute to the under-representation of minority groups [77, 78, 79, 43]. An interdisciplinary approach can lead to the overcoming of classical issues linked to data collection in the development of corpora and benchmark datasets for HS detection. With the insights gathered during the focus group and survey phases, NLP researchers can formulate more precise queries for data collection, identify the most representative contexts for sourcing data, and determine the linguistic forms relevant to the target phenomena and identities. Additionally, collaborating with communities in earlier stages can help researchers gain legitimacy for data collection and facilitate direct access to authentic data.

Informed Annotation Phase The sociolinguistic survey, composed of input sentences and precise questions, is a fundamental support for the creation and definition of an annotation scheme. Surveys can serve as large-scale pilot annotations beyond the research team, representing a step forward in collaborative approaches [46]. Additionally, designing large-scale surveys targeted at specific communities helps move beyond the traditional reliance on annotation teams primarily composed of individuals from Western Societies.

Dataset Creation and NLP Experimental Tasks As anticipated above, the traditional data gathering approach and the FATA proposal merge in the final stages of the conventional NLP workflow, i.e. the dataset creation that precedes the experimental phase. Although both approaches pursue the same objectives, it is important to emphasize the different quality of the data collected. In fact, in the FATA proposal, data are supposed to be more informed, fair and controlled as they are gathered by actively involving marginalized and vulnerable communities which are target of HS. Representative data lead to the construction of more accurate datasets, which improve the performance of experimental tasks.

4. Conclusion and Future Works

This contribution introduces an innovative approach that integrates theoretical insights from Queer Linguistics and Intersectionality with methodological tool from the social sciences. In this position paper, we present the First Ask Then Act (FATA) proposal, which takes into account different perspectives and experiences relevant to the phenomenon under study. By incorporating established sociolinguistic techniques, such as focus groups and surveys, into the NLP data collection workflow, the FATA approach aims to enhance the quality, fairness, and representativeness of the resulting data. Nowadays, not taking into consideration the sociocultural context in which data are produced can lead researchers to observe phenomena from a detached point of view, without deeply considering their ethical impact on society. By actively engaging target communities in an effort to empower under-represented people [80], researchers will be able to obtain more accurate and reliable findings, without risking the collection of misleading data. For this reason, FATA proposal would lead to more authentic and bias aware studies, overcoming ethical issues through the direct involvement of marginalized people and vulnerable communities, which are often target of HS.

As future work, we plan to present a comprehensive case study on the full adoption of the proposed approach investigating slur reclamation across different languages. By involving communities and researchers from various research fields, we plan to show how the FATA approach can function as a valuable tool for the HS detection community. In addition, we aim to promote the adoption of FATA in different research fields by asking researchers who want to tackle studies that impact people's daily lives and investigate social phenomena to primarily ask and gather opinions by actively involving people from different groups. This procedure leads to fully understand the topic and ensure more representative and fair studies.

5. Limitations

In this section, we discuss the main limitations of this research, recognizing that the proposed methodology is experimental and may require further development.

Firstly, with regards to the sociolinguistic methods, focus group interviews lead to the collection of qualitative data from a group of people which could have been involuntarily cherry-picked by the researchers; whereas the large-scale survey allows the collection of a large amount of quantitative opinions, although the use of closed-ended questions may result in oversimplifications and introduce researcher biases. Furthermore, in order to involve people with different points of view, educational backgrounds and experiences, it becomes necessary to "get out" from the researchers' personal bubble. Moreover, in some cases, the main limit is the impossibility of collecting data from a completely different perspective, as people who are not interested in the topic under investigation could not fill out a survey or do not want to be interviewed. This means that the answers collected could be quite aligned with researchers' positionality. Another issue related to the involvement of people is the possibility to be influenced by social desirability, a tendency through which subjects give socially desirable responses by over-reporting behaviors that make them appear good and under-reporting those that make them look bad [81]. This is problematic because it may lead people to hide their authentic ideas and positions in order to be accepted by the digital community [82] or by researchers.

Considering time and cost limits, it is not always possible to conduct both qualitative and quantitative studies. As far as cost limitations are concerned, there are some paid platforms that can be used to ask people to fill in a survey or to express their opinions through annotation tasks. Unfortunately, at the moment, these platforms can not always reach people with diverse backgrounds and sociodemographic characteristics [83]. Indeed, some minority languages do not provide sufficient data and people belonging to specific target groups may be underrepresented. Finally, recognizing the internal diversity of communities and the limits of generalization, we advocate for fine-grained approaches that engage different individuals within the same community to more accurately capture the complexity of the phenomenon under study and its facets.

Acknowledgments

The work of Viviana Patti and Cristina Bosco is partially supported by "HARMONIA" project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 under the NextGenerationEU programme.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J. Pierrehumbert, HateCheck: Functional tests for hate speech detection models, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 41–58. URL: https://aclanthology.org/2021.acl-long.4/. doi:10.18653/v1/2021.acl-long.4.

[2] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate speech detection in Italian social media text, in: K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, Z. Talat (Eds.), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 252–260. URL: https://aclanthology.org/2022.woah-1.24/. doi:10.18653/v1/2022.woah-1.24.

[3] F. M. Plaza-del arco, D. Nozza, D. Hovy, Respectful or toxic? using zero-shot learning with language models to detect hate speech, in: Y.-l. Chung, P. R{\"ottger}, D. Nozza, Z. Talat, A. Mostafazadeh Davani (Eds.), The 7th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 60–68. URL: https://aclanthology.org/2023.woah-1.6/. doi:10.18653/v1/2023.woah-1.6.

[4] J. S. Malik, H. Qiao, G. Pang, A. van den Hengel, Deep learning for hate speech detection: a comparative study, International Journal of Data Science and Analytics (2024) 1–16.

[5] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1049–1065. URL: https://aclanthology.org/2023.findings-acl.67/. doi:10.18653/v1/2023.findings-acl.67.

[6] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, V. Patti, Emotionally informed hate speech detection: A multi-target perspective, Cogn. Comput. 14 (2022) 322–352. URL: https://doi.org/10.1007/s12559-021-09862-5. doi:10.1007/S12559-021-09862-5.

[7] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Inf. Process. Manag. 57 (2020) 102360. URL: https://doi.org/10.1016/j.ipm.2020.102360. doi:10.1016/J.IPM.2020.102360.

[8] A. Muti, F. Ruggeri, K. A. Khatib, A. Barrón-Cedeño, T. Caselli, Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 21091–21107. URL: https://aclanthology.org/2024.emnlp-main.1174/. doi:10.18653/v1/2024.emnlp-main.1174.

[9] M. Z. U. Rehman, S. Zahoor, A. Manzoor, M. Maqbool, N. Kumar, A context-aware attention and graph neural network-based multimodal framework for misogyny detection, Information Processing & Management 62 (2025) 103895.

[10] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, M. Ullah, Enhancing misogyny detection in bilingual texts

using explainable ai and multilingual fine-tuned transformers, Complex & Intelligent Systems 11 (2025) 39.

[11] A. Mohasseb, E. Amer, F. Chiroma, A. Tranchese, Leveraging advanced nlp techniques and data augmentation to enhance online misogyny detection, Applied Sciences 15 (2025) 856.

[12] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023–learning with disagreement for sexism identification and characterization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 316–342.

[13] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305/. doi:10.18653/v1/2023.semeval-1.305.

[14] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the first shared task on homotransphobia detection in italian, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper26.pdf.

[15] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, T. Alcántara, M. Soto, C. Macias, Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population, Procesamiento del Lenguaje Natural 73 (2024) 393–405.

[16] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1049–1065. URL: https://aclanthology.org/2023.findings-acl.67/. doi:10.18653/v1/2023.findings-acl.67.

[17] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, L. Wang, Mitigating hallucination in large multi-modal models via robust instruction tuning, arXiv preprint arXiv:2306.14565 (2023).

[18] W. Sun, H. Xu, X. Yu, P. Chen, S. He, J. Zhao, K. Liu, ItD: Large language models can teach themselves induction through deduction, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2719–2731. URL: https://aclanthology.org/2024.acl-long.150/. doi:10.18653/v1/2024.acl-long.150.

[19] A. Brown, What is hate speech? part 2: Family resemblances, Law and Philosophy 36 (2017) 561–613.

[20] L. Anderson, M. R. Barnes, Hate speech, in: E. N. Zalta (Ed.), Stanford Encyclopedia of Philosophy, The Metaphysics Research Lab, Philosophy Department, Stanford University, 2022. URL: https://plato.stanford.edu/.

[21] M. Yoder, L. Ng, D. W. Brown, K. Carley, How hate speech varies by target identity: A computational analysis, in: A. Fokkens, V. Srikumar (Eds.), Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 27–39. URL: https://aclanthology.org/2022.conll-1.3/. doi:10.18653/v1/2022.conll-1.3.

[22] S. Parker, D. Ruths, Is hate speech detection the solution the world wants?, Proceedings of the National Academy of Sciences 120 (2023) e2209384120.

[23] L. Draetta, C. Ferrando, M. Cuccarini, L. James, V. Patti, ReCLAIM project: Exploring Italian slurs reappropriation with large language models, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 335–342. URL: https://aclanthology.org/2024.clicit-1.40/.

[24] E. W. Pamungkas, V. Basile, V. Patti, Do you really want to hurt me? predicting abusive swearing

in social media, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6237–6246. URL: https://aclanthology.org/2020.lrec-1.765.

[25] E. W. Pamungkas, V. Basile, V. Patti, Investigating the role of swear words in abusive language detection tasks, Lang. Resour. Evaluation 57 (2023) 155–188. URL: https://doi.org/10.1007/s10579-022-09582-8. doi:10.1007/S10579-022-09582-8.

[26] E. Zsisku, A. Zubiaga, H. Dubossarsky, Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination, in: Proceedings of the 16th ACM Web Science Conference, 2024, pp. 241–249.

[27] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678. URL: https://aclanthology.org/P19-1163/. doi:10.18653/v1/P19-1163.

[28] T. Dias Oliva, D. M. Antonialli, A. Gomes, Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online, Sexuality & Culture 25 (2021) 700–732.

[29] T. Jay, Do offensive words harm people?, Psychology, public policy, and law 15 (2009) 81.

[30] C. Bianchi, Slurs and appropriation: An echoic account, Journal of Pragmatics 66 (2014) 35–44.

[31] E. Nossem, Queer, frocia, femminiellə, ricchione et al.–localizing "queer" in the italian context, gender/sexuality/italy 6 (2019).

[32] B. Cepollaro, D. L. de Sa, The successes of reclamation, Synthese 202 (2023) 205.

[33] J. Mun, C. Buerger, J. T. Liang, J. Garland, M. Sap, Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–22.

[34] D. Hovy, The social and the neural network: How to make natural language processing about people again, in: M. Nissim, V. Patti, B. Plank, C. Wagner (Eds.), Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 42–49. URL: https://aclanthology.org/W18-1106/. doi:10.18653/v1/W18-1106.

[35] D. Nguyen, Collaborative growth: When large language models meet sociolinguistics, Language and Linguistics Compass 19 (2025) e70010.

[36] D. Nguyen, L. Rosseel, When social meaning meets nlp: How can nlp models inform sociolinguistic research and vice versa?, in: Sociolinguistics Symposium 24, 2022.

[37] D. Yang, D. Hovy, D. Jurgens, B. Plank, Socially aware language technologies: Perspectives and practices, Computational Linguistics 51 (2025) 689–703. URL: https://aclanthology.org/2025.cl-2.10/. doi:10.1162/coli_a_00556.

[38] J. Grieve, S. Bartl, M. Fuoli, J. Grafmiller, W. Huang, A. Jawerbaum, A. Murakami, M. Perlman, D. Roemling, B. Winter, The sociolinguistic foundations of language modeling, Frontiers in Artificial Intelligence 7 (2025) 1472411.

[39] J. M. Pérez, P. Miguel, V. Cotik, Exploring large language models for hate speech detection in Rioplatense Spanish, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 7174–7187. URL: https://aclanthology.org/2025.findings-naacl.400/. doi:10.18653/v1/2025.findings-naacl.400.

[40] S. Nagar, F. A. Barbhuiya, K. Dey, Towards more robust hate speech detection: using social context and user data, Social Network Analysis and Mining 13 (2023) 47.

[41] T. Chaturvedi, S. SR, P. Duraisamy, Exploring the relationship between social context of speech in race, gender and autonomic detection of hate speech on social media, International Journal of Health and Allied Sciences 13 (2024) 2.

[42] D. Nguyen, L. Rosseel, J. Grieve, On learning and representing social meaning in nlp: a sociolinguistic perspective, in: Proceedings of the 2021 Conference of the North American Chapter of the

Association for Computational Linguistics: Human language technologies, 2021, pp. 603–612.
[43] V. Hofmann, P. R. Kalluri, D. Jurafsky, S. King, AI generates covertly racist decisions about people based on their dialect, Nature 633 (2024) 147–154.
[44] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10671–10682. URL: https://aclanthology.org/2022.emnlp-main.731/. doi:10.18653/v1/2022.emnlp-main.731.
[45] M. Orlikowski, P. Röttger, P. Cimiano, D. Hovy, The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1017–1029. URL: https://aclanthology.org/2023.acl-short.88/. doi:10.18653/v1/2023.acl-short.88.
[46] S. Frenda, G. Abercrombie, V. Basile, A. Pedrani, R. Panizzon, A. T. Cignarella, C. Marco, D. Bernardi, Perspectivist approaches to natural language processing: a survey, Language Resources and Evaluation (2024) 1–28.
[47] S. Casola, S. Lo, V. Basile, S. Frenda, A. Cignarella, V. Patti, C. Bosco, et al., Confidence-based ensembling of perspective-aware models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Houda Bouamor, Juan Pino, Kalika Bali, 2023, pp. 3496–3507.
[48] P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. Von Vacano, C. Kennedy, The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022, 2022, pp. 83–94.
[49] G. Rizzi, The many facets of hateful content detection: from perspectivism to bias, 2025. URL: https://boa.unimib.it/retrieve/105a571f-7ffa-44f3-8e60-b19ba15344ee/phd_unimib_794865.pdf, doctoral thesis, PhD program Computer Science, Università degli Studi Milano Bicocca.
[50] M. Madeddu, S. Frenda, M. Lai, V. Patti, V. Basile, Disaggreghate it corpus: A disaggregated italian dataset of hate speech, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023, volume 3596 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3596/paper29.pdf.
[51] J. Kurrek, H. M. Saleem, D. Ruths, Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020, pp. 138–149.
[52] A. Field, S. L. Blodgett, Z. Waseem, Y. Tsvetkov, A survey of race, racism, and anti-racism in NLP, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1905–1925. URL: https://aclanthology.org/2021.acl-long.149/. doi:10.18653/v1/2021.acl-long.149.
[53] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, S. Mohamed, Power to the people? opportunities and challenges for participatory ai, in: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2022, pp. 1–8.
[54] F. Delgado, S. Yang, M. Madaio, Q. Yang, The participatory turn in ai design: Theoretical foundations and the current state of practice, in: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2023, pp. 1–23.
[55] M. Sloane, E. Moss, O. Awomolo, L. Forlano, Participation is not a design fix for machine learning (pp. 1–7), in: Proceedings of the International Conference on Machine Learning, Vienna, Austria, 2020.
[56] T. Caselli, R. Cibin, C. Conforti, E. Encinas, M. Teli, Guiding principles for participatory design-inspired natural language processing, in: Proceedings of the 1st Workshop on NLP for Positive

Impact, Association for Computational Linguistics (ACL), 2021, pp. 27–35.
[57] K. Crenshaw, Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics, The University of Chicago Legal Forum 140 (1989) 139–167.
[58] K. Crenshaw, Mapping the margins: Intersectionality, identity politics, and violence against women of color, Stanford Law Review 43 (1991) 1241–1299. URL: http://www.jstor.org/stable/1229039.
[59] P. H. Collins, Black feminist thought: Knowledge, consciousness, and the politics of empowerment, Routledge, 2000.
[60] C. Ferrando, M. Madeddu, V. Patti, M. Lai, S. Pasini, G. Telari, B. Antola, Exploring YouTube comments reacting to femicide news in Italian, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 356–365. URL: https://aclanthology.org/2024.clicit-1.43/.
[61] J. P. Lalor, Y. Yang, K. Smith, N. Forsgren, A. Abbasi, Benchmarking intersectional biases in nlp, in: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2022, pp. 3598–3609.
[62] M. A. Stranisci, R. Damiano, E. Mensa, V. Patti, D. Radicioni, T. Caselli, WikiBio: a semantic resource for the intersectional analysis of biographical events, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12370–12384. URL: https://aclanthology.org/2023.acl-long.691/. doi:10.18653/v1/2023.acl-long.691.
[63] A. Leidinger, R. Rogers, How are llms mitigating stereotyping harms? learning from search engine studies, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, volume 7, 2024, pp. 839–854.
[64] H.-W.-S. Bao, P. Gries, Intersectional race–gender stereotypes in natural language, British Journal of Social Psychology (2024).
[65] I. Spada, M. Lai, V. Patti, Inters8: A corpus to study misogyny and intersectionality on twitter., in: Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR, 2023.
[66] C. Casula, S. Salto, A. Ramponi, S. Tonelli, Delving into qualitative implications of synthetic data for hate speech detection, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 19709–19726.
[67] H. Motschenbacher, Language, Gender and Sexual Identity: Poststructuralist Perspectives, John Benjamins, 2010.
[68] K. Davis, Intersectionality as buzzword: A sociology of science perspective on what makes a feminist theory successful, Feminist theory 9 (2008) 67–85.
[69] N. Lykke, Feminist studies: A guide to intersectional theory, methodology and writing, Routledge, 2010.
[70] M. Bucholtz, K. Hall, Language and identity, A companion to linguistic anthropology 1 (2004) 369–394.
[71] E. Esposito, C. Pérez-Arredondo, A. Zottola, Intersecting inequalities: towards a critical discursive approach, Journal of Gender Studies (2024) 1–10.
[72] M. Bloor, M. Thomas, J. Frankland, Focus groups in social research, SAGE Publications Ltd, 2000.
[73] R. A. Krueger, M. A. Casey, J. Donner, S. Kirsch, J. N. Maack, Social analysis: selected tools and techniques, World Dev 36 (2001) 4–23.
[74] L. Milroy, M. Gordon, Sociolinguistics: Method and interpretation, John Wiley & Sons, 2008.
[75] A. Marra, C. Ferrando, L. Draetta, B. Cepollaro, V. Patti, How is the reclamation of slurs perceived in Italian? A sociolinguistic survey to inform future NLP studies, Linguistik Online. Special issue on Gender-inclusive language in a multilingual Europe. Institutional policies, their applications and AI-related developments (2025). In press.
[76] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, J. Xu, Bias and unfairness in information retrieval systems:

New challenges in the llm era, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6437–6447.

[77] R. Navigli, S. Conia, B. Ross, Biases in large language models: origins, inventory, and discussion, ACM Journal of Data and Information Quality 15 (2023) 1–21.

[78] H. Kotek, D. Q. Sun, Z. Xiu, M. Bowler, C. Klein, Protected group bias and stereotypes in large language models, arXiv preprint arXiv:2403.14727 (2024).

[79] H. Luo, H. Huang, Z. Deng, X. Liu, R. Chen, Z. Liu, Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm, arXiv preprint arXiv:2407.15240 (2024).

[80] L. Havens, M. Terras, B. Bach, B. Alex, Situated data, situated systems: A methodology to engage with power relations in natural language processing research, in: M. R. Costa-jussà, C. Hardmeier, W. Radford, K. Webster (Eds.), Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 107–124. URL: https://aclanthology.org/2020.gebnlp-1.10/.

[81] D.-H. A. Kwak, X. Ma, S. Kim, When does social desirability become a problem? detection and reduction of social desirability bias in information systems research, Information & Management 58 (2021) 103500.

[82] F. Massara, F. Ancarani, M. Costabile, F. Ricotta, Social desirability in virtual communities, International Journal of Business Administration 3 (2012) 93–100.

[83] Y. Y. Chiu, L. Jiang, B. Y. Lin, C. Y. Park, S. S. Li, S. Ravi, M. Bhatia, M. Antoniak, Y. Tsvetkov, V. Shwartz, Y. Choi, CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 25663–25701. URL: https://aclanthology.org/2025.acl-long.1247/. doi:10.18653/v1/2025.acl-long.1247.