Introduction to the Special Issue of the TAL Journal on Abusive Language: Linguistic Resources, Methods and Applications

Delphine Battistelli* — **Farah Benamara**** — **Viviana Patti*****

* *MoDyCo-CNRS, Université Paris Nanterre*
** *Université de Toulouse, IRIT-CNRS, Toulouse INP and IPAL-CNRS Singapore*
*** *Dipartimento di Informatica, Università degli Studi di Torino, Italy*

ABSTRACT. *Abusive language and the propagation of harmful stereotypes have unfortunately become commonplace occurrences on various social media platforms, partly due to users' freedom, anonymity and the lack of regulation provided by these platforms. The sheer volume and often implicit nature of such unwanted content make manual moderation of these user spaces a formidable task. Various scientific communities (Computational Social Science, Natural Language Processing and Computational Linguistics) interested in its at least partial automation have taken up the problem over the past ten years. This special issue aims to encourage interdisciplinary submissions in the field of abusive language detection discussing the limitations of the current approaches and directions for future work.*

KEYWORDS: *Abusive language, Linguistic resources, Automatic detection*

TITRE. *Introduction au numéro spécial de la revue TAL sur le discours de haine : ressources linguistiques, méthodes et applications*

RÉSUMÉ. *Les discours de haine ainsi que la propagation de stéréotypes qui les accompagnent bien souvent sont légion sur les médias sociaux en raison de l'anonymat de leurs utilisateurs, mais aussi du fait du manque de réglementation fournie par les plateformes. Le volume considérable et la nature souvent implicite de ces contenus indésirables font de la modération manuelle une tâche extrêmement complexe. Les sciences sociales computationnelles, le traitement automatique des langues et la linguistique computationnelle se sont emparées de la problématique depuis une dizaine d'années. Ce numéro spécial a pour objectif d'encourager les soumissions interdisciplinaires autour de la tâche de détection de discours de haine tout en abordant les limites des approches actuelles ainsi que les orientations futures.*

MOTS-CLÉS : *Discours de haine, Ressources linguistiques, Détection automatique*

1. Introduction

1.1. *Abusive Language Detection: a Well Established Interdisciplinary Research Field*

Abusive language, hate speech, and the propagation of harmful stereotypes have unfortunately become commonplace occurrences on various social media platforms, due to users' freedom and anonymity and the absence of regulation provided by these platforms. The sheer volume and often implicit nature of such unwanted content make manual moderation of these user spaces a formidable task. Consequently, the Computational Social Science, Natural Language Processing (NLP) and Computational Linguistics communities have proposed numerous works to create resources, datasets, and models aimed at automating the task of abusive language detection (henceforth ALD) (Talat and Hovy, 2016; Fortuna and Nunes, 2018; Vidgen *et al.*, 2019; Fortuna *et al.*, 2020), making it a significant and well-established research area in NLP, with a substantial body of literature.

At the international level, many dedicated workshops have been organized, such as the workshop on Online Abuse and Harms (WOAH) @ACL 2022, ACL 2023, NAACL 2024 (47, 55, 56 submissions respectively) and the workshop on Trolling, Aggression and Cyberbullying @LREC 2020 (70 submissions). We also cite well-attended shared tasks such as HateEval (Basile *et al.*, 2019), OffensEval (Zampieri *et al.*, 2019; Zampieri *et al.*, 2020) and ToxicSpan@ SemEval 2019, 2020 and 2021. For example, 74 (resp. 70) teams submitted papers at HateEval (resp. OffensEval), HateEval being co-organized by one of the coordinator of this special issue. Finally, two special issues of the Journal of Online Social Networks and Media, volume 27, 2022 (Detecting, Understanding and Countering Online Harms) and the Journal of Personal and Ubiquitous Computing, volume 27 (2023) (Intelligent Systems for Tackling Online Harms).

At the national level (i.e., France), most special issues/workshops are multidisciplinary, with a particular focus on approaches from social science and linguistics. For example, the Draine multidisciplinary workshops organized by a French consortium on combating extreme and hate discourse. [1] We also cite "Analyse et exploration des données sociales" (analysis and exploration of social data) (ALIAS) workshop series @TALN 2018 and INFORSID 2019 proposed by the ALIAS GDR-MADICS, a CNRS action on cyberviolence and extreme ideology in social media, [2] founded by the two French coordinators of this special issue. We finally cite a special issue of the Journal MOTS in 2021. [3]

1. https://groupedraine.github.io/.
2. https://www.madics.fr/event/1520426929-3916/.
3. https://shs.cairn.info/journal-mots?lang=en.

1.2. *Abusive Language: a Complex Phenomenon*

Following Poletto *et al.,* (Poletto *et al.*, 2021a), we use here "Abusive Language" (AL) as an umbrella term to refer to the various forms of harmful language, such as toxic, offensive language, hate speech, and stereotypes. The reader can refer to Vidgen *et al.,* and Madukwe *et al.,* for a discussion on the lack of universal definitions and its impact on automatic detection (Vidgen *et al.*, 2019; Madukwe *et al.*, 2020). For comprehensive overviews of this field, we recommend surveys such as Schmidt *et al.,* Fortuna and Nunes, Vidgen and Derczynski, Poletto *et al.,* Yin and Zubiaga, and Pamungkas *et al.,* (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Vidgen and Derczynski, 2020; Poletto *et al.*, 2021a; Yin and Zubiaga, 2021; Pamungkas *et al.*, 2023).

AL is topically focused (misogyny, sexism, racism, xenophobia, homophobia, etc.), and each specific manifestation of hate speech targets different vulnerable groups based on characteristics such as gender (misogyny, sexism), ethnicity, race, religion (xenophobia, racism, islamophobia), sexual orientation (homophobia), and so on. Most automatic abusive language detection approaches cast the problem into a binary classification task by neglecting three crutial aspects: (1) the topical focus or the target-oriented nature of hate speech ; (2) the degree of engagement of users in toxic content (denunciation, approbation, reporting and neutral attitude, etc.) ; (3) the question of stereotypes. Furthermore, most of the work (resources, classifiers) is developed for English.

Thus, the scientific challenges are numerous. For Computational Linguistics, the challenge is also linked to the development of methods capable of processing heterogeneous (different topics, various structures and volume) and noisy content (possible presence of abbreviations, smileys or even-sentences in several languages). For Machine Learning methods, a major challenge remains adaptation to a field in constant evolution both in its content (e.g., emerging topics in propaganda rhetoric) and its form. A transverse lock is the constitution of a coherent knowledge base supported by a formal model highlighting both the indices and risk factors provided by sociological models as well as their linguistic anchoring in the content retrieved from the Internet. These challenges show that addressing the threat posed by message and idea propagation to societal security requires a deeper understanding of the linguistic and extra-linguistic content (see for instance Ricardo *et al.,* Chiril *et al.,* Dragos *et al.,* (Ricardo *et al.*, 2018; Chiril *et al.*, 2022; Dragos *et al.*, 2022) on the role of emotion categories in toxic languages; or Poletto *et al.,* and Batistelli et *al.,* about the question of degrees of or engagement in hatefulness (Poletto *et al.*, 2019; Battistelli *et al.*, 2024)).

2. Abusive Language Detection: Current Research and Future Directions

As we said before, ALD has received a growing attention within the field of NLP (Poletto *et al.*, 2021b; Plaza-del Arco *et al.*, 2023; Röttger *et al.*, 2021; Malik

et al., 2024; Nozza *et al.*, 2022), emerging as a fundamental tool for many purposes. Such purposes are ranging from the development of platforms for hate speech monitoring in social media to map vulnerable groups and support policy actions (Capozzi *et al.*, 2020; Laurent, 2020), to the recognition of new targets and vulnerable identities that may become targets of hate speech at a certain historical moment or social climate (Guillén-Pacho *et al.*, 2024); from supporting anti-discrimination educational programs in schools (D'Errico *et al.*, 2024; Cignarella *et al.*, 2023; Cignarella *et al.*, 2024) to moderating online content to prevent the proliferation of hate speech before it causes harm, a purpose as relevant as ever, also considering the recent integration of the Revised EU Code of Conduct on Countering Illegal Hate Speech Online (*Code of Conduct+*) into the Digital Services Act (DSA) regulatory framework, which imposes stricter obligations on online platforms regarding the detection and removal of illegal hate speech. [4]

Recently, the adaptability and flexibility of transformer-based models and Large Language Models (LLMs) led various scholars to focus more and more on exploring and detecting the different nuances that AL could assume depending on diverse contexts, topical focuses and targets. This has encouraged the development of increasingly precise models capable of capturing the specific shapes that AL assumes depending on the affected target, such as misogyny (Rehman *et al.*, 2025; Jiang *et al.*, 2024; Hashmi *et al.*, 2025; Muti *et al.*, 2024; Mohasseb *et al.*, 2025; Pamungkas *et al.*, 2020b), sexism (Plaza *et al.*, 2023; Kirk *et al.*, 2023), homophobic and transphobic discourses (Nozza *et al.*, 2023; Gómez-Adorno *et al.*, 2024). However, even though this research field is now widespread and state-of-the-art models achieve good results, detecting and moderating online abuse remains a complex task, with an increasing awareness of the intertwining of technical, social, legal, and ethical challenges (Cao *et al.*, 2024; Dong *et al.*, 2024; Elesedy *et al.*, 2024).

It remains challenging to provide a univocal definition of what constitutes hate speech (Korre *et al.*, 2025) and to determine the extent to which certain terms should be considered harmful. Different scholars highlighted that AL is commonly a context-dependent phenomenon (Anderson and Barnes, 2022; Brown, 2017; Yoder *et al.*, 2022), and it is often simplistic to classify hate speech using clear-cut boundaries (Parker and Ruths, 2023; Draetta *et al.*, 2024), noting that some terms can assume different meanings depending on the background and the communicative intent of the speaker (Pamungkas *et al.*, 2020a; Pamungkas *et al.*, 2023; Zsisku *et al.*, 2024). For instance, contrastive non-hate variations, such as counter-speech (Yu *et al.*, 2022; Cepollaro *et al.*, 2023; Bonaldi *et al.*, 2024), often blur the line between harmful and not-harmful language.

To properly support content moderation, AL detection systems must be sophisticated enough to identify also hard cases. Recent studies (Dias Oliva *et al.*, 2021; Zsisku *et al.*, 2024; Sap *et al.*, 2019) highlighted that state-of-the-art ALD models

4. `https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online`.

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 in distinguishing between abusive and not-abusive swearing contexts (Pamungkas *et al.*, 2023), disregarding the multifaceted nature of slurs, which are often used with positive social functions (Jay, 2009). In line with this, a still open challenge is the detection of reclamatory uses of slurs (Cepollaro and de Sa, 2023), where members of target groups re-purpose the terms historically used to derogate their group, to express belonging and identity, manifesting solidarity and subverting structures of discrimination. This phenomenon is mostly overlooked in NLP (Zsisku *et al.*, 2024; Draetta *et al.*, 2024; Röttger *et al.*, 2021), and this feeds into the risk of removing legal speech in content moderation, with the paradoxical outcome of hurting the categories of users that one would like to protect. This can be taken as a concrete example for the need to develop socially relevant AL detection models, able to recognize authentic uses in different contexts, embracing new practice of inclusive design in the development of ALD corpora.

Other open challenges, also relevant for mitigating the under-moderation risks in the current ALD systems, are related to the need for a deeper exploration of the nuanced ways online harms manifest (for instance analyzing the relationship between the linguistic expressions of gender-based violence (GBV) in news and responsibility perception (Minnema *et al.*, 2022; Ferrando *et al.*, 2024), and the capability of the ALD systems to recognize also implicit manifestations of abusive language, as the ones featured by the presence of figurative language and sarcastic devices (Frenda *et al.*, 2022; Frenda *et al.*, 2023)).

Looking at the challenge of monitoring users' opinions and hate in online social platforms across time, the availability of large annotated corpora from social media and the development of powerful classification approaches have contributed in an unprecedented way to tackle the issue. Such linguistic data are strongly affected by events and topic discourse, and this aspect is crucial when detecting phenomena such as hate speech, especially from a diachronic perspective. In this context, temporal robustness of hate speech detection and monitoring systems is still a challenge. First findings on data from the real case study of the "Contro l'Odio" platform for monitoring hate speech against immigrants in the Italian Twittersphere (Florio *et al.*, 2020) highlighted the limits that supervised classification models encounter on data that are heavily influenced by events. Future approaches to be investigated could rely, on the one hand, on computational approaches to lexical semantic change detection (Tahmasebi *et al.*, 2018), on the other hand, on techniques of Longitudinal Evaluation of Model Performance that have been recently applied in the context of sentiment analysis in the LongEval CLEF 2023 challenge (Alkhalifa *et al.*, 2023).

Finally, interdisciplinary research and involvement of social scientists, cultural scholars, and practitioners seem to be more and more the key to address the NLP challenges related to the positive final aim of promoting inclusive and fair language,

as several ethical questions arise, particularly concerning how certain linguistic uses are perceived by the target communities. Understanding such perceptions and integrating participatory design methods (Caselli *et al.*, 2021) is crucial for achieving, on the one hand, a more accurate representation of language in ALD datasets (revising common practices in data collection and annotation (Frenda *et al.*, 2024; Madeddu *et al.*, 2023)) and, by extension, cultural diversity in NLP models.

3. Submission Topics

Motivated by the interest of the community in the problem of ALD, we invited papers from Natural Language Processing, Machine Learning, Computational Social Sciences, and Linguistics. We explicitly encouraged interdisciplinary submissions including linguistics resources, methods, end-user applications but also position papers on the actual state of the art in the field discussing the limitations of the current approaches and directions for future work. The topics covered by the special issue include, but are not limited to:

– linguistic resources and evaluation: annotation scheme, corpus linguistics studies, new datasets, with a particular interest on the French language and/or multilingual resources;

– formal/conceptual approaches for AL as inspired by sociological and psychological models;

– models and methods: supervised and non supervised approaches, including LLMs;

– role of contextual phenomena, including discourse, extra-linguistic (e.g., cultural aspects) context;

– models for cross-lingual and multimodal detection;

– new approaches beyond binary classification: target-oriented ALD, degree of user engagement;

– dynamics of online AL in social media, propaganda propagation;

– bias detection and removal in resource creation, datasets and methods;

– application of ALD tools in education, social media content moderation, etc.;

– social, legal, and ethical implications of detecting, monitoring and moderating AL.

The call for papers for this special issue has been launched in February 2024, with a deadline fixed to mid-June 2024.

4. Reviewing and Selection of Papers

Five papers (two in French and three in English) have been submitted, covering a large spectrum in the field ranging from linguistic resource creation, corpus-based

analysis, and automatic detection. We received submissions from Senegal, India, Germany, Italy and France. Each article has been reviewed by three experts: two members of the special issue scientific committee and one member of the journal editorial board. The first round of reviews has been discussed with the editorial board and the guest editors, resulting in the selection of three papers for a second round of reviews, among which two are in English. The final decisions were made in February 2025 where three papers have been accepted, resulting in a selection rate of 60%.

5. Accepted Papers

The aim of this special issue was to report on some recent and innovative methodological angle of attack of what is referred to as Abusive Language circulating on the internet. The accepted papers contribute to this end. Each of them proposes a new dataset related to abusive language (one for French, one for German and one for English) with rigorous indications about the ways the resource has been built. They also offer a set of classification experiments aiming at characterizing and distinguishing abusive language from other types of language. It appears clearly that the classification tasks are necessarily closely linked to how the datasets have been constructed; thus, the ways of investigating correlations between linguistic characteristics and abusive language are necessarily different but offer both interesting results. In a little more detail, the content of the articles is as follows:

– *CyberAgressionAdo-Large: French Multiparty Chat Dataset to Address Online Hate* (by Anaïs Ollagnier, Elena Cabrio, Serena Villata and Valerio Basile) describes a dataset of conversations written by French teenagers involving cyberbullying situations. The adopted methodology for building this dataset consisted in organizing, in close collaboration with sociologists and experts in education, role-playing games addressing different topics (homophobia, religion, etc.) and with annotations belonging to several dimensions (hate, target, verbal abuse, etc.). The paper details the annotation procedure and presents statistical insights to measure divergences across groups of annotations and a study of the most frequent annotated patterns;

– *Comparaison de méthodes pour la détection du discours des incels sur Reddit (Comparing methods for detecting incels' speech on Reddit)* (by Camille Demers and Dominic Forest) addresses the problem of analyzing and then detecting incels' comments in English-speaking forum Reddit. The hypothesis is that incel communities' speech can be violent and therefore be considered as abusive language, particularly against women. The dataset is created by labelling comments according to their community label, not according to their content. Then the learning experiments are based on the Bag-of-Communities method in which a comment is labeled according to the subreddit it originated from. The authors also proposes a set of lexical-based analysis to identify specific lexical units that are more likely to be predictive of incel-type content;

– *Automated Speech Act Classification in Offensive German Language Tweets* (by Melina Plakidis, Elena Leitner and Georg Rehm) presents a manually annotated

dataset of tweets in German according to speech act theory. The hypothesis is that the annotation of speech acts could improve the detection of abusive language. The data used come from the 2018 and 2019 editions of GermEval, a community shared task that focuses on abusive language phenomena. The authors present a correlation study between speech acts and hate-speech annotations with the annotations from the shared task, observing a difference in distribution between the categories. The dataset is then used to train a classifier.

Acknowledgements

We gratefully thank the TAL editors-in-chief, especially Maxime Amblard and Cécile Fabre, for inviting us to coordinate this special issue, and for their supports and guidelines along the whole editing process. We also thank the journal editorial board and the special issue reviewing committee for their work and reactivity: Elena Cabrio (University of Côte d'Azur), Marie Candito (University of Paris Cité), Tommaso Caselli (Faculty of Arts, Rijksuniveristeit Groningen), Vincent Claveau (CNRS IRISA), Valentina Dragos (ONERA), Benoît Favre (University of Aix-Marseille), Claire Hugonnier (University of Grenoble Alpes), Irina Illina (University of Lorraine), Véronique Moriceau (Toulouse University), Frédérique Segond (Inria and INALCO), Didier Schwab (University of Grenoble Alpes), Mathieu Valette (INALCO), Samuel Vernet (University of Aix-Marseille), and François Yvon (CNRS and Sorbonne University).

The work of Farah Benamara is partially supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. The work of Viviana Patti was also partially supported by "HARMONIA" project – M4-C2, I1.3 Partenariati Estesi – Cascade Call – FAIR – CUP C63C22000770006 – PE PE0000013 under the NextGenerationEU programme. Farah Benamara and Viviana Patti have also been supported by the International project STERHEOTYPES (Studying European Racial Hoaxes and sterEOTYPES) funded by the Compagnia di San Paolo and VolksWagen Stiftung under the Challenges for Europe call for Project (CUP: B99C20000640007). The work of Delphine Battistelli was partially supported by the project FLYER (Artificial intelligence for extremist content analysis) – ANR-19-ASTR-0012.

6. References

Alkhalifa R., Bilal I., Borkakoty H., Camacho-Collados J., Deveaud R., El-Ebshihy A., Espinosa-Anke L., Gonzalez-Saez G., Galuščáková P., Goeuriot L. *et al.*, "Extended Overview of the CLEF-2023 LongEval Lab on Longitudinal Evaluation of Model Performance", *CEUR Workshop Proceedings*, vol. 3497, CEUR-WS, p. 2181-2203, 2023.

Anderson L., Barnes M. R., "Hate Speech", *in* E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, The Metaphysics Research Lab, Philosophy Department, Stanford University, 2022.

Basile V., Bosco C., Fersini E., Nozza D., Patti V., Rangel Pardo F. M., Rosso P., Sanguinetti M., "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter", *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, p. 54-63, June, 2019.

Battistelli D., Dragos V., Mekki J., "Annotating social data with speaker/user engagement. Illustration on online hate characterization in French", *Fortino, G., Kumar, A., Swaroop, A., Shukla, P. (eds) Proceedings of Third International Conference on Computing and Communication Networks: ICCCN 2023*, Lecture Notes in Networks and Systems, vol 917. Springer, Singapore, p. 317-330, 2024.

Bonaldi H., Chung Y.-L., Abercrombie G., Guerini M., "NLP for Counterspeech against Hate: A Survey and How-To Guide", *in* K. Duh, H. Gomez, S. Bethard (eds), *Findings of the Association for Computational Linguistics: NAACL 2024*, Association for Computational Linguistics, Mexico City, Mexico, p. 3480-3499, June, 2024.

Brown A., "What is hate speech? Part 2: Family resemblances", *Law and Philosophy*, vol. 36, p. 561-613, 2017.

Cao Y. T., Domingo L.-F., Gilbert S., Mazurek M. L., Shilton K., Daumé Iii H., "Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators through a User-Centric Method", *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, p. 3567-3587, November, 2024.

Capozzi A. T., Lai M., Basile V., Poletto F., Sanguinetti M., Bosco C., Patti V., Ruffo G., Musto C., Polignano M., Semeraro G., Stranisci M., "Contro L'Odio: A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media", *IJCoL (Torino)*, vol. 6, n^{o} 1, p. 77-97, 2020.

Caselli T., Cibin R., Conforti C., Encinas E., Teli M., "Guiding Principles for Participatory Design-inspired Natural Language Processing", *in* A. Field, S. Prabhumoye, M. Sap, Z. Jin, J. Zhao, C. Brockett (eds), *Proceedings of the 1st Workshop on NLP for Positive Impact*, Association for Computational Linguistics, Online, p. 27-35, August, 2021.

Cepollaro B., de Sa D. L., "The successes of reclamation", *Synthese*, vol. 202, n^{o} 6, p. 205, 2023.

Cepollaro B., Lepoutre M., Simpson R. M., "Counterspeech", *Philosophy Compass*, vol. 18, n^{o} 1, p. e12890, 2023.

Chiril P., Pamungkas E., Benamara F., Moriceau V., Patti V., "Emotionally Informed Hate Speech Detection: A Multi-target Perspective", *Cognitive Computation*, vol. 14, p. 322–352, 2022.

Cignarella A. T., Chierchiello E., Ferrando C., Frenda S., Lo S. M., Marra A., "From Hate Speech to Societal Empowerment: A Pedagogical Journey Through Computational Thinking and NLP for High School Students", *in* S. Al-azzawi, L. Biester, G. Kovács, A. Marasović, L. Mathur, M. Mieskes, L. Weissweiler (eds), *Proceedings of the Sixth Workshop on Teaching NLP*, Association for Computational Linguistics, Bangkok, Thailand, p. 54-65, August, 2024.

Cignarella A. T., Frenda S., Lai M., Patti V., Bosco C., "DeactivHate: An Educational Experience for Recognizing and Counteracting Online Hate Speech", *IJCoL (Torino)*, vol. 9, n^{o} 2, p. 1007-1023, 2023.

D'Errico F., Bosco C., Paciello M., Benamara F., Cicirelli P. G., Patti V., Moriceau V., Taulé M., "SteRHeotypes Project. Detecting and Countering Ethnic Stereotypes emerging from Italian, Spanish and French Racial hoaxes", *in* A. Bonet-Jover, R. Sepúlveda-Torres, R. M. Guillena, E. Martínez-Cámara, E. L. Pastor, Á. Rodrigo-Yuste, A. Atutxa (eds), *Proceedings of the Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations (SEPLN-CEDI-PD 2024) co-located with the 7th Spanish Conference on Informatics (CEDI 2024), A Coruña, Spain, June 19-20, 2024*, vol. 3729 of *CEUR Workshop Proceedings*, CEUR-WS.org, p. 77-81, 2024.

Dias Oliva T., Antonialli D. M., Gomes A., "Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online", *Sexuality & Culture*, vol. 25, p. 700-732, 2021.

Dong Z., Zhou Z., Yang C., Shao J., Qiao Y., "Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey", *in* K. Duh, H. Gomez, S. Bethard (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, Mexico, p. 6734-6747, June, 2024.

Draetta L., Ferrando C., Cuccarini M., James L., Patti V., "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, p. 335-342, December, 2024.

Dragos V., Battistelli D., Étienne A., Constable Y., "Angry or Sad? Emotion Annotation for Extremist Content Characterisation", *LREC*, European Language Resources Association, p. 193-201, 2022.

Elesedy H., Esperanca P. M., Oprea S. V., Ozay M., "LoRA-Guard: Parameter-Efficient Guardrail Adaptation for Content Moderation of Large Language Models", *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, p. 11746-11765, November, 2024.

Ferrando C., Madeddu M., Patti V., Lai M., Pasini S., Telari G., Antola B., "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, p. 356-365, December, 2024.

Florio K., Basile V., Polignano M., Basile P., Patti V., "Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media", *Applied Sciences*, 2020.

Fortuna P., Nunes S., "A Survey on Automatic Detection of Hate Speech in Text", *ACM Computing Surveys*, vol. 51, n° 4, p. 85:1-85:30, July, 2018.

Fortuna P., Soler J., Wanner L., "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets", *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 6786-6794, May, 2020.

Frenda S., Abercrombie G., Basile V., Pedrani A., Panizzon R., Cignarella A. T., Marco C., Bernardi D., "Perspectivist approaches to natural language processing: a survey", *Language Resources and Evaluation*, vol. 59, p. 1-28, 2024.

Frenda S., Cignarella A. T., Basile V., Bosco C., Patti V., Rosso P., "The unbearable hurtfulness of sarcasm", *Expert Systems with Applications*, vol. 193, p. 116398, 2022.

Frenda S., Patti V., Rosso P., "When Sarcasm Hurts: Irony-Aware Models for Abusive Language Detection", *in* A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (eds), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, vol. 14163 of *Lecture Notes in Computer Science*, Springer, p. 34-47, 2023.

Gómez-Adorno H., Bel-Enguix G., Calvo H., Ojeda-Trueba S., Andersen S. T., Vásquez J., Alcántara T., Soto M., Macias C., "Overview of homo-mex at iberlef 2024: Hate speech detection towards the Mexican Spanish speaking LGBT+ population", *Procesamiento del Lenguaje Natural*, vol. 73, p. 393-405, 2024.

Guillén-Pacho I., Longo A., Stranisci M. A., Patti V., Badenes-Olmedo C., "The Vulnerable Identities Recognition Corpus (VIRC) for Hate Speech Analysis", *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, CEUR, 2024.

Hashmi E., Yayilgan S. Y., Yamin M. M., Ullah M., "Enhancing misogyny detection in bilingual texts using explainable AI and multilingual fine-tuned transformers", *Complex & Intelligent Systems*, vol. 11, n^{o} 1, p. 39, 2025.

Jay T., "Do offensive words harm people?", *Psychology, public policy, and law*, vol. 15, n^{o} 2, p. 81, 2009.

Jiang A., Vitsakis N., Dinkar T., Abercrombie G., Konstas I., "Re-examining Sexism and Misogyny Classification with Annotator Attitudes", *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, Miami, Florida, USA, p. 15103-15125, November, 2024.

Kirk H., Yin W., Vidgen B., Röttger P., "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, p. 2193-2210, July, 2023.

Korre K., Muti A., Ruggeri F., Barrón-Cedeño A., "Untangling Hate Speech Definitions: A Semantic Componential Analysis Across Cultures and Domains", *Findings at NAACL 2025*, 2025.

Laurent M., "Project Hatemeter: helping NGOs and Social Science researchers to analyze and prevent anti-Muslim hate speech on social media", *Procedia Computer Science*, vol. 176, p. 2143-2153, 2020. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020.

Madeddu M., Frenda S., Lai M., Patti V., Basile V., "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*, vol. 3596 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

Madukwe K., Gao X., Xue B., "In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets", *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Association for Computational Linguistics, Online, p. 150-161, November, 2020.

Malik J. S., Qiao H., Pang G., van den Hengel A., "Deep learning for hate speech detection: a comparative study", *International Journal of Data Science and Analytics*, vol. 12, p. 1-16, 2024.

Minnema G., Gemelli S., Zanchi C., Caselli T., Nissim M., "Dead or Murdered? Predicting Responsibility Perception in Femicide News Reports", *in* Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (eds), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online only, p. 1078-1090, November, 2022.

Mohasseb A., Amer E., Chiroma F., Tranchese A., "Leveraging Advanced NLP Techniques and Data Augmentation to Enhance Online Misogyny Detection", *Applied Sciences*, vol. 15, n° 2, p. 856, 2025.

Muti A., Ruggeri F., Khatib K. A., Barrón-Cedeño A., Caselli T., "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, p. 21091-21107, November, 2024.

Nozza D., Bianchi F., Attanasio G., "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), p. 252-260, July, 2022.

Nozza D., Cignarella A. T., Damo G., Caselli T., Patti V., "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*, vol. 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

Pamungkas E. W., Basile V., Patti V., "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, p. 6237-6246, May, 2020a.

Pamungkas E. W., Basile V., Patti V., "Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study", *Inf. Process. Manag.*, vol. 57, n° 6, p. 102360, 2020b.

Pamungkas E. W., Basile V., Patti V., "Towards multidomain and multilingual abusive language detection: a survey", *Pers. Ubiquitous Comput.*, vol. 27, n° 1, p. 17-43, 2023.

Parker S., Ruths D., "Is hate speech detection the solution the world wants?", *Proceedings of the National Academy of Sciences*, vol. 120, n° 10, p. e2209384120, 2023.

Plaza-del Arco F. M., Nozza D., Hovy D. *et al.*, "Respectful or toxic? using zero-shot learning with language models to detect hate speech", *The 7th Workshop on Online Abuse and Harms (WOAH)*, Association for Computational Linguistics, 2023.

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

Poletto F., Basile V., Sanguinetti M., Bosco C., Patti V., "Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review", *Language Resources and Evaluation*, vol. 55, nº 2, p. 477-523, June, 2021a.

Poletto F., Basile V., Sanguinetti M., Bosco C., Patti V., "Resources and benchmark corpora for hate speech detection: a systematic review", *Language Resources and Evaluation*, vol. 55, p. 477-523, 2021b.

Poletto F., Valerio B., Bosco C., Patti V., Stranisci M., "Annotating hate speech: Three schemes at comparison", *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, vol. 2481 of CEUR Workshop Proceedings, 2019.

Rehman M. Z. U., Zahoor S., Manzoor A., Maqbool M., Kumar N., "A context-aware attention and graph neural network-based multimodal framework for misogyny detection", *Information Processing & Management*, vol. 62, nº 1, p. 103895, 2025.

Ricardo M., Marco G., João A. J., Paulo N., Pedro H., "Hate Speech Classification in Social Media Using Emotional Analysis", *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, p. 61-66, 2018.

Röttger P., Vidgen B., Nguyen D., Waseem Z., Margetts H., Pierrehumbert J., "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, p. 41-58, August, 2021.

Sap M., Card D., Gabriel S., Choi Y., Smith N. A., "The risk of racial bias in hate speech detection", *Proceedings of the 57th annual meeting of the association for computational linguistics*, p. 1668-1678, 2019.

Schmidt A., Wiegand M., "A Survey on Hate Speech Detection Using Natural Language Processing", *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Association for Computational Linguistics, Valencia, Spain, p. 1-10, April, 2017.

Tahmasebi N., Borin L., Jatowt A., "Survey of Computational Approaches to Diachronic Conceptual Change", *CoRR*, 2018.

Talat Z., Hovy D., "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter", *Proceedings of the NAACL Student Research Workshop*, Association for Computational Linguistics, San Diego, California, p. 88-93, June, 2016.

Vidgen B., Derczynski L., "Directions in Abusive Language Training Data, a Systematic Review: Garbage in, Garbage Out", *PLOS ONE*, vol. 15, nº 12, p. e0243300, December, 2020.

Vidgen B., Harris A., Nguyen D., Tromble R., Hale S., Margetts H., "Challenges and Frontiers in Abusive Content Detection", *Proceedings of the Third Workshop on Abusive Language Online*, Association for Computational Linguistics, Florence, Italy, p. 80-93, August, 2019.

Yin W., Zubiaga A., "Towards Generalisable Hate Speech Detection: A Review on Obstacles and Solutions", *PeerJ Computer Science*, vol. 7, p. e598, June, 2021.

Yoder M., Ng L., Brown D. W., Carley K., "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), p. 27-39, December, 2022.

Yu X., Blanco E., Hong L., "Hate Speech and Counter Speech Detection: Conversational Context Does Matter", *in* M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (eds), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, United States, p. 5918-5930, July, 2022.

Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R., "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)", *in* J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (eds), *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, p. 75-86, June, 2019.

Zampieri M., Nakov P., Rosenthal S., Atanasova P., Karadzhov G., Mubarak H., Derczynski L., Pitenis Z., Çöltekin Ç., "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)", *in* A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (eds), *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 1425-1447, December, 2020.

Zsisku E., Zubiaga A., Dubossarsky H., "Hate Speech Detection and Reclaimed Language: Mitigating False Positives and Compounded Discrimination", *Proceedings of the 16th ACM Web Science Conference*, p. 241-249, 2024.