HODIAT: A Dataset for Detecting Homotransphobic Hate Speech in Italian with Aggressiveness and Target Annotation

Greta Damo[1], **Alessandra Teresa Cignarella**[2], **Tommaso Caselli**[3], **Viviana Patti**[4], **Debora Nozza**[5]
[1]Université Côte d'Azur, [2]Ghent University,
[3]University of Groningen, [4]University of Turin, [5]Bocconi University
Correspondence: greta.damo@univ-cotedazur.fr, alessandrateresa.cignarella@ugent.be

Abstract

The escalating spread of homophobic and transphobic rhetoric in both online and offline spaces has become a growing global concern, with Italy standing out as one of the countries where acts of violence against LGBTQIA+ individuals persist and increase year after year. This short paper study analyzes hateful language against LGBTQIA+ individuals in Italian using novel annotation labels for *aggressiveness* and *target*. We assess a range of multilingual and Italian language models on this new annotation layers across zero-shot, few-shot, and fine-tuning settings. The results reveal significant performance gaps across models and settings, highlighting the limitations of zero- and few-shot approaches and the importance of fine-tuning on labelled data, when available, to achieve high prediction performance.

***Warning**: this paper contains obfuscated examples some readers may find upsetting and offensive.*[1]

1 Introduction

The rise of homophobic and transphobic discourse in online and offline spaces has been recently escalating, posing a global concern. Reports from organizations such as ILGA-Europe and Human Rights Watch document a «worrying increase in anti-LGBTQIA+ hate crimes», fueled by a strong growing opposition to the so-called *"gender ideology"*. In several European countries, policies restricting LGBTQIA+ rights have gained traction. Italy is among the countries where the LGBTQIA+ community faces institutional opposition, with the government taking steps against homoparentality further reinforcing queer discrimination. Furthermore, Italy still lacks a national law criminalizing homo-transphobic hate crimes. The proposed DDL Zan law, which aimed to extend anti-discrimination protections to LGBTQIA+ individuals, was blocked in 2021 due to opposition from far-right parties. This legislative vacuum leaves many victims of anti-LGBTQIA+ violence without proper legal protection (Viggiani et al., 2020).

The consequences are tangible both offline and online, where hate speech is often perpetrated, underscoring the urgent need for effective computational tools to detect and mitigate homotransphobic discourse. Despite the critical need, computational research on homophobia and transphobia detection in Italian is underdeveloped. In a previous research, we organized the HODI shared task (Nozza et al., 2023), providing the first benchmark dataset for homotransphobia detection in Italian, while other research, like QUEEREOTYPES (Cignarella et al., 2024), focused on LGBTQIA+ stereotypes, addressing different aspects of the same issue. This study builds on previous work that highlights the importance of aggressiveness and target identification in understanding hate speech dynamics and their potential escalation into real-world violence (Fersini et al., 2020; Basile et al., 2019).

Our contributions are the following:

(1) We propose an extension of the pre-existing HODI dataset, namely HODIAT, enriched with aggressiveness and target annotations to improve the granularity of homotransphobia detection in Italian.[2] Rather than enforcing a harmonized gold standard, we release disaggregated annotations to preserve the subjectivity of annotator perspectives.

(2) A comprehensive evaluation of state-of-the-art NLP models, including GPT-4o-mini, Llama3, Qwen, Minerva, Llamantino, XLM-T, HATE-ITA and ModernBERT, for the detection of homotransphobia, aggressiveness, and target in Italian across three settings: zero-shot, few-shot, and fine-tuning.

[1]Examples have been obfuscated with a Python package for obfuscating profanities: PrOf by Nozza and Hovy (2023).

[2]https://github.com/HODI-EVALITA/HODI_2023.

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 124–135

2 Related Work

Despite growing interest in hate speech detection, research specifically addressing LGBTQIA+ communities remains limited. Developing a hate speech model that effectively covers different targets and languages has proven challenging (Nozza et al., 2023). Indeed, Locatelli et al. (2023) conduct a cross-lingual study on homotransphobia in Twitter discourse, demonstrating that while homotransphobia is a global issue, it manifests through distinct cultural expressions, shaped by factors such as misinformation, cultural prejudices, demographic bias and religious beliefs.

Further exploring the multilingual nature of hate speech, Kumaresan et al. (2024) introduce a dataset for homotransphobia detection in Telugu, Kannada, and Gujarati. Chan et al. (2024) examine the challenges of multilingual LGBTQIA+ hate speech detection, particularly how translation affects detection across English, Italian, Chinese, and English-Tamil code-mixed text. Their findings indicate that fine-tuning consistently improves model performance across languages, whereas translation has mixed effects on detection accuracy.

Moreover, shared tasks have played a crucial role in advancing hate speech detection for the LGBTQIA+ community. The LT-EDI@EACL series focuses on identifying homophobia, transphobia, and non-anti-LGBTQIA+ content across Tamil, English, and code-mixed English-Tamil (Chakravarthi et al., 2024, 2023, 2022). HOMO-MEX is dedicated to the detection of phobic messages towards the Mexican Spanish speaking LGBTQIA+ community (Bel-Enguix et al., 2023; Gómez-Adorno et al., 2024). For Italian, the HODI shared task (Nozza et al., 2023) represents the first initiative focused on homotransphobia detection.

While explicit forms of hate speech have been extensively studied, implicit and subtle forms, such as sarcasm and coded language, have often been overlooked. Recent efforts have focused on detecting implicit forms of hate speech (ElSherief et al., 2021; Muti et al., 2024; Damo et al., 2024b). In a follow-up study, Damo et al. (2024a) also demonstrated that including explanations while detecting implicit hateful messages enhances models' performance.

However, two critical components of hate speech, i.e. detecting aggressiveness and identifying targets of abusive content, are areas that remain under-explored, despite some attention in previous work (Basile et al., 2019; Ibrohim and Budi, 2019; Kumar et al., 2020; Caselli et al., 2021).

Our work builds upon these prior studies by enhancing the HODI dataset for detecting homotransphobia in Italian Twitter discourse, with a particular focus on aggression and target identification. Our study contributes to the growing body of research on online abuse and discrimination against the LGBTQIA+ community and to the creation of resources for languages other than English.

3 Dataset

HODIAT builds on our previous work, the HODI dataset (Nozza et al., 2023), which includes 6,000 Italian tweets binary-labeled (0/1) for hate toward LGBTQIA+ people and annotated with hateful text spans (rationales).[3] This work introduces two novel annotation layers addressing factors that are crucial yet often overlooked in existing research: aggressiveness and target type.

Three independent annotators, all members of the LGBTQIA+ community, were recruited to enrich the data (for details, see positionality statement in Appendix A). They annotated whether the homophobic content **targets an individual or the LGBTQIA+ community as a whole** and they were asked to provide a binary label **(aggressive vs. non-aggressive)** to capture the intensity of hate speech.

The annotation guidelines, including the working definitions of "hate speech", "aggressiveness" and "target", were adapted from established practices in related tasks (Basile et al., 2019; Fersini et al., 2020). The annotation guidelines are available in the same repository of the dataset.[4] These additional layers allow us to (1) differentiate between hate speech aimed at discrediting or isolating individuals versus that which undermines or marginalizes the entire community and (2) gain more insights into the potential harm and urgency associated with the textual content. Below are some examples of possible annotations:

• **TYPE OF TARGET**. It specifies whether the hateful comment addresses a generic group of LGBTQIA+ people or whether it is directed towards a specific individual. The two possible labels are GROUP and INDIVIDUAL.

> Che paese di m*rda, è più importante dar la libertà ai fr*ci di sposarsi che dare il

[3] Please, refer to Nozza et al. (2023) for more details about the original dataset and the shared task.

[4] https://github.com/HODI-EVALITA/HODI_2023.

diritto al lavoro alla gente sia essa fr*cia etera o aliena. Pazzesco
*What a sh*t country, it's more important to give freedom to f*ggots to marry than to give the right to work to people whether they are f*ggots, hetero, or alien. Insane* [**GROUP**]

A [PERS] c*lattone di merda .....ti va bene che non ti incontro per strada ,ti sputerei in faccia m*rdaaaa
A [PERS] *you sh*t f*ggot.....you're lucky I don't meet you on the street, I'd spit in your face, piece of sh*t.* [**INDIVIDUAL**]

• **AGGRESSIVENESS**. The comment contains a message spreading, inciting or promoting violence against LGBTQIA+ people, or a message legitimizing an aggressive action or behaviour that intimidates them. This is a binary category, the possible labels are AGGRESSIVE and NON-AGGRESSIVE.

Il prossimo r*cchione che fa sta roba lo tiro sotto con l'auto
*@user The next f*ggot who does this, I'll run them over with the car.* [**AGGRESSIVE**]

sembra un tr*vione di quelli potenti
*looks like a tr*nny, a huge one*
[**NON-AGGRESSIVE**]

3.1 Agreement Analysis

The annotation was done in three batches, with inter-annotator agreement (IAA) calculated at three time points. Agreement scores are reported in Table 1 using Fleiss' κ. The IAA for target is consistently high across all three batches, with values ranging from 0.71 to 0.74. These values demonstrate strong agreement among annotators, indicating that the task of identifying the message's target was performed reliably.

Fleiss' κ	**1st batch**	**2nd batch**	**3rd batch**
target	0.7066	0.7303	0.7388
aggressiveness	0.5109	0.3345	0.4895

Table 1: Fleiss' κ IAA scores for target and aggressiveness annotations across the three annotation batches.

In contrast, agreement for aggressiveness annotation is notably lower, with Fleiss' κ values ranging from 0.33 to 0.51. The second batch shows the lowest agreement ($\kappa = 0.33$), indicating moderate to fair agreement at best. This suggests that aggressiveness is more subjective.

3.2 The New Dataset: HODIAT

The HODIAT dataset consists of a total of 6,000 Italian tweets (5,000 train + 1,000 test) for homotransphobia detection. Table 2 shows the label distribution (hatefulness, aggressiveness, and target).

	Train	Test	Train %	Test %
Hateful	2,008	511	40.16	51.10
Non-Hateful	2,992	489	59.84	48.90
Individual Target	1,415	336	70.47	65.75
Group Target	593	175	29.53	34.25
Aggressive	104	20	5.18	3.91
Non-Aggressive	1,904	491	94.82	96.09

Table 2: Label distribution. Target and aggressiveness percentages are based on the number of hateful tweets.

To understand the specific behaviours of homotransphobic aggressiveness and targeting, we compare the distribution to HatEval (Basile et al., 2019), a dataset with similar annotations. HatEval, available in English and Spanish with target categories such as women and immigrants, serves as a benchmark against which we compare our dataset's label distribution, revealing several emerging patterns.

Regarding the expression of **target** in HatEval, messages directed at immigrants see an overwhelming majority of the group label rather than individual (94.11% vs. 5.89%). Messages in Spanish present a similar behaviour (86.28% group vs. 13.72% individual). When the target is women, the pattern reverses: in English 64.94% of the hateful content is directed at individuals, (and 35% to general targets), and this individual focus is even more pronounced in Spanish (87.58%).

In HODIAT, hate is more frequently directed at individuals (70.47%) than at groups or generic references (29.53%) similar to hate speech against women in HatEval. These patterns might be related to the intrinsic nature of homotransphobia, which (similarly to misogyny) seems to be often triggered by an individual's perceived violation of social norms, as explained by Manne (2017). On the other hand, racist manifestations of hatred (as in the distribution of the HatEval dataset) seem to operate differently. For instance, in populist rhetoric, it usually manifests through *in-group* versus *out-group* dynamics (Comandini and Patti, 2019), therefore correlating more with the GROUP label.

Aggressiveness in HatEval varies by target group and language. In English, 55.08% of messages targeting immigrants are aggressive, while only

	Hatefulness			Aggressiveness			Target		
	zero-shot	**few-shot**	**fine-tuning**	**zero-shot**	**few-shot**	**fine-tuning**	**zero-shot**	**few-shot**	**fine-tuning**
GPT-4o-mini	0.62	0.65	X	0.27	0.38	X	0.73	0.75	X
LLaMA3	0.56	0.52	0.79	0.12	0.50	0.94	0.34	0.64	0.75
Qwen	0.56	0.55	0.68	0.27	0.58	0.93	0.72	0.68	0.59
Minerva	0.40	0.35	0.73	0.84	0.83	0.94	0.54	0.52	0.77
Llamantino	0.57	0.56	0.50	0.23	0.43	0.93	0.71	0.47	0.61
ModernBERT	—	—	0.74	—	—	0.95	—	—	0.75
HATE-ITA	—	—	0.78*	—	—	0.95*	—	—	0.80*
XLM-T	—	—	**0.82***	—	—	**0.96***	—	—	**0.85***

Table 3: Weighted macro F1 scores by model and dimension. (X) = Output not available due to proprietary model restrictions. (—) = no experiment performed. (*) marks statistically significant results ($p < 0.01$) (see Appendix D).

30.06% are aggressive when targeting women. In Spanish, aggression is higher overall: 68.58% for immigrants and 87.58% for women.

In the HODIAT dataset, this dimension is highly imbalanced, with the aggressive class representing only 5.18% of the total instances. This indicates a strong predominance of non-aggressive content, and presents challenges in drawing definitive conclusions. In future work, we will investigate whether this imbalance might be influenced by the data sampling process or potential annotator bias. Additionally, the limited number of aggressive instances has important implications for our experiments, which we address in Section 5.

4 Experiments

Our experimental setup combines various Large Language Models (LLMs), including GPT-4o-mini (OpenAI, 2024), LLaMa (Meta, 2024), Llamantino (Polignano et al., 2024), Minerva (Sapienzanlp, 2024), and Qwen (Team, 2024), along with transformer-based models such as Modern BERT (Warner et al., 2024), XLM-T (Barbieri et al., 2022), and HATE-ITA (Nozza et al., 2022). We apply the LLMs in zero-shot, few-shot (5 examples), and fine-tuned settings. Finally, we predict the three labels all at once in zero-shot and few-shot settings. Appendix B contains further details about the experimental setting.

5 Results

Our evaluation reveals several key trends in the performance of different models in the hateful, aggressive, and target classification tasks. Table 3 shows the F1 scores for each model across the different settings (zero-shot, few-shot, fine-tuning).

Fine-tuning consistently outperforms both zero-shot and few-shot approaches, particularly for hatefulness and aggressiveness tasks. `XLM-T` achieves the best overall performance and all other models also show improvements (e.g. `Llama3` demonstrates a 0.23 increase in hatefulness and a substantial 0.82 rise in aggressiveness). The improvement in target classification is less pronounced across all models.

Few-shot learning offers limited advantages over zero-shot methods. In some cases, few-shot performance is even lower than zero-shot. This suggests that few-shot prompting does not effectively leverage in-context learning and may introduce noise. `Minerva` shows a strong but imbalanced performance, achieving the highest zero-shot F1 for aggressiveness (0.84), outperforming all other models (scoring below 0.30). However, `Minerva`'s performance suffers from severe class imbalance, particularly for aggressiveness (where scores vary widely between labels) and to a lesser extent for target classification. This makes its predictions less reliable and inconsistent, especially in hatefulness classification, where its zero-shot score is only 0.40. In contrast, `XLM-T` maintains a similar performance across labels. This contributes to its robust fine-tuning performance across all tasks, making it the top-performing model, particularly in target classification and aggressiveness. Similarly, `ModernBERT` shows solid fine-tuning results, confirming that transformer-based models benefit greatly from fine-tuning, especially in tasks like target classification.

The `HATE-ITA` results emphasize the challenges of transferring hate speech tasks across languages and targets. `HATE-ITA` is an `XLM-T` model trained on English and available Italian datasets (focused on hate speech against immigrants and women). Fine-tuning the model solely on HODI (i.e., `XLM-T`) outperforms training on these diverse datasets.

On the other hand, Llamantino, which shows competitive zero-shot and few-shot results, struggles when fine-tuned. Its hatefulness score decreases from 0.57 (zero-shot) to 0.50 (fine-tuning), diverging from the general trend where fine-tuning typically leads to improved performance. This suggests that fine-tuning may not always be the most effective strategy for every model, especially when dealing with certain tasks.

Target classification proved to be the most fluctuating task, with varying performances between models. Minerva achieves the highest LLMs fine-tuned score (0.77), and XLM-T outperforms all models with a score of 0.85. Other models showed fluctuating results, suggesting that target classification may require additional optimization or task-specific tuning to achieve consistent performance.

A notable finding is that training the 3 labels jointly (hatefulness, aggressiveness, and target) tends to worsen performance compared to training them separately, as shown in Table 4 in Appendix C.

6 Conclusion

This paper introduces an enhanced dataset, HODIAT, with additional aggressiveness and target annotations, designed to refine the granularity of homotransphobia detection in Italian, extending the HODI dataset (Nozza et al., 2023). Our findings based on testing several state-of-the-art LLMs and encoder-based models reinforce the importance of fine-tuning when labelled data is available, as it consistently outperforms both zero-shot and few-shot learning. However, even with fine-tuned models, class imbalance remains a challenge, particularly in tasks like aggressiveness and target classification. This highlights the need for further research on addressing bias and improving the stability of model performance across different classes.

Moreover, Minerva's strong but imbalanced performance, along with XLM-T's consistent performance across all tasks, highlights the importance of considering each model's strengths and weaknesses when selecting the best approach for a specific task.

Ethical Considerations and Limitations

This research comes with some ethical considerations and limitations that shall be acknowledged.

First, our study is conducted exclusively on data in Italian and on a task-specific dataset. This inevitably limits the generalizability of our findings to other languages, cultures, and textual genres. We attempted to mitigate this limitation by drawing comparisons with the outcomes of the HatEval dataset (which includes English and Spanish data and targets different social groups: women and immigrants).

Due to the proprietary nature of some language models used in our experiments, we were unable to carry out a fully controlled and uniform experimental setting. Specifically, we were not able to perform fine-tuning on GPT-4o-mini, due to the model's implemented safety guardrails. The output returned the following message *«The job failed due to an invalid training file. This training file was blocked because too many examples were flagged by our moderation API for containing content that violates OpenAI's usage policies in the following categories: hate. Use the free OpenAI Moderation API to identify these examples and remove them from your training data. See* https://platform.openai.com/docs/guides/moderationformoreinformation.*»* This raises ethical concerns about transparency and reproducibility in NLP research, particularly when working with commercial, black-box systems.

Furthermore, while we aimed to include multiple models of different size, it is possible that with more computational resources, larger models could have been utilized, potentially leading to improved performance.

Another limitation concerns the distribution of labels within our dataset (particularly with respect to aggressiveness) which is highly imbalanced. While such imbalance could potentially skew model performance and evaluation metrics, it may also reflect real-world distributions, where aggressive content is relatively rare. Therefore, class imbalance is not inherently problematic, but it requires careful consideration in both modeling and interpretation.

Moreover, our experimental setup relies on evaluation against a gold standard obtained via majority voting. We acknowledge that this approach inevitably reduces the plurality of interpretations on a sensitive and subjective topic, potentially obscuring individual perspectives.

Finally, we recognize that our positionality as researchers may have influenced both our methodological choices and the interpretation of the results. Our opinions are our own and reflect our personal backgrounds, which we detail in our positionality statement (see details in Appendix A). We encour-

age readers to interpret our findings within this context and we welcome critical engagement with our work.

Acknowledgments

The work of G. Damo is supported by the French government, under the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA- 0002. The work of A. T. Cignarella is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions (grant agreement No. 101146287, RAINBOW). The work of V. Patti is partially supported by "HARMONIA" project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 under the NextGenerationEU programme. D. Nozza's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). D. Nozza is a member of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, and S. Ojeda-Trueba. 2023. Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish-speaking lgbtq+ population. *Procesamiento del lenguaje natural*, 71.

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: the Dutch Abusive Language Corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66, Online. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Malliga S, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jimenez-Zafra, Jose Antonio Garcia-Diaz, Rafael Valencia-Garcia, and Nitesh Jindal. 2023. Overview of second shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 38–46, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Fai Leui Chan, Duke Nguyen, and Aditya Joshi. 2024. "is hate lost in translation?": Evaluation of multilingual LGBTQIA+ hate speech detection. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 146–152, Canberra, Australia. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Manuela Sanguinetti, Simona Frenda, Andrea Marra, Cristina Bosco, and Valerio Basile. 2024. QUEEREOTYPES: A multi-source Italian corpus of stereotypes towards LGBTQIA+ community members. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13429–13441, Torino, Italia. ELRA and ICCL.

Gloria Comandini and Viviana Patti. 2019. An Impossible Dialogue! Nominal Utterances and Populist Rhetoric in an Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 163–171, Florence, Italy. Association for Computational Linguistics.

Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2024a. PEACE: Providing Explanations and Analysis for Combating Hate Expressions. In *27th European Conference on Artificial Intelligence*, volume 392 of *Frontiers in Artificial Intelligence and Applications*. IOS Press Ebooks.

Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2024b. Unveiling the Hate: Generating Faithful and Plausible Explanations for Implicit and Subtle Hate Speech Detection. In *International Conference on Applications of Natural Language to Information Systems*, pages 211–225. Springer.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic Misogyny Identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, T. Alcántara, M. Soto, and C. Macías. 2024. Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish-speaking lgbt+ population. *Procesamiento del lenguaje natural*, 73.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Dhruv Sharma, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. Dataset for identification of homophobia and transphobia for Telugu, Kannada, and Gujarati. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4404–4411, Torino, Italia. ELRA and ICCL.

Davide Locatelli, Greta Damo, and Debora Nozza. 2023. A cross-lingual study of homotransphobia on Twitter. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 16–24, Dubrovnik, Croatia. Association for Computational Linguistics.

Kate Manne. 2017. *Down Girl: The Logic of Misogyny*. Oxford University Press, New York. Online edition, Oxford Academic, accessed 7 June 2025.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Meta. 2024. Llama-3.1-8B-Instruct. `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`. Accessed in March 2025.

Arianna Muti, Federico Ruggeri, Khalid Al Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. 2024. Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Giuseppe Attanasio. 2022. HATE-ITA: Hate speech detection in italian social media text. In *Proceedings of the 6th Workshop on Online Abuse and Harms*. Association for Computational Linguistics.

Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, Tommaso Caselli, and Viviana Patti. 2023. HODI at EVALITA 2023: overview of the homotransphobia detection in Italian task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, pages 1–8. CEUR-WS.org.

Debora Nozza and Dirk Hovy. 2023. The state of profanity obfuscation in natural language processing scientific publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3897–3909, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o-mini-2024-07-18. `https://openai.com`. Accessed via ChatGPT in March 2025.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Preprint*, arXiv:2405.07101.

Sapienzanlp. 2024. sapienzanlp/minerva-7b-instruct-v1.0. `https://huggingface.co/sapienzanlp/Minerva-7B-instruct-v1.0`. Accessed in March 2025.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Giacomo Viggiani et al. 2020. Quando l'odio (non) diventa reato. il punto sul fenomeno dei crimini d'odio di matrice omotransfobica in italia. *GenIUS*, 1:1–20.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

A Positionality Statement

Our positionality: This paper is authored by a team of researchers specializing in Natural Language Processing, with diverse yet complementary academic and personal backgrounds. All of us are native Italian speakers; three currently reside in different EU countries, while two are based in Italy. Our professional roles span from junior researchers to senior academics within the European university system. Our collective expertise includes theoretical linguistics, philosophy of language, computer science, data science, natural language processing, and digital humanities. We are united by a shared research focus on hate speech and abusive language, each exploring different facets of these phenomena. Beyond our academic work, we are actively engaged in feminist, LGBTQIA+ advocacy, and anti-hate speech activism. These commitments inform our research perspectives and reinforce our dedication to ethical and socially responsible NLP.

Annotator 1 self-describes as a 28-year-old white Italian non-binary person. Their native language is Italian. They identify as a member of the LGBTQIA+ community and have experienced homotransphobia first-hand. They have a background in social and computer sciences, and are currently pursuing a PhD in Natural Language Processing.

Annotator 2 self-describes as a 31-year-old white Italian man. His native language is Italian. He is part of the LGBTQIA+ community and identifies as gay. He has experienced homotransphobia first-hand. He has a background in Law.

Annotator 3 self-describes as a 27-year-old white Italian man. His native language is Italian. He is part of the LGBTQIA+ community and identifies as bisexual. He has experienced homotransphobia first-hand. He has a background in philosophy and computer science, with a focus on Natural Language Processing.

B Experimental Settings

Data preprocessing. To ensure data quality, we preprocess the tweets by removing special characters, normalizing text, and applying tokenization. We also removed URLs and we anonymized the users mentions. We also analyze the class distribution to identify and address potential imbalances.

Models. Our experimental setup includes a combination of Large Language Models (LLMs) and transformer-based models. We tested a range of LLMs, including GPT-4o-mini, LLaMa, Llamantino, Minerva, and Qwen. All these models are open source and accessible through the Hugging Face platform, except for GPT4, which is accessed through the OpenAI API platform[5]. For GPT, we use the `gpt-4o-mini-2024-07-18` version (OpenAI, 2024), while for the other LLMs we used the following versions: `Llama-3.1-8B-Instruct` (Meta, 2024), `LLaMAntino-3-ANITA-8B-Inst-DPO-ITA` (Polignano et al., 2024), `Minerva-7B-instruct-v1.0` (Sapienzanlp, 2024), and `Qwen2.5-7B-Instruct` (Team, 2024). We choose models of comparable size with learning parameters comprised between 7 and 8 billion parameters. Furthermore, we employed the instruct version of the models, in order to be able to prompt them. These models were also chosen as some of them are multi-lingual, trained also on Italian (GPT, Llama, Qwen, and others are specific for Italian only (Llamantino and Minerva). The hyperparameters used for zero- and few-shots settings are: `max_new_tokens` set to 10, since the output should only be a label, and the task is specified as `text-classification`.

Concerning the transformer-based models, we choose Modern BERT, XLM-T (a cross-lingual transformer model specifically trained on Tweets data), and HATE-ITA (a fine-tuned version of the XLM-T model on English and Italian datasets). Specifically, for Modern BERT we use the `ModernBERT-base` version (Warner et al., 2024), for XLM-T we use the `twitter-xlm-roberta-base` version (Barbieri et al., 2022), while for HATE-ITA we use the `hate-ita` version (Nozza et al., 2022).

Fine-tuning. For the fine-tuned models, we train both transformer-based models and selected LLMs on our dataset under the following conditions: LLMs are fine-tuned for three epochs, except for

[5]https://openai.com/api/

Llamantino, which required five epochs to converge on the Target Detection task. XLM-T and modern BERT models are trained with 3, 5, 7, and 10 epochs, and also use early stopping based on validation loss. A 10% portion of the training set is reserved for validation during all fine-tuning experiments. We employ a `batch size` of 16 for modern BERT and XLM-T models, and a batch size of 8 for the LLMs. We use the `AdamW` optimizer across all fine-tuned models, with `learning rate` of 2e-5. All models are implemented in `PyTorch` using the Hugging Face Transformers library. Training scripts were executed on one NVIDIA A100 GPU. Finally, fine-tuning could not be performed with GPT-4o Mini due to limitations imposed by its safety guardrails, as indicated by the error [6] message given as output.

B.1 Prompting strategy

We employ the following prompts for instructing the LLMs.

Hatefulness classification. For zero-shot and fine-tuning settings, we utilize the following prompt:

You are a hate speech detection model. Given the `MESSAGE`*, classify it as either* `HATEFUL` *or* `NOT HATEFUL`*. Output only the label:* `HATEFUL` *or* `NOT HATEFUL`.

For the few-shot setting, we use the following prompt:

Given the `MESSAGE`*, classify it as either* `HATEFUL` *or* `NOT HATEFUL`*. Consider both classifications equally. Do not assume a message is hateful unless it clearly meets the criteria. Use the examples below as guidance:*

- *"@user_abcdef amo mi mandi ricettina voglio cucinarli al mio r*cchione"* → `NOT HATEFUL`
- *"@user_ab Vi ho sm*rdato ritieniti fortunato che non ti ho taggato e non ho fatto i vostri nomi nel tweet sennò veramente finivi sottoterra r*cchione di m*rda"* → `HATEFUL`
- *"beyonce ha rotto in c*lo tutti i record. be brava. e con zero promo."* → `NOT HATEFUL`
- *"@user_abcdefghi @user_abcdefghij Ritornatene in America b*stardo rotto in c*lo hai scassato la m*nchia buffone"* → `HATEFUL`
- *"scusate ma a me fa proprio strano che ci siano più etero che gay in quel local gay in quella città"* → `NOT HATEFUL`
- *"@user_abcdefgh Solita ch*cca che va allo stadio per fare le foto, a cantare la prossima volta"* → `HATEFUL`
- *"non potete capire come una ragazza lesbica possa guardare la propria migliore amica etero"* → `NOT HATEFUL`
- *"vogliamo mandare i gay (lgbt commnunity) nei forni clematori"* → `HATEFUL`
- *"i brividi che mi sono saliti al sentire amica futura madre dire "spero mio figlio non diventi ch*cca" cioè cominciamo bene proprio"* → `NOT HATEFUL`
- *"1 chi ti da il permesso di seguirmi 2 non fare la lesbica sotto i miei commenti che la f*ssa te la taglio e te la faccio mangiare a colazione [URL]"* → `HATEFUL`

Output only the label: `HATEFUL` *or* `NOT HATEFUL`.

Aggressiveness classification. For zero-shot and fine-tuning settings, we use the following prompt:

You are an aggressiveness detection model. Given the `MESSAGE`*, classify it as either* `AGGRESSIVE` *or* `NOT AGGRESSIVE`*. Output only the label:* `AGGRESSIVE` *or* `NOT AGGRESSIVE`.

For the few-shot setting, we use the following prompt:

Given the `MESSAGE`*, classify it as either* `AGGRESSIVE` *or* `NOT AGGRESSIVE`*. Consider both classifications equally. Do not assume a message is aggressive unless it clearly meets the criteria. Use the examples below as guidance:*

- *"@user_ab Vi ho sm*rdato ritieniti fortunato che non ti ho taggato e non ho fatto i vostri nomi nel tweet sennò veramente finivi sottoterra r*cchione di m*rda"* → `AGGRESSIVE`

[6] «The job failed due to an invalid training file. This training file was blocked because too many examples were flagged by our moderation API for containing content that violates OpenAI's usage policies in the following categories: hate. Use the free OpenAI Moderation API to identify these examples and remove them from your training data. See https://platform.openai.com/docs/guides/moderation for more information.»

- *"@user_abcdefghi @user_abcdefghij Ritornatene in America b*stardo rotto in c*lo hai scassato la m*nchia buffone"* → NOT AGGRESSIVE
- *"@user_abcdefgh Solita ch*cca che va allo stadio per fare le foto, a cantare la prossima volta"* → NOT AGGRESSIVE
- *"vogliamo mandare i gay (lgbt commnunity) nei forni crematori"* → AGGRESSIVE
- *"1 chi ti da il permesso di seguirmi 2 non fare la lesbica sotto i mie commenti che la f*ssa te la taglio e te la faccio mangiare a colazione [URL]"* → NOT AGGRESSIVE

Output only the label: AGGRESSIVE *or* NOT AGGRESSIVE.

Target classification. For zero-shot and fine-tuning settings, we use the following prompt:

You are a target detection model. Given the MESSAGE*, consider if it is directed towards an individual or a group. Classify it as either* INDIVIDUAL *or* GROUP*. Output only the label:* INDIVIDUAL *or* GROUP.

For the few-shot setting, we use the following prompt:

Given the MESSAGE*, consider if it is directed towards an individual or a group. Classify it as either* INDIVIDUAL *or* GROUP*. Consider both classifications equally. Use the examples below as guidance:*

- *"@user_ab Vi ho sm*rdato ritieniti fortunato che non ti ho taggato e non ho fatto i vostri nomi nel tweet sennò veramente finivi sottoterra r*cchione di m*rda"* → INDIVIDUAL
- *"@user_abcdefghi @user_abcdefghij Ritornatene in America b*stardo rotto in c*lo hai scassato la m*nchia buffone"* → INDIVIDUAL
- *"@user_abcdefgh Solita ch*cca che va allo stadio per fare le foto, a cantare la prossima volta"* → INDIVIDUAL
- *"vogliamo mandare i gay (lgbt commnunity) nei forni crematori"* → GROUP
- *"1 chi ti da il permesso di seguirmi 2 non fare la lesbica sotto i mie commenti che la f*ssa te la taglio e te la faccio mangiare a colazione [URL]"* → INDIVIDUAL

Output only the label: INDIVIDUAL *or* GROUP.

Combined classification. For zero-shot setting, we use the following prompt:

Given the MESSAGE*, classify it as either* HATEFUL *or* NOT HATEFUL*. Consider also if it is directed towards an individual or a group and classify it as either* INDIVIDUAL *or* GROUP*. Finally, classify it as either* AGGRESSIVE *or* NOT AGGRESSIVE*. Return the output in the following format with only one label for each classification:*

- *hatefulness:* HATEFUL *or* NOT HATEFUL
- *aggressiveness:* AGGRESSIVE *or* NOT AGGRESSIVE
- *target:* INDIVIDUAL *or* GROUP

For the few-shot setting, we use the following prompt:

Given the MESSAGE*, classify it as either* HATEFUL *or* NOT HATEFUL*. Consider also if it is directed towards an individual or a group and classify it as either* INDIVIDUAL *or* GROUP*. Finally, classify it as either* AGGRESSIVE *or* NOT AGGRESSIVE*. Return the output in the following format with only one label for each classification:*

- *hatefulness:* HATEFUL *or* NOT HATEFUL
- *aggressiveness:* AGGRESSIVE *or* NOT AGGRESSIVE
- *target:* INDIVIDUAL *or* GROUP

Use the examples below as guidance:

- *"@user_abcdef amo mi mandi ricettina voglio cucinarli al mio r*cchione"* → NOT HATEFUL
- *"@user_ab Vi ho sm*rdato ritieniti fortunato che non ti ho taggato e non ho fatto i vostri nomi nel tweet sennò veramente finivi sottoterra r*cchione di m*rda"* → HATEFUL, INDIVIDUAL, AGGRESSIVE

- *"beyonce ha rotto in c*lo tutti i record. be brava. e con zero promo."* → NOT HATEFUL
- *"@user_abcdefghi @user_abcdefghij Ritornatene in America b*stardo rotto in c*lo hai scassato la m*nchia buffone"* → HATEFUL, INDIVIDUAL, NOT AGGRESSIVE
- *"scusate ma a me fa proprio strano che ci siano più etero che gay in quel local gay in quella città"* → NOT HATEFUL
- *"@user_abcdefgh Solita ch*cca che va allo stadio per fare le foto, a cantare la prossima volta"* → HATEFUL, INDIVIDUAL, NOT AGGRESSIVE
- *"non potete capire come una ragazza lesbica possa guardare la propria migliore amica etero"* → NOT HATEFUL
- *"vogliamo mandare i gay (lgbt commnunity) nei forni clematori"* → HATEFUL, GROUP, AGGRESSIVE
- *"i brividi che mi sono saliti al sentire amica futura madre dire "spero mio figlio non diventi ch*cca" cioè cominciamo bene proprio"* → NOT HATEFUl
- *"1 chi ti da il permesso di seguirmi 2 non fare la lesbica sotto i miei commenti che la f*ssa te la taglio e te la faccio mangiare a colazione [URL]"* → HATEFUL, INDIVIDUAL, AGGRESSIVE

C Results

Joint classification. Interestingly, training all three labels (hatefulness, aggressiveness, and target classification) simultaneously generally leads to worse performance compared to training them individually, as demonstrated in Table 4. From there we can see that it is particularly apparent with Llama3 and Qwen, where individual task performance is stronger than when all labels are handled simultaneously. This reinforces the idea that task-specific optimization may be more effective in some cases than multi-task training.

D Statistical significance test

To assess statistical significance, we applied McNemar's test (McNemar, 1947) to the paired binary outputs of the two best performing models, i.e. XLM-T and HATE-ITA, on the shared test set for the three tasks of hate speech detection, aggressiveness detection, and target detection.

For the **hate speech** detection task, McNemar's test yielded a statistic of 69.0 and a p-value of 0.0051. This indicates a statistically significant difference $(p < 0.01)$ between the models' predictions. Despite similar overall performance, the models make different errors on individual instances of hate speech, which suggests they classify instances differently in terms of specific categories.

For the **aggressiveness** detection task, McNemar's test resulted in a statistic of 0.0 and a p-value of 1.53 x 10^{-5}. This test also revealed a statistically significant difference $(p < 0.01)$ in the predictions of the models. The difference is highly significant, suggesting that the two models diverge in their approach to classifying aggressive behaviour in the test set.

For the **target** detection task, McNemar's test showed a statistic of 22.0 and a p-value of 7.01 x 10^{-5}. This result also points to a statistically significant difference $(p < 0.01)$ in the models' predictions, indicating that the models disagree on which instances are considered to contain targets for aggression or hate speech.

Predictions all together									
	Hatefulness			**Aggressiveness**			**Target**		
	zero-shot	**few-shot**	**fine-tuning**	**zero-shot**	**few-shot**	**fine-tuning**	**zero-shot**	**few-shot**	**fine-tuning**
GPT-4o-mini	***0.61***	***0.57***	—	0.24	***0.64***	—	***0.81***	***0.81***	—
LLaMA3	0.55	0.44	—	0.32	0.07	—	0.73	0.76	—
Qwen	0.32	0.51	—	***0.94***	0.41	—	0.17	0.65	—
Minerva	N/A	N/A	—	N/A	N/A	—	N/A	N/A	—
Llamantino	0.58	0.50	—	0.19	0.10	—	0.77	0.74	—

Table 4: F1 scores for all the models used performing all the three classification tasks together at the same time. The results are divided by the different settings: zero-shot, few-shot, and fine-tuning. (—) = no experiment performed