

TAOH

Together Against Online Hate Speech

Deliverable 2.1 Localised Hate Speech Model for Dutch, German and Spanish

About TAOH

The main goal of TAOH - Together Against Online Hate Speech - is to use AI-driven data collection and analysis alongside trainings to strengthen the resilience of civil society, media and authorities in reducing hate speech in the Netherlands, Germany and Spain.

Consortium Members

- SPF - Sharing Perspectives Foundation - Netherlands | Coordinator
- UvA - University of Amsterdam - Netherlands
- Build Up - How to Build Up Germany - Germany
- IBH - ichbinhier e.V. - Germany
- FundiPau - Fundació par la Pau - Spain



**Co-funded by
the European Union**

Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

Violence, Intolerance and Impoliteness thematic Corpus (Dutch, German, Spanish)

This document describes the development of lexicons for the classification of hate speech in German, Dutch and Spanish. These lexicons are part of the project “Together Against Online Hate Speech”, by the following partners: Sharing Perspectives Foundation, BuildUp, ichbinhier e.V., FundiPau, and the University of Amsterdam.

The model was developed by annotating the following datasets:

- German corpus: 1.452.268 comments (598.865 from Instagram, 730.971 from Tiktok, 122.432 from X)
- Spanish corpus: 2.445.553 comments (457.036 from Instagram, 1.953.819 from Tiktok, 34.698 from X)
- Dutch corpus: 5.419.981 comments (917.543 from Instagram, 589.511 from Tiktok, 3.912.927 from X)

Model output

The model classifies language in the following categories:

“Gender” [GEN], “Culture” [CUL], “Race” [RAC], “Sexual orientation” [SOR], “Religion” [REL], “Ableism” [ABL], “Violent language” [VIO], “Insults and personal attacks” [INS], “Vulgar language” [VUL], “Irony and sarcasm” [IRO], “Sexual intimidation” [INT], “Politics” [POL], “Untruth” [UNT], “Crime” [CRI].

Together Against Online Hatespeech (TAOH) – Violence, Intolerance and Impoliteness thematic Corpus (VIIC)

Version 1.1 March 25th, 2026 Van Haperen, Sander (corresponding author: s.p.f.vanhaperen@uva.nl); Ruijgrok, Kris; Van Der Lans, Joyce; Gray, Caitlin; Van Loo, Nick; Shepherd, Anna Claire

Citation:

van Haperen, S.P.F.; Ruijgrok, Kris; van der Lans, J.R.; Gray, Caitlin; van Loo, Nick; Shepherd, Anna Claire (2026). Violence, Intolerance and Impoliteness thematic Corpus. University of Amsterdam. Online resource. <https://doi.org/10.21942/uva.31876402>

Abstract

This document describes the development of lexicons for the classification of hate speech in German, Dutch and Spanish. These lexicons are part of the project “Together Against Online Hate Speech”, by the following partners: Sharing Perspectives Foundation, BuildUp, ichbinhier e.V., FundiPau, and the University of Amsterdam.

Repository Contents

Repository: https://huggingface.co/sanhap/Violence_Intolerance_Impoliteness_VIIC

- A German language lexicon of themes prevalent in violence, intolerance and impoliteness (VIIC_DE);
https://huggingface.co/sanhap/Violence_Intolerance_Impoliteness_VIIC/resolve/main/20260327_TAOH_VIIC_DE_final.json
- A Spanish language lexicon of themes prevalent in violence, intolerance and impoliteness (VIIC_ES);
https://huggingface.co/sanhap/Violence_Intolerance_Impoliteness_VIIC/resolve/main/20260327_TAOH_VIIC_ES_final.json
- A Dutch language lexicon of themes prevalent in violence, intolerance and impoliteness (VIIC_NL);
https://huggingface.co/sanhap/Violence_Intolerance_Impoliteness_VIIC/resolve/main/20260327_TAOH_VIIC_NL_final.json
- The data statement related to VIIC v1.0 (see below);

Project Description

These lexicons are part of the project “Together Against Online Hate Speech”, by the following partners: Sharing Perspectives Foundation, Build Up, IchBinHier, FundiPau, and the University of Amsterdam.

The project is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union.

Neither the European Union nor the granting authority can be held responsible for them. Together Against Online Hate Speech (TAOH) aims to tackle the impact of hate speech in online public spaces by empowering civil society organisations, media, law enforcement, public authorities and individuals to address polarisation and create safer, more inclusive digital spaces.

The main goal is to strengthen the resilience of civil society, media and authorities in reducing hate speech, particularly targeting ethnic minorities, women, gender-diverse and LGBTQ+ individuals. Targets of hate speech will benefit from more accessible resources for support and offline spaces. Initially focused on the Netherlands, Germany and Spain, TAOH's methods will be replicable across the EU. Together Against Online Hate Speech combines AI-enabled language classification with proven human-led dialogue interventions and policy influence to effectively understand and address hateful and discriminatory online speech. The goal of this approach is to address the full spectrum of engagement needed to counter hate speech comprehensively. Content moderation and reporting to authorities is important to address the type of content that is illegal or violates platform policies. However, punishment and content deletion is not enough because hate speech is often legal, for example using coded language or dog whistles. These instances are just as harmful and need to be engaged with. Our approach allows the project to address the full spectrum of hate speech by combining a range of methodologies relevant to individuals, organisations and law enforcement, as well as for policy.

Data Statement for VIIC v1.0

Data set name: Violence, Intolerance and Impoliteness thematic Corpus v1.0

Data set developer(s): Sander van Haperen, Kris Ruijgrok, Joyce van der Lans, Anna Claire Shepherd, Nick van Loo, Caitlin Gray, Pilar Richter, Kamil Nakad, Lea Bund, Paula Magnanimo, Claudia Castro, Anonymous Annotator 1.

Data statement author(s): Sander van Haperen (University of Amsterdam)

Others who contributed to the procedure described in this document: Casper van der Heijden, Bart van der Velden, Christina Hübers, Rita Costa Cots, Mahmoud Bastati, Pablo Aguiar, Maria Raduà, Julie Hawke, Helena Puig Larrauri, and Benjamin Cerigo.

A. Data Collection Strategy

The corpus is the result of the following procedure. First, based on a review of recent related research, we formulated the following definition of “hate speech”: “publicly incite to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin.’ . We also include reference to gender, sexual orientation, and physical ability. At the outset, we expect such language includes categories such as: “Dehumanisation, sexism, politics, racism, religion, violence, and threats” (adapted from EOHH, 2026). This definition provided us with sensitizing concepts and informed calibration among annotators.

Second, we identified accounts of individuals and organizations deemed most likely to receive various forms of online hate, intolerance and incivility, for each of the three countries: Spain, Germany, and the Netherlands. Case selection within the Spanish, German and Dutch settings followed a purposive sampling strategy with the aim of compiling a diverse and comprehensive set of social media accounts that are most likely to receive online hate speech. The overarching objective was to construct a dataset that captures a wide spectrum of hateful expressions, thereby supporting the development of robust detection models sensitive to both explicit and subtle forms of hate. Across all cases, the selection began with the identification of prominent individuals and organizations operating in domains characterized by high public visibility and engagement. Accounts were selected from domains including politics, media, entertainment and culture, public service, academia, activism, sports, and civil society organizations. Within each domain internal diversity was ensured by incorporating subgroups such as elected officials, political parties, journalists, athletes, and influencers. We relied on familiarity with the settings and supplementary sources such as news reports to identify individuals known to receive high volumes of online hate, ensuring that our selection was grounded in observable patterns of harassment. In particular, attention was paid to figures involved in recent or ongoing political, cultural, or media controversies, as well as those associated with controversial issues such as migration, geopolitical conflicts or national debates which are known to increase exposure to online hostility. This sampling strategy aligned with our definition of hate speech while capturing its diverse manifestations across social groups. We identified 377 accounts for Spain, 217 accounts for Germany, and 360 accounts for the Netherlands.

For each of these seed accounts we collected data from social media profiles on X, Instagram, and TikTok. Comments and reactions posted on these accounts, or posts mentioning the account, in the period between 1-1-2024 to 31-12-2025 are included in the selection criteria.

Accounts without an official or identifiable social media presence were excluded. Only publicly available posts and comments are included. No data that was set to “private” or had been deleted between the time of posting and our study are included in our data collection.

This procedure led to three datasets: German corpus: 1.452.268 comments (598.865 from Instagram, 730.971 from Tiktok, 122.432 from X) Spanish corpus: 2.445.553 comments (457.036 from Instagram, 1.953.819 from Tiktok, 34.698 from X) Dutch corpus: 5.419.981 comments (917.543 from Instagram, 589.511 from Tiktok, 3.912.927 from X)

B. Annotation procedure

To identify different forms of hate speech, we adopted an abductive approach to coding (Timmermans & Tavory, 2012). First, three expert annotators were involved in consecutive iterations of coding random samples from these datasets to identify categories of hate speech. Our shared definition of hate speech provided sensitizing concepts and informed discussions about semantics and our objects of analysis. To identify recurring linguistic patterns within user interactions we interpreted posts across accounts within similar domains or sharing identity markers, documenting common keywords, phrases, and expressions of hate speech. This process also included the identification of more subtle forms of harmful language such as coded expressions, slurs, slang, and dogwhistles.

Second, after each round of coding samples of 500 posts, we convened to compare assigned labels, evaluate our definition of hate speech, engage with prior literature, and adjust our code book or add to it. Based on our starting definition we initially expected to find the following categories: race, colour, religion, descent or national or ethnic origin, gender, sexual orientation, and physical ability. We then sought to complement these categories inductively, which led to inclusion of the categories “Untruth”, “Anti-institutional distrust”, “Conspiracies”, “Polarisation”, “Ukraine”, “Russia”, “Autocracy”, “Middle-East”, “Climate”, “Crime”, “Sports”, “Insults”, “Vulgarity”, “Irony and Sarcasm”, “Sexual intimidation”, “Informative”, “Critical”, and “Positivity”.

Third, the resulting list of terms was subsequently reviewed and refined by individuals familiar with each country’s online context to ensure cultural and linguistic relevance. Aside from identifying different categories of hate speech and topic of the post, we initially also scored a measure of intensity from 1 to 10, which was ultimately abandoned after critical reflection on our differing interpretations of intensity. In line with our abductive approach, we iterated between coding and literature review, which led us to distinguish incivility and

impoliteness from hate speech (Nithyanand et al., 2017; Theocharis et al., 2020). We think of incivility defined as “a lack of respect for the social and cultural norms that govern personal interactions and the functioning of democratic systems” (Bentivegna & Rega, 2024). Such disrespect may surface as interpersonal impoliteness or as discourses that delegitimize democratic systems and their actors, including forms such as stigmatization, role-based attacks, demonization, name-calling, threats, incitement to violence, vulgarity, allegations of illegality, and offensive language (Bentivegna & Rega, 2024). After four iterations of coding, we were satisfied the code book and the identified categories adequately covered the data encountered in the fourth sample. At that point, our coding and discussions did not generate new categories or labels, which we took as a saturation point.

Fourth, we applied the code book to annotate out-of-sample data. Twelve expert annotators were trained and involved in applying the code book on new random samples, combined with eight calibration sessions to discuss disagreements over labels. Disagreements were resolved when consensus was reached following discussion among annotators.

Fifth, this was followed by several rounds of axial coding with the goal of relating categories back to our understanding of hate speech as it developed throughout the coding process (Green & Thorogood, 2005). To do so, we considered relationships between categories. As a result of axial coding, several categories were merged. Of others, we determined they were not sufficiently salient or relevant in light of the project’s goals. The final core categories represented in these lexicons are: “Gender” [GEN], “Culture” [CUL], “Race” [RAC], “Sexual orientation” [SOR], “Religion” [REL], “Ableism” [ABL], “Violent language” [VIO], “Insults and personal attacks” [INS], “Vulgar language” [VUL], “Irony and sarcasm” [IRO], “Sexual intimidation” [INT], “Politics” [POL], “Untruth” [UNT], “Crime” [CRI].

Sixth, three rounds of selective coding served to identify keywords (n-gram) from the data where at least three human annotators agreed on the relevant category, and cross-referenced with existing hate speech lexicons: HADES (Tulkens et al, 2016) and HurtLex v1.2 (Bassignana et al, 2018).

Finally, manual curation of the list of keywords led to the exclusion of specific words carrying ambiguous meanings. An example is the word “Let”, which in Dutch signifies a person from Latvia and was therefore part of the “Culture” category, but also more commonly occurs as “Let op” [watch out]. Similarly the word “Frans” [French] occurs more commonly in Dutch as a first name [e.g. Frans Timmermans] than as a reference to French

culture.

The resulting list of keywords per category are the current lexicons.

Limitations remain as a result of prioritizing breadth and diversity. The focus on prominent figures implies that local or less visible actors may be underrepresented. We sought to account for this by purposive sampling on identity. Characteristics such as race, ethnicity, gender identity, sexual orientation, religion, migration background, language, and political ideology were taken into account to ensure the dataset includes groups disproportionately targeted by hate speech but who may be less prominent in the public debate. We sought to represent minority populations specific to each national context, such as ethno-linguistic minorities in Spain and individuals engaged in debates surrounding historically sensitive topics in Germany. Another limitation of our approach is overlap between sectors (such as politics and activism) which complicated efforts to balance across categories. Ultimately, this led to some categories being strongly represented (such as politics) at the expense of others (such as sports). We believe our approach ensures extensive coverage across social domains and identity groups, while accounting for contextual and cultural variation in online hate expression.

Overall, this methodology provides a comprehensive and adaptable framework for cross-national data collection. By systematically identifying accounts operating in the public domain and who are exposed to a variety of online attacks, the resulting dataset reflects the dynamic and rapidly evolving nature of hate speech. With appropriate attention to local context, this approach is replicable and well-suited to supporting the development of detection systems capable of identifying both overt and nuanced forms of online abuse.

C. Language varieties

Dutch, BCP-47 language tag: nl-NL, language variety description: Netherlands and Belgium (Vlaams) German, BCP-47 language tag: de-DE, language variety description: Germany. Spanish, BCP-47 language tag: ES-es, language variety description: Spain.

D. Speaker demographic

N/A

E. Annotator Demographic

N/A

F. Speech situation

N/A

G. Text Characteristics

Posts and comments on X (formerly Twitter); short messages and reactions of approximately 50 words or fewer; may contain multimedia materials, external URL links, and mentions of other users. Time period of collection illustrated in the section Data selection strategy. Posts and comments on Instagram; typically short captions and reactions of approximately 50 words or fewer; may contain multimedia materials, external URL links, and mentions of other users. Time period of collection illustrated in the section Data selection strategy. Posts and comments on TikTok; typically short captions and reactions of approximately 50 words or fewer; may contain multimedia materials, external URL links, and mentions of other users. Time period of collection illustrated in the section Data selection strategy.

H. Recording quality

N/A

I. Sources

- Bentivegna, S., & Rega, R. (2024). Politicians Under Fire: Citizens' Incivility Against Political Leaders on Social Media. *Social Media + Society*, 10(4), 20563051241298415. <https://doi.org/10.1177/20563051241298415>
- Elisa Bassignana, Valerio Basile, Viviana Patti. Hurtlelex: A Multilingual Lexicon of Words to Hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It 2018)*
- European Observatory of Online Hate (2026) "Definitions: Hate speech", <https://eoooh.eu/faq>, [last accessed 19/9/2025]
- Green, J., & Thorogood, N. (2005). Analysing qualitative data. *Principles of social research*, 75-89.
- Kralj Novak, P., Scantamburlo, T., Pelicon, A., Cinelli, M., Mozetič, I., Zollo, F. (2022). Handling Disagreement in Hate Speech Modelling. In: Ciucci, D., et al. *Information Processing and Management of Uncertainty in Knowledge-Based*

Systems. IPMU 2022. Communications in Computer and Information Science, vol 1602. Springer, Cham. https://doi.org/10.1007/978-3-031-08974-9_54

- Nithyanand, R., Schaffner, B., & Gill, P. (2017). Online Political Discourse in the Trump Era (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1711.05303>
- Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The Dynamics of Political Incivility on Twitter. Sage Open, 10(2), 2158244020919447. <https://doi.org/10.1177/2158244020919447>
- Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. Sociological theory, 30(3), 167-186.
- Tulkens, Stéphan, et al. "The automated detection of racist discourse in dutch social media." Computational linguistics in the Netherlands journal 6 (2016): 3-20.

About data statement document

A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

License

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY 4.0)

You are free to: Share — copy and redistribute the material in any medium or format Adapt — remix, transform, and build upon the material The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms: Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. NonCommercial — You may not use the material for commercial purposes. ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits. Notices: You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation. No warranties are given.

The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material. <https://creativecommons.org/licenses/by-nc-sa/4.0/>