ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

# A Taxonomy of Online Hate Speech

Luiz Rogério Lopes Silva

Federal University of Paraná, Brazil

Rodrigo Eduardo Botelho Francisco

Federal University of Paraná, Brazil

Maria do Carmo Duarte de Freitas

Federal University of Paraná, Brazil

## Abstract

This article presents a taxonomy for the representation of online hate speech from a Knowledge Organization perspective, specifically through the concept of event as proposed by the Unified Foundational Ontology (UFO). The methodology articulates taxonomic development approaches, the use of literary and documentary warrants, and Dahlberg's Concept Theory for terminological and conceptual stabilization. The taxonomy comprises 202 classes organized into 10 analytical dimensions, interconnected by 46 properties, and instantiated in 1,697 occurrences, covering participants, discursive propositions, motivations, resources, contexts, consequences, and moderation. The structure was implemented in a digital interface that allows for interactive exploration of the taxonomy and serves as the conceptual foundation for the development of the OntoHate ontology, which is also available in the system via SPARQL queries. The results indicate that taxonomies can function as an initial stage in the conceptual organization of complex informational events, providing support for ontological modeling and content moderation applications on digital platforms.

Keywords: Knowledge Organization; taxonomy; hate speech; ontology; digital platforms

## 1.0 Introduction

The academic study of hate speech consolidated around legal and philosophical disputes that emerged in the United States during the 1980s and 1990s, organized around three distinct analytical frameworks. The first, harm-based, centers on the silencing effects and the psychological and social damage inflicted on historically oppressed minorities (Matsuda 1989). The second, content-based, holds that restrictions should target the stigmatizing character of messages that assail human dignity (Waldron 2012). The third, intent-based, emphasizes the deliberate purpose of the speaker to incite violence or promote hostility (Benesch 2014). Despite their different points of emphasis, these frameworks share a common difficulty, the absence of fixed and necessary properties that would allow hate speech to be identified independently of its context of occurrence. Brown

(2015), Brown (2017a), and Brown (2017b) make this difficulty explicit by arguing that hate speech is better understood as a set of overlapping traits that vary across social and institutional contexts than as a stable category with clear boundaries. This observation indicates that any system for representing the phenomenon must be sensitive to the domain that produces it and to the conditions under which each occurrence takes place.

For the purposes of representation and retrieval, this sensitivity to the domain is precisely what Knowledge Organization offers. Hjørland (2021) demonstrates that the philosophical foundations of a classificatory system determine what it is capable of capturing and that classificatory criteria lose validity when isolated from the domain that produced them. This premise applies directly to the objective of this study, since a taxonomic system for hate speech cannot operate with abstract categories imported from outside the domain. It must reflect the practices, relevance criteria, and needs of the communities that intervene in the problem, including moderators, legal scholars, researchers, activists, and victims. It is in this sense that the taxonomy proposed here anchors its classificatory decisions in the real conditions of production, circulation, and moderation of hate speech in digital environments, treating the domain not as a stable set of terms to be organized, but as an arena of situated practices that defines what counts as hate in each social and technological configuration.

From the perspective of hate speech circulating online, the complexity of the phenomenon intensified with the digitization of public communication, which amplified the reach, speed, and scale of its circulation. Implicit forms of hostility such as sarcasm, coded speech, and stereotypes systematically evade lexical marker-based systems, because their hostility resides not in isolated words but in the relations among utterance, context, and shared cultural code (ElSherief et al. 2021). Recent advances have sought to overcome this limitation through multidimensional taxonomies that incorporate target, intent, and context as classificatory axes (Vidgen et al. 2021), multimodal approaches that extend detection beyond text (Gómez et al. 2020), and sociotechnical analyses that reposition platforms as active agents in defining what circulates (Gillespie 2018; Roberts 2019). What remains incipient, however, is the formalization of these dimensions within a Knowledge Organization framework that integrates domain analysis and formal ontologies into a unified conceptual structure.

This article proposes a taxonomy of online hate speech grounded in the concept of informational event, shifting the focus from isolated lexical analysis to the conditions of production, circulation, and consumption of messages. The proposal does not represent the discovery of an unexplored problem, but a formalization, within Knowledge Organization, of questions already recognized in philosophy, law, and computer science. Its originality lies in the integration of domain analysis (Hjørland 2021), formal ontology (Guizzardi et al. 2022), and informational event modeling into a unified conceptual structure. The taxonomy is part of the OntoHate project, under development since 2020, and constitutes the first enrichment stage of the initial ontology defended in 2022.

It is worth emphasizing that the proposed hierarchy does not replace existing approaches but complements them by adding the event dimension to the representation of the domain.

Earlier initiatives, such as those of Miró Llinares (2016), Salminen et al. (2018), Ousidhoum et al. (2019), and Zampieri et al. (2019), demonstrated the viability of classificatory structures for hate speech, but operate predominantly with stable categories applied to textual properties. The present proposal advances by representing the ontological dependencies among event participants, by incorporating moderation as a constitutive dimension of the phenomenon, and by offering a formally interoperable foundation that allows the taxonomy to be extended to new linguistic and cultural contexts.

## 2.0 Hate Speech as an Informational Event

The Unified Foundational Ontology is a foundational ontology organized into microtheories and a stratified architecture aimed at providing ontological grounding for fundamental notions of conceptual modeling (Guizzardi et al. 2022). Its structure is divided into three main fragments, namely UFO-A, which addresses endurants; UFO-B, which addresses perdurants; and UFO-C, which addresses social entities. Among its core distinctions is the difference between endurants and perdurants. Endurants are individuals that exist in time with all their parts and can change qualitatively while maintaining their numerical identity. Perdurants, referred to here as events, are individuals that unfold in time accumulating temporal parts, are manifestations of dispositions, and exist only in the past. An event does not exist outside the circumstances that constitute it, since it depends on its participants, the relations among them, and the specific conditions under which it unfolds. The difference between the two categories is ontological in nature, since endurants can change while persisting, whereas events cannot genuinely change, as any variation implies distinct temporal parts or changes in the endurants that participate in them.

Grounded in this distinction, this study defines an informational event as a situated occurrence in which agents, contents, intentions, contexts, and sociotechnical infrastructures articulate to produce meaning and social effects. Unlike an informational object, which can be described independently of its conditions of circulation, an informational event is constituted by the relations that produce it and by the participants that depend on it ontologically. Applied to hate speech, this definition implies that no utterance is an intrinsic bearer of the quality of hate. That status emerges exclusively from the dynamic relation among enunciator, target, intentionality, context of production, and effects generated. Hate speech, understood as an informational event, is therefore not a property solely of the message, but an occurrence that depends on the configuration of elements that constitute it.

This distinction has direct theoretical consequences for knowledge representation. A text, an image, or a user profile are endurants, as they exist completely at any given moment and can be described independently of the circumstances in which they circulate. It is not denied that hateful content can be identified through properties of the textual object, but such identification is not reliably possible when dealing with implicit forms of hostility. The same utterance may constitute an affective expression between interlocutors who share a cultural code or a coordinated act of symbolic violence, depending on the relational configuration in which it emerges. Lexicon-based solutions therefore carry the risk of false

positives, when content that activates lexical markers without genuine hostility is incorrectly flagged, or false negatives, when veiled forms of hate such as irony, stereotypes, and coded speech go undetected because they do not trigger those markers.

The trajectory of hate speech representation studies reveals a progression from lexical approaches toward increasingly relational and contextual models. Early systematic automated detection initiatives, such as those of Waseem and Hovy (2016) and Davidson et al. (2017), operated with lists of proscribed terms and binary category annotations, treating hate as an intrinsic property of the text. This model quickly revealed structural limitations, as the ambiguity of the concept and the impossibility of fixing its boundaries in a stable way became evident in both computational and philosophical research (Brown 2015; Nikunen et al. 2025).

In response to these limitations, the literature advanced toward more modular and multidimensional definitions. Melis et al. (2025) systematized this movement by organizing existing definitions into a taxonomy of fourteen conceptual elements, demonstrating that the choice of definition and the degree of specificity of encoded elements such as target, intention, and consequences directly affect model classification performance. This finding reinforces that hate speech is not a stable category, but a phenomenon whose representation depends on the relational elements one decides to include. Vidgen et al. (2021) proposed a dynamic dataset generation process in which human annotators and computational models collaborate iteratively, incorporating variables such as target, intention, and context as classificatory axes and recognizing that hostility does not reside exclusively in words, but in the relations among who speaks, to whom, and under what circumstances.

More recently, architectures such as BERT-OTA have integrated structured ontological knowledge with transformational attention mechanisms, demonstrating that the incorporation of relations among domain categories significantly improves the detection of implicit forms of hate (Abusaqer et al. 2026). This body of developments demonstrates that the interaction between content, context, actors, and technical infrastructure is already a consolidated area of research. The sociotechnical analyses of Gillespie (2018) and Roberts (2019) reinforce this point by repositioning platforms as active agents in defining what circulates, showing that moderation decisions are not external responses to the phenomenon but conditions that determine whether and how the hate event is realized. What remains incipient is the formalization of these dimensions into a unified conceptual structure with a logical and semantic foundation capable of supporting representation, retrieval, and interoperability between information systems.

Building on the works cited above, this study observes that UFO-C offers precisely the constructs needed to formalize what the progression of the literature had already signaled as necessary but had not yet organized on ontological grounds. If early lexical models treated hate as a property of the text, and if more recent literature advanced by incorporating targets, intentions, contexts, and platform infrastructures as analytical dimensions, UFO-C makes it possible to take the next step, translating these dimensions into formally specified ontological dependencies. Constructed on top of UFO-A, which

represents the endurants participating in the event such as agents, targets, and digital artifacts, and UFO-B, which represents the event itself and its temporal conditions of occurrence, UFO-C addresses notions such as Beliefs, Desires, Intentions, Actions, and Social Roles. In UFO-C, actions are intentional events caused by Intentions and oriented toward satisfying the propositional content of some Belief or Desire of the agent, which makes this microtheory directly applicable to the study of hostile communicative phenomena.

This architecture is directly applicable to the concept of informational event proposed in this study. Hate speech is not a property of the message, but a situated occurrence in which a social agent, driven by specific beliefs and intentions, directs hostile content toward a target with protected characteristics, through particular resources, in a context that determines whether and how the event is realized. UFO-C makes it possible to formally represent each of these elements and the dependencies among them, which distinguishes an ontologically grounded representation from a classification system based on isolated attributes. It is this capacity to capture the relational configuration that constitutes each hate occurrence as such that provides the ontological foundation for the taxonomy developed in the following section.

## 3.0 Methodological Procedures

This study is part of the OntoHate project, initiated in 2020 and dedicated to the construction of a formal ontology of online hate speech in the Portuguese language. The first version of the ontology, defended as a doctoral thesis at the Federal University of Paraná in 2022 (Silva 2022), organized the domain around superclasses grounded in the Unified Foundational Ontology, namely *Agente_odiador, Agente_moderador, Destinatário_vítima* and *Manifestação_do_discurso_de_ódio*, all anchored in the *Universal_de_eventos* class of UFO-C and formalized in OWL with a medium degree of formal rigor. The second version of OntoHate is currently under development and foresees the semantic expansion of the original model, the deepening of axiomatization, and the development of applications oriented toward content moderation on digital platforms. The taxonomy presented here constitutes the first methodological exercise of this second version, functioning simultaneously as a domain organization instrument and as a conceptual foundation for subsequent ontological formalization. For its development, a structured and multiphase methodological approach was adopted, inspired by the proposals of Lambe (2007) and Hedden (2016) and guided by the Ontoforinfoscience protocol (Mendonça 2015), following five sequential stages: scope definition, terminological survey, conceptual analysis, taxonomic structuring, and interdisciplinary validation.

The first stage consisted of defining the scope, target audience, and objectives of the taxonomy. The scope was delimited to hate speech in digital environments, with emphasis on the categories necessary to describe the phenomenon as an informational event, which implies representing not only the content of messages but also the agents, conditions of production, circulation mechanisms, and social effects generated. The taxonomy expands

the superclasses of the original version into ten interdependent analytical dimensions: *Núcleo, Natureza, Proposição Odienta, Manifestação, Estratégias Discursivas, Recursos, Motivação, Contexto, Consequência* and *Moderação*. These dimensions are interconnected by formal properties that capture the relationships among event participants, their intentions, the resources mobilized, and the effects produced, maintaining the anchoring in the UFO-C constructs present in the original version, especially the classes *event, intentional mode, mental mode, desire, belief, intention, social agent* and *physical agent*. The delimitation of scope was accompanied by the establishment of criteria for the inclusion and exclusion of terms, with the aim of avoiding unplanned domain expansion and preserving the conceptual coherence of the structure. Categories with consolidated literary warrant or stabilizable through Dahlberg's Concept Theory were included, while those that constituted unplanned domain expansion or conceptual overlap with existing categories were excluded.

In the second stage, a terminological survey was conducted based on literary and documentary warrant, drawing on interdisciplinary and multilingual sources. The sources consulted included scientific literature on hate speech and digital communication published between 2018 and 2024, specialized ontologies and vocabularies, platform policies and technical reports, and national and international legal frameworks. Collaborative hate speech annotation repositories were also consulted, including the datasets of Waseem and Hovy (2016) and Davidson et al. (2017) in English, the Hurlex lexicon, available in more than 50 languages including Portuguese, and Hatebase in its multilingual version, as well as the Portuguese-language glossaries of Silva and Sampaio (2018) and Lopes-Silva et al. (2022). The purpose of this survey was to identify both consolidated terms and recurrent conceptual gaps characteristic of rapidly evolving digital phenomena. In the subsequent filtering and selection stage, the primary criterion applied was the availability of terms in Portuguese from a Brazilian perspective, which resulted in the exclusion of the Waseem and Hovy (2016) and Davidson et al. (2017) repositories from the final terminological base, as they contain exclusively English-language entries, and in the preferential incorporation of the remaining repositories consulted.

The collected terms were then subjected to terminological and conceptual analysis, including the control of lexical variation, synonymy, and ambiguity. At this stage, the Concept Theory proposed by Dahlberg (1978) and Dahlberg (1992) was used as methodological support for the formulation of operational definitions in cases where literary warrant proved insufficient. For each concept without a pre-established definition, triangulation was applied among the reference item, its predicable characteristics, and its verbal designator, enabling the stabilization of emerging concepts and the semantic consistency of the system. Efforts were also directed toward verifying the correspondence between stabilized terms and UFO-C constructs, ensuring that each taxonomic category maintained formal ontological anchoring.

The following stage involved organizing the terms into the ten interdependent dimensions that structure the taxonomy. This organization approximates faceted approaches and allows the phenomenon to be represented through complementary dimensions without requiring a single linear hierarchy, supporting applications in indexing, semantic

annotation, comparative analysis, and interoperability between information systems. The construction was conducted iteratively, with successive refinements based on internal coherence tests, including the verification of generalization and specialization relationships and the identification of possible conceptual overlaps. The ontological relations formalized in the original version of OntoHate, namely *é_um, participa_de, utiliza, ocorre_em, é_vítima_de* and *concretiza_em*, served as formal links between the taxonomic categories and the ontological model, ensuring semantic continuity between the two versions of the project.

Table 1 presents the taxonomic cores resulting from this process, accompanied by their component terms and respective literary and conceptual warrants. The terms are maintained in Portuguese because the taxonomy was originally constructed, validated, and made available in that language, accessible through the OntoHate digital interface (https://luquizis.github.io/Ontohate/). Maintaining the terms in Portuguese is therefore a condition of fidelity to the artifact: readers who access the ontology will find classes and instances named in Portuguese and will be able to identify the English conceptual equivalent of each term through the Object Properties of the model, in accordance with the multilingual structure adopted. Translating the terms in the table would disconnect the textual representation from the ontological artifact, compromising the traceability between the instrument described in the article and the system available for consultation and reuse.

Table 1 Taxonomic cores and literary warrants of the online hate speech taxonomy

| **Taxonomic core** | **Component terms** | **Literary and conceptual warrant** |
|---|---|---|
| **Taxonomic core (central)** | *Universal de evento; Evento de discurso de ódio; Discurso de ódio online; Agente odiador; Hater; Troll; Incels; Dispositivo odiador; Haterbot; Trollbot; Agente sintético; Agente híbrido; Rede de agentes odiadores; Agente influenciador; Agente denunciante; Agente moderador automático; Agente testemunha; Agente moderador; Destinatário vítima; Hater produtor; Hater distribuidor; Hater fantasiador; Hater oportunista; Hater enganador* | UFO – Unified Foundational Ontology Guizzardi et al. (2022); studies on hate speech and cyberhate; Gillespie (2018) (platform moderation); Awan (2016) (typologies of haters); Concept Theory Dahlberg (1978) and Dahlberg (1992) (concept formation and hierarchical organization). |
| **Motivation** | *Diversão; Intenção; Desejo; Aprovação social; Pertencer ao grupo; Arrogância moral; Modo intencional; Modo mental;* | UFO (theory of mental modes and intentionality); literature on motivations |

| | | |
|---|---|---|
| | *Crença; Crença na superioridade; Crença religiosa; Crença na normalidade; Crença na ordem social; Crença na tradição; Proposição odienta* | for hate speech; sociological studies on radicalization; Concept Theory Dahlberg (1978) and Dahlberg (1992) (concept delimitation and conceptual ordering). |
| **Consequence** | *Dano provocado; Dano físico; Pressão alta; Insônia; Dificuldade de respiração; Dano psicológico; Depressão; Sentimento de medo; Raiva interna; Dano social; Desumanização; Falta de poder; Manutenção de estereótipos negativos; Exclusão de espaços; Retirada de oportunidades expressivas; Silenciamento no debate público; Restrições à capacidade de se identificar como membro do grupo* | Gelber and McNamara (2016) (social impacts of hate speech); Brown (2015) (discursive and social effects); Virtual Health Library (BIREME et al. n.d.) (psychosocial effects); Concept Theory Dahlberg (1978) and Dahlberg (1992) (hierarchical structure of consequences). |
| **Target / Victim** | *Pessoa com característica protegida; Mulher cisgênero; Transgênero; Travesti; Gay; Lésbica; Bissexual; Assexual; Pansexual; Queer; Negro; Indígena; Judeu; Muçulmano; Ateu; Agnóstico; Dissidente religioso; Candomblecista; Pessoa com deficiência; Gordo; Obeso; Idoso; Pessoa em situação de rua; Trabalhador sem terra; Estrangeiro; Imigrante; Migrante; Refugiado; Adversário político; Grupo de pessoas com característica protegida; População negra; População indígena; População LGBTQIA+; Movimento dos trabalhadores sem terra; Coletivo político* | GSSO – Gender, Sex and Sexual Orientation Ontology (http://www.ontobee.org/ontology/GSSO) ; IBGE (2011) (race/color classifications); orld Health Organization (2025) (health terminologies); legal frameworks for the protection of minorities; specialized glossaries; Concept Theory Dahlberg (1978) and Dahlberg (1992) (terminological harmonization). |
| **Nature** | *Natureza do discurso de ódio; Natureza principal; Natureza interseccional; Natureza afetiva; Natureza afiliativa; Natureza dispositiva; Natureza por conduta; Natureza por atributos;* | Brown (2015) (typologies of hate speech); Silva and Sampaio (2018) (discursive categories); literature on discrimination and |

| | | |
|---|---|---|
| | *Discurso LGBTQIfóbico; Discurso xenofóbico; Discurso intolerante religioso; Discurso antissemita; Discurso político partidário; Discurso aporofóbico; Discurso misógino; Discurso capacitista; Discurso racialista; Discurso etarista; Discurso de ódio por aparência* | intersectionality; Concept Theory Dahlberg (1978) and Dahlberg (1992) (hierarchy and intersectionality). |
| **Manifestation** | *Manifestação do discurso de ódio; Manifestação explícita; Manifestação velada; Linguagem pejorativa; Linguagem discriminatória; Linguagem excludente* | Silva and Sampaio (2018) (discursive forms); linguistic studies on hate speech; Concept Theory Dahlberg (1978) and Dahlberg (1992) (conceptual delimitation). |
| **Context** | *Contexto do discurso de ódio online; Contexto político; Contexto religioso; Contexto acadêmico; Contexto artístico; Contexto local; Contexto global; Contexto de evento gatilho; Contexto de normas culturais; Contexto algorítmico; Contexto transplataforma; Contexto de plataforma fechada* | Studies on platforms and digital governance; literature on algorithmic mediation; studies in digital communication; Concept Theory Dahlberg (1978) and Dahlberg (1992) (contextual levels). |
| **Resource** | *Texto odiento; Imagem de ódio; Vídeo de ódio; Meme de ódio; Comentário odiento; Hashtag odienta; Recurso do discurso de ódio; Recursos de forma; Recursos técnicos; Bot social; Conteúdo multimodal; Deepfake odiento; Emoji como recurso odiento; Hashtag coordenada; Estratégias discursivas; Etnofaulismo; Incitação à violência; Ameaça; Argumento de superioridade do odiador; Argumento de inferioridade da vítima; Argumento de normalidade do tratamento; Estereótipo negativo; Expulsão; Exclusão; Animalização; Subumanização; Demonização; Conspiração; Inimigos da nação; Direitos negados;* | Silva and Sampaio (2018) (discursive strategies); empirical studies on hate speech in digital platforms; SaferNet Brasil (n.d.) (cybercrime reporting indicators); literature on multimodality and memes; Concept Theory Dahlberg (1978) and Dahlberg (1992) (discursive resources and tones). |

| | | |
|---|---|---|
| | *Desinformação; Humor; Eufemismo odiento; Antonímia odienta; Supremacia branca; Calúnia; Difamação; Qualidade expressiva; Tom do discurso; Tom insultuoso; Tom ameaçador; Tom humorístico; Tom desumanizante; Tom moralizante; Tom de deboche; Tom sarcástico; Tom eufemístico; Tom aparentemente neutro; Tom incitador* | |
| **Moderation** | *Plataforma digital; Infraestrutura técnica; Sistema algorítmico; Política de uso; Regra comunitária; Governança de plataforma; Sistema de moderação; Moderação humana; Moderação automatizada; Indicador de risco odiento; Nível de severidade; Evidência computacional; Decisão de moderação; Alcance; Amplificação; Audiência; Bystander* | Gillespie (2018) (platform governance); Roberts (2019) (content moderation work); literature on algorithmic governance; Concept Theory Dahlberg (1978) and Dahlberg (1992) (decision-making articulation). |

Source: the authors

The validation of the instrument occurred in two distinct moments. In the first, the pre-glossary of terms, concepts, and synonyms produced in the original version of OntoHate was submitted to a panel of hate speech specialists. Twelve professionals and four organizations working at the intersection of human rights, technology, and hate speech were initially invited; eight accepted the invitation and six completed the validation process. The panel was composed of professionals from the fields of Linguistics, Health, Social Work, and Sociology, including researchers with consolidated experience in violence, vulnerable populations, digital communication, and hate speech in social networks. The identity of the evaluators was preserved in accordance with the principle of blind peer review, given the physical and moral risks associated with the public exposure of researchers who address this subject. Each specialist evaluated 150 terms, their definitions, and synonyms, with the possibility of validating, rejecting, or suggesting improvements. The terms and definitions were considered satisfactory by all validators, and all suggestions were accepted, including additions to definitions, orthographic and grammatical corrections, and the inclusion of the term "incel" in the agents subdomain.

In the second moment, the expanded taxonomic structure was submitted to ontological verification and validation conducted by two specialists in ontological modeling, affiliated with federal higher education institutions and with experience in the development of computational ontologies in the field of Information Science. This evaluation verified validation criteria such as ontological commitment, degree of domain representativeness, and instrument expansiveness, as well as verification criteria such as completeness,

integrity, consistency, precision, and documentation of formalized classes and relations. The complete axiomatization of all classes and the development of risk scoring systems for automated moderation constitute prospective stages of the project, planned for subsequent versions of OntoHate.

## 4.0 Results: Visualization and Exploration of the Taxonomy

The main result of this study can be observed on the OntoHate project website (https://luquizis.github.io/Ontohate), where the taxonomy has been implemented as a navigable Knowledge Organization artifact. It is recommended that the reading of this section be accompanied by navigation through the interface, since interaction facilitates the understanding of the conceptual relationships represented in the system and allows the reader to explore the dimensions independently or in combination.

Figure 1 presents the main interface of the system, which allows users to visualize the complete taxonomy or generate specific visualizations for each analytical dimension, examining subsets of classes and their hierarchical relationships within and across dimensions.

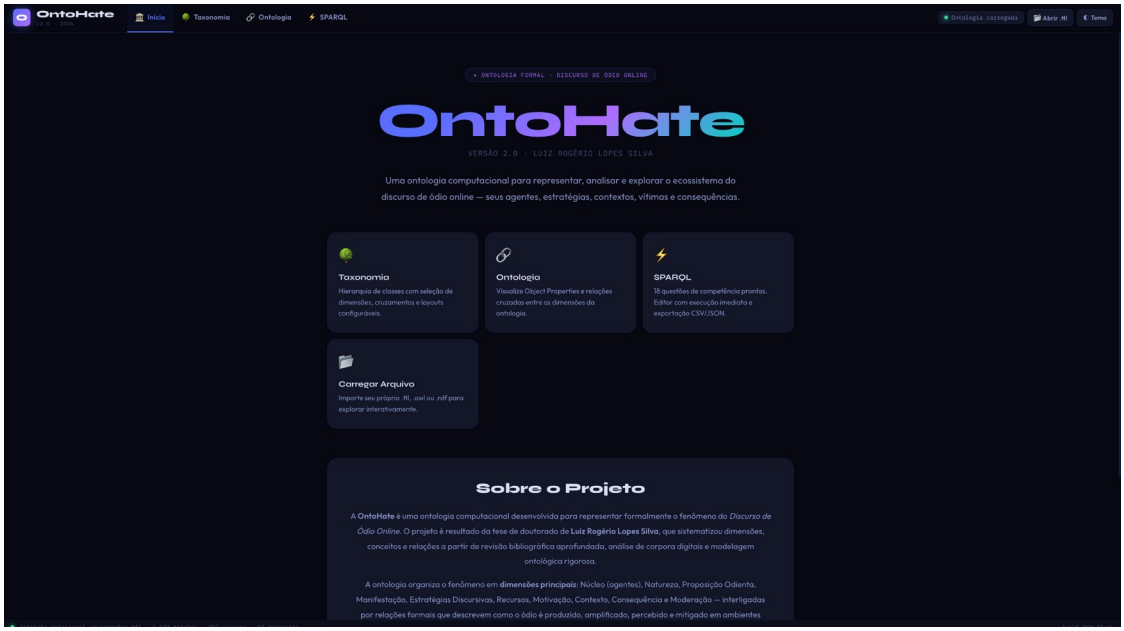


Figure 1. OntoHate project interface

Source: the authors.

The interface materializes the conceptual and methodological decisions described in the previous sections, operationalizing the definition of online hate speech as an informational event through a navigable and interoperable structure. The ten analytical dimensions,

namely *Núcleo, Destinatário-Vítima, Natureza, Proposição Odienta, Manifestação, Estratégias Discursivas, Recursos, Motivação, Contexto, Consequência* and *Moderação*, are interconnected by the formal ontological relations of the original version of OntoHate, ensuring semantic continuity between the two versions of the project. Each dimension gathers sets of concepts that describe specific aspects of the occurrence in digital environments and can be explored independently or in articulation with the others, which gives the system a modular character compatible with different analytical and operational purposes.

This modularity is a central property of the instrument. Since hate speech does not conform to a single linear classificatory hierarchy, the domain requires a structure capable of simultaneously representing its multiple constitutive dimensions. The faceted organization adopted allows the reader to examine, for example, the subclasses of agents without losing sight of their relationship with the motivations that guide their actions or with the contexts in which the event takes place. It is this articulation between dimensions, and not only the internal hierarchy of each one, that constitutes the central contribution of the taxonomy to the representation of the domain.

Figure 2 illustrates a fragment of the taxonomy demonstrating how the dimensions *Núcleo*, *Motivação* and *Destinatário-Vítima* are hierarchically organized. In the *Núcleo* dimension, represented in green, the hierarchy departs from *Discurso de Ódio Online* and unfolds into subclasses such as *Agente Odiador*, with specializations including *Hater*, *Troll*, *Incels*, *Haterbot* and *Rede de Agentes Odiadores*, and *Destinatário-Vítima*, which branches into *Pessoa com Característica Protegida*, gathering groups such as *População Negra*, *Indígena*, *LGBTQIA+*, *Imigrantes*, *Refugiados* and *Pessoas com Deficiência*, among others. In the *Motivação* dimension, represented in orange, the hierarchy departs from *Modo Intencional* and *Modo Mental*, unfolding into categories such as *Desejo*, *Intenção*, *Diversão* and *Crença*, the latter specialized into subcategories including *Crença na Superioridade*, *Crença Religiosa*, *Crença na Normalidade*, *Crença na Ordem Social* and *Crença na Tradição*. The modular structure allows each dimension to be analyzed independently or in relation to the others, showing how different aspects of the informational event articulate in the representation of the domain.

The taxonomic hierarchy represented in the figure evidences a fundamental property of the instrument, namely the capacity to progressively specify the participants of the event and their characteristics, from the most general to the most specific level, without losing semantic coherence across levels. This progressive specificity is what distinguishes the taxonomy from simpler classifications based on flat categories, and is also what makes it viable as a conceptual foundation for future computational applications. Among these applications, planned for subsequent versions of the project, are specialized ontological queries and risk scoring mechanisms for automated moderation that systematically consider factors such as intentional intensity, the profile of the agents involved, and the contextual reach of each event. These applications are not yet implemented in the current artifact and will be documented in subsequent articles of the project.

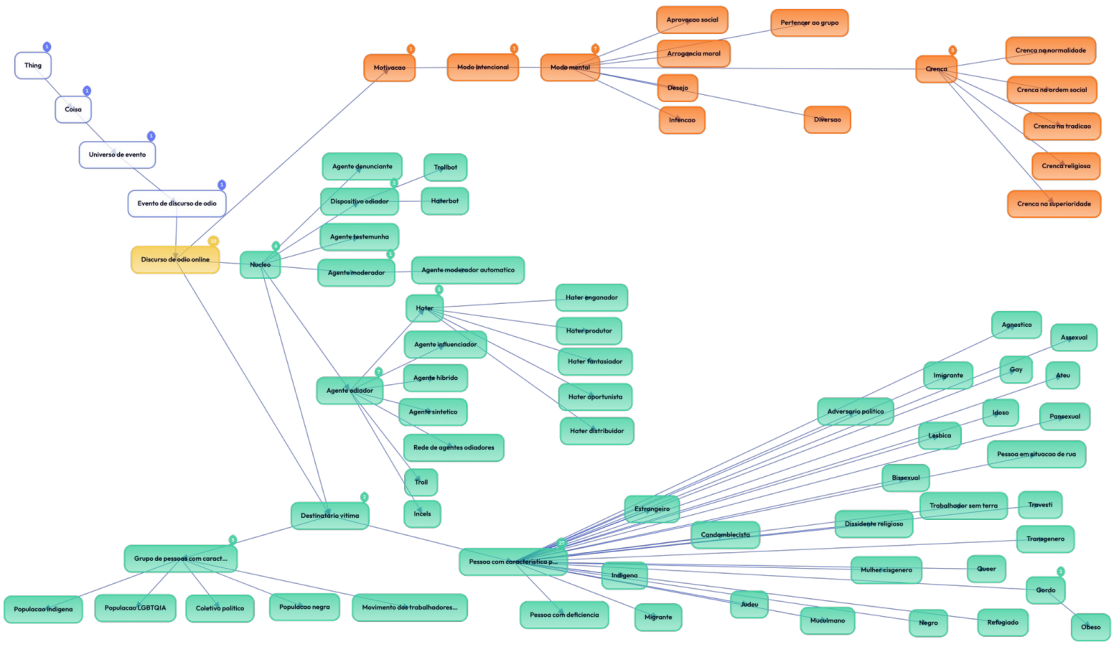


Figure 2. Taxonomic structure showing a left-to-right visualization of one dimension of online hate speech.

Source: the authors.

It is worth noting that the OntoHate project website offers the reader access to a broader set of artifacts beyond the taxonomy described here. By navigating the interface, it is possible to find the formal ontology in OWL, with its classes, relations and axioms, as well as a SPARQL query module that allows semantic information about the domain to be retrieved through thirty competency questions with answers generated in natural language. These components are under continuous development and will be documented and discussed in subsequent articles of the project. The present study is delimited to the description and analysis of the taxonomic stage, which constitutes, as described in the methodology, the first methodological exercise of the second version of OntoHate and the conceptual foundation upon which the remaining artifacts will be progressively formalized.

## 5.0 Final considerations

Representing online hate speech as an informational event within a taxonomy required the identification and conceptual organization of the elements that structure its occurrence in digital environments. The main objective of this study was to develop a classificatory system capable of systematically organizing the domain through categories and relationships that describe the articulation between participants, propositions, motivations, contexts, manifestations, and consequences. By treating hate speech as an event rather than a textual property, the taxonomy makes explicit the relational and

situated nature of the phenomenon, moving beyond approaches that reduce it to the presence of specific lexical markers.

From the perspective of Knowledge Organization, this work demonstrates the value of taxonomies as an initial instrument for complex and dynamic domains. The structure developed here makes explicit the main components of the event and their interactions, revealing relational patterns that often escape traditional classificatory approaches. The faceted and modular organization adopted allows different dimensions of the domain to be examined independently or in articulation, supporting analytical and operational applications that require sensitivity to context, intentionality, and the sociotechnical conditions of circulation. At the same time, the relational and multidimensional nature of the phenomenon points to the need for more formal representations to enable computational analysis at scale, which motivates the evolutionary continuity of the project.

The taxonomy presented here establishes the conceptual foundations for the transition from the initial classificatory structure to the second version of the OntoHate ontology, which will computationally formalize the classes and relationships outlined in this study. Through this ontology, advanced semantic queries become feasible directly within the project's digital interface, enabling the identification of complex risk patterns across large volumes of digital content. This progression also points toward the development of graduated risk scoring systems for automated moderation, which will support proportional and contextually informed decisions, overcoming the limitations of traditional binary classification approaches. These developments are planned for subsequent versions of the project and will be documented in future publications.

The taxonomy also presents limitations that must be acknowledged. The primary limitation concerns the complexity of the domain, which involves social, political, and cultural categories in constant evolution, potentially requiring revision or refinement as new forms of hate speech emerge in digital environments. A second limitation relates to the analytical character of the proposed structure, which seeks to represent the dimensions of the informational event but does not aim to capture all empirical variations observed across specific platforms or linguistic communities. These limitations are inherent to the first stage of an iterative ontological project and are expected to be progressively addressed in subsequent versions of OntoHate.

Although the taxonomy reflects a conceptual framework grounded in the literature analyzed between 2018 and 2024 and in the methodological choices adopted in this study, it is intended as a dynamic and collaborative starting point. As such, it remains open to incorporating the continuous transformations of hate speech in Brazilian digital platforms and to supporting the development of broader analytical and intervention-oriented solutions in the future. The OntoHate project, initiated in 2020 and now entering its second developmental phase, positions itself as a long-term contribution to the field of Knowledge Organization, offering a formal and evolving instrument for the representation of one of the most pressing challenges of contemporary digital communication.

AI USAGE STATEMENT: The authors used a generative AI assistant in two ways during the preparation of this work. First, to support language editing and revision of the text, focusing on clarity, grammar, and style and on assisting with translation between Portuguese and English, notably in the introduction and in Section 2. Second, to support the development of the website that hosts the OntoHate ontology, available at luquizis.github.io/Ontohate/, assisting with the writing and structuring of the page's code. The conceptualization, the design of the ontology, the analysis, and all scholarly claims are the authors' own. The authors reviewed and edited every AI-assisted passage of text and code and take full responsibility for the final content.

References

Abusaqer, Mahmoud, Jamil Saquer, and Mukulika Ghosh. 2026. “BERT-OTA: Enhancing Hate Speech Detection with Ontology-Guided Transformer Attention.” IEEE Access 14: 3345–3358. doi: 10.1109/ACCESS.2026.3650874.

Awan, Imran. 2016. “Islamophobia on Social Media: A Qualitative Analysis of the Facebook Pages of Far-Right Political Parties.” Information, Communication & Society 19: 1–18.

Benesch, Susan. 2014. “Countering Dangerous Speech: New Ideas for Genocide Prevention.” Working paper. Washington, DC: United States Holocaust Memorial Museum.

BIREME, PAHO, and WHO. n.d. Biblioteca Virtual em Saúde (BVS). Accessed February 4, 2026. https://bvsalud.org/.

Brown, Alexander. 2015. Hate Speech Law: A Philosophical Examination. London: Routledge.

Brown, Alexander. 2017a. “What Is Hate Speech? Part 1: The Myth of Hate.” Law and Philosophy 36: 419–468.

Brown, Alexander. 2017b. “What Is Hate Speech? Part 2: Family Resemblances.” Law and Philosophy 36: 561–613.

Dahlberg, Ingetraut. 1978. “A Referent-Oriented Analytical Concept Theory for INTERCONCEPT.” International Classification 5: 142–151.

Dahlberg, Ingetraut. 1992. “Knowledge Organization: Its Scope and Possibilities.” Knowledge Organization 19: 67–71.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. “Automated Hate Speech Detection and the Problem of Offensive Language.” In Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017), 512–515. Palo Alto, CA: AAAI Press. doi: 10.1609/icwsm.v11i1.14955.

ElSherief, Mai, 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, edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 345–363. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.29.

Gelber, Katharine, and Luke McNamara. 2016. "Evidencing the Harms of Hate Speech." Social Identities 22: 324–341. https://doi.org/10.1080/13504630.2015.1128810.

Gillespie, Tarleton. 2018. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven: Yale University Press.

Gomez, Raul, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. "Exploring Hate Speech Detection in Multimodal Publications." In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 1459–1467. Piscataway, NJ: IEEE. https://doi.org/10.1109/WACV45572.2020.9093414.

Guizzardi, Giancarlo, Alessander Botti Benevides, Claudenir M. Fonseca, Daniele Porello, João Paulo A. Almeida, and Tiago Prince Sales. 2022. "UFO: Unified Foundational Ontology." Applied Ontology 17: 167–210. doi: 10.3233/AO-210256.

Hedden, Heather. 2016. The Accidental Taxonomist. 2nd ed. Medford, NJ: Information Today.

Hjørland, Birger. 2021. "Information Retrieval and Knowledge Organization: A Perspective from the Philosophy of Science." Information 12, no. 3: 135. doi: 10.3390/info12030135.

IBGE (Instituto Brasileiro de Geografia e Estatística). 2011. Características étnico-raciais da população: um estudo das categorias de classificação de cor ou raça 2008. Rio de Janeiro: IBGE. https://biblioteca.ibge.gov.br/visualizacao/livros/liv49891.pdf.

Lambe, Patrick. 2007. Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness. Oxford: Chandos Publishing.

Lopes-Silva, Luiz Rogério, Rodrigo Eduardo Botelho Francisco, Vânia Moreno de Oliveira, and Tarcísio Ramos. 2022. "Thou Shalt Not Take the Lord's Name in Vain: A Methodological Proposal to Identify Religious Hate Content on Digital Social Networks." International Journal of Communication 16: 2198–2219.

Matsuda, Mari J. 1989. "Public Response to Racist Speech: Considering the Victim's Story." Michigan Law Review 87: 2320–2381.

Melis, Matteo, Gabriella Lapesa, and Dennis Assenmacher. 2025. "A Modular Taxonomy for Hate Speech Definitions and Its Impact on Zero-Shot LLM Classification Performance." arXiv:2506.18576. doi: 10.48550/arXiv.2506.18576.

Mendonça, Fabrício Martins. 2015. OntoForInfoScience: metodologia para construção de ontologias pelos cientistas da informação – uma aplicação prática no desenvolvimento da ontologia sobre componentes do sangue humano (HEMONTO). Ph.D. dissertation. Belo Horizonte: Universidade Federal de Minas Gerais. https://repositorio.ufmg.br/handle/1843/BUBD-A35H3K.

Miró Llinares, Fernando. 2016. "Taxonomía de la Comunicación Violenta y el Discurso del Odio en Internet." Revista Electrónica de Ciencia Penal y Criminología 18, no. 13: 1–29.

Nikunen, Kaarina, Paula Haara, Heidi Kosonen, Aleksi Knuutila, Reeta Pöyhtäri, and Tuija Saresma. 2025. "Contested Meanings of Hate Speech and the Post-Truth Condition on Digital Platforms." Social Media + Society 11, no. 2. doi: 10.1177/20563051251341794.

Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. "Multilingual and Multi-Aspect Hate Speech Analysis." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 4675–4684. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1474.

Roberts, Sarah T. 2019. Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven: Yale University Press.

Salminen, Joni, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. "Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media." In Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018), 330–339. Palo Alto, CA: AAAI Press. doi: 10.1609/icwsm.v12i1.15028.

SaferNet Brasil. n.d. "Indicadores do Canal de Denúncias de Crimes Cibernéticos." Accessed February 4, 2026. https://indicadores.safernet.org.br/.

Silva, Luiz Rogério Lopes. 2022. OntoHate: Uma Ontologia do Discurso de Ódio em Plataformas Digitais no Brasil. PhD dissertation. Curitiba: Universidade Federal do Paraná.

Silva, Luiz Rogério Lopes, and Rafael Cardoso Sampaio. 2018. "Impeachment, Facebook e Discurso de Ódio: A Incivilidade e o Desrespeito nas Fanpages das Senadoras da República." Esferas 1, no. 10: 95–107. doi: 10.31501/esf.v1i10.9334.

Vidgen, Bertie, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. "Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection." In 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), edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 1667–1682. Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.132.

Waldron, Jeremy. 2012. The Harm in Hate Speech. Cambridge, MA: Harvard University Press.

Waseem, Zeerak, and Dirk Hovy. 2016. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.” In Proceedings of the NAACL Student Research Workshop, edited by Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou, 88–93. San Diego, CA: Association for Computational Linguistics. doi: 10.18653/v1/N16-2013.

World Health Organization. 2025. “Obesity and Overweight.” Fact sheet, December 8. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. “SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval).” In Proceedings of the 13th International Workshop on Semantic Evaluation, edited by Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, 75–86. Minneapolis, MN: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2010.