Modeling Linguistic Violence: An Ontology-Based Framework for the Computational Analysis of Violence Manifested in Language

Brenda Salenave Santana[1], **Ana Marilza Pernas**[1], **and Aline A. Vanin**[2],
[1] PPGC – UFPEL, Pelotas, RS, Brazil
[2] Dept. of Education and Humanities, UFCSPA, Porto Alegre, RS, Brazil
brenda@inf.ufpel.edu.br marilza@inf.ufpel.edu.br alinevanin@ufcspa.edu.br

Abstract

The conceptual ambiguity among terms like 'hate speech', 'toxic speech', and 'dangerous speech' creates a significant bottleneck for both research and automated moderation. Traditional NLP models, often focused on lexical cues, struggle to differentiate these nuanced forms of linguistic violence, especially when the harm is implicit. This paper addresses this gap with a twofold objective. First, we conduct a conceptual review and propose a unified ontology that differentiates these concepts—including verbal aggression and cyberbullying—based on their core attributes, such as their target, intent, and associated rhetorical hallmarks. Second, we propose a computational methodology designed to operationalize this ontology. Our framework uses a multi-stage NLP pipeline that leverages semantic analysis, specifically Semantic Role Labeling and Named Entity Recognition, to deconstruct speech acts into their core components (e.g., target and action). This component-based approach allows for a granular classification that can robustly distinguish between seemingly similar phenomena, such as a general insult and a targeted identity-based attack. This methodology is particularly promising for low-resource languages, such as Portuguese, as it relies on core semantic tasks for which multilingual models are available, rather than requiring massive, task-specific labeled datasets.

1 Introduction

Linguistic violence can manifest in different ways, either through the direct use of insulting language aimed at an individual who is seen as a target, or through discourse that, while not directed at a specific person, legitimizes or glorifies violent practices—such as speeches that exalt white supremacy. Nevertheless, as pointed out by Butler (2021), it is not just circumstances that make words hurt. Alternatively, as stated by the authors, we could be led to claim that all words are susceptible to being words that hurt, depending on how they are used, and that the use of words is not reducible to the circumstances of their utterance. As uttered by Toni Morrison upon receiving the Nobel Prize for Literature in 1993[1], "Oppressive language does more than represent violence; it is violence".

In the computational literature and platform policy, these harmful linguistic practices are often conceptualized through a range of overlapping terms, such as dangerous speech, toxic speech, and hate speech. A significant challenge in this field is that these critical terms are frequently used interchangeably, with "blurred boundaries" among them. For hate speech in particular, there is no single, consensually accepted definition in the literature. This ambiguity is a critical bottleneck for research and moderation, as differences in definition inevitably lead to inconsistencies in content classification.

This problem is exacerbated when attempting to detect linguistic violence automatically. Current approaches, often focused on lexical analysis, struggle to fight against implicit, harmful texts. Understanding the context, irony, or implied meaning necessary to distinguish these concepts requires moving beyond surface-level features.

This study has a twofold objective to address this conceptual and methodological gap. First, we conduct a conceptual review of the basic terms for linguistic violence, including hate speech, toxic speech, dangerous speech, and verbal aggression. Based on this review, we propose a unified ontology that differentiates these concepts based on their core attributes, such as their target, intention, and associated rhetorical hallmarks.

Second, we propose a computational methodology to operationalize this ontology. This framework leverages deep semantic analysis, specifically, preprocessing tasks like Semantic Role Labeling

[1] See https://www.nobelprize.org/prizes/literature/1993/morrison/lecture/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 291–301

(SRL) and Named Entity Recognition (NER), to decompose speech acts into core components (e.g., who is the target and what is the action). This allows for a granular classification that can robustly distinguish between seemingly similar phenomena, such as a general insult (toxic speech) and a targeted, identity-based attack (hate speech).

Our approach positions the detection of linguistic violence within the broader context of structured prediction and multi-label classification. By moving beyond flat, single-label paradigms, we propose a framework that recognizes the overlapping nature of harmful speech, such as the intersection between identity-based hate and incitement to violence. Specifically, our framework differentiates these concepts along their primary axes: Dangerous Speech (by its Intent: to incite violence), Toxic Speech (Effect: conversation degradation), Hate Speech (Target: protected group), Verbal Aggression (Act: a personal insult), and Cyberbullying (Pattern: repetitive harm).

This paper is structured as follows. Section 2 reviews the conceptual foundation of linguistic violence, defining the key rhetorical hallmarks and differentiating the core concepts. Section 3 formally presents our proposed ontology, organizing these concepts into a unified framework based on their attributes and associated hallmarks. Section 4 details the computational methodology designed to operationalize this ontology, presenting a multi-stage Natural Language Processing (NLP) pipeline that leverages semantic analysis. Section 5 presents a case study followed by a discussion of the potentialities of the proposal. Finally, Section 6 provides our concluding remarks and outlines directions for future work.

2 Conceptual Foundation: The Building Blocks of Linguistic Violence

In the literature, different but similar terms can be framed as symbolically harmful speeches (e.g., dangerous, toxic, hate, etc.). When dealing with hate speech-related literature, different terms come up with blurred boundaries (Jahan and Oussalah, 2021). However, although similar, these concepts differ from each other. Below, we present the conceptualization of the main associated topics. The order in which the definitions are presented follows our understanding of the most comprehensive concept to the least comprehensive concept.

This section discusses the theoretical foundations underlying violent and discriminatory language phenomena, highlighting their conceptual intersections and distinctions. While related, each type of harmful discourse—hate speech, toxic speech, and violent language—presents unique linguistic and pragmatic hallmarks that guide the development of the proposed ontology.

2.1 Hallmarks of Linguistic Violence

As cited in these given definitions and also from the study of research (de Barros, 2014; Leader Maynard and Benesch, 2016), some hallmark characteristics are pointed out for those symbolically violent speeches, as they are described in the sequence.

1. *Dehumanization:* Speakers may convince their listeners to deny others any of the moral respect they offer to those who are "fully" human by portraying other groups of people as anything other than human, or less than human (Leader Maynard and Benesch, 2016). Dehumanizing purposes prepare followers to condone or commit violence by making their targets' deaths and misery seem less critical or even helpful or required.

2. *Accusation in a Mirror:* Believing that you, your family, your community, or even your society is facing an existential threat from another group makes it seem not just reasonable (as dehumanization does) but necessary to fend off that threat.

3. *Break of Social Contracts of Integrity or Purity*

   a) *Threat to group Integrity or Purity:* claim that members of another group can cause irreparable damage to the integrity or purity of one's own group.

   b) *Assertion of Attack Against Women and Girls:* It is the suggestion that women or girls of the in-group have been or will be threatened, harassed, or defiled by members of an out-group. In many cases, the purity of women symbolizes the group's purity, identity, or way of life.

   c) *Sanction Speech:* Such discourses intend to punish subjects considered bad complaints of certain social contracts. Those who defend ideas out of what is socially expected are a target of these speeches;

4. *Questioning In-Group Loyalty:* In a dangerous speech, out-group or target group members are generally identified; some of it never mentions them, instead characterizing in-group members as insufficiently loyal or even treacherous to be sympathetic to the out-group.

5. *Figures of Opposition*

    a) *Passionate Hate and Aversion to the Different ones:* Speeches in which the passions of hatred and fear prevail in relation to what is considered different. These occur from antipathy to homophobia, racism, xenophobia, and misogyny, among others;

    b) *Themes and Figures of Opposition:* Speeches that develop themes and figures from the opposition between equality or identity and difference.

To better organize these rhetorical strategies, we grouped these hallmarks into five major sets (Dehumanization; Accusation in a Mirror; Break of Social Contracts of Integrity or Purity; Questioning In-Group Loyalty; and the Construction Figures of Opposition). Table 1 presents these rhetorical strategies and maps these macro-hallmarks to the linguistic violence concepts (discussed in the following sections) with which they are most frequently associated. Here, Verbal Aggression and Abusive Language are used interchangeably.

2.2 Dangerous Speech

Coined by Susan Benesch, Dangerous Speech is defined as any form of expression (speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group (Leader Maynard and Benesch, 2016). It is described as a "narrower and more precisely bounded category" than hate speech, the most prevalent term in academic discourse (Benesch, 2023).

Dangerousness is not a binary concept, but instead falls on a spectrum (Benesch, 2023). Speech (including words, sounds, and images) can be "more or less dangerous" depending on factors such as the influence of the speaker, the audience, the message itself, and the specific social and historical context, as Benesch (2023) states. This model emphasizes the cumulative effect of rhetoric, where repeated messages act as "drops of petrol," slowly shifting the mindset of an audience to become more susceptible to the next message (Benesch, 2023). Furthermore, according to the author, repeated exposure can convince members of an audience that such ideas are "widely accepted," even if they do not personally believe them.

2.3 Toxic Speech

Tirrell (2017) states toxic speech as a mechanism by which speech acts and discursive practices can inflict harm, making sense of claims about harms arising from speech devoid of slurs, epithets, or a narrower class that the author calls 'deeply derogatory terms.' Building on this definition, we characterize toxic speech as a broader 'community-harm' concept. It degrades the conversational environment, which can occur even when the speech is not directed at a specific individual.

Toxic speech is characterized by hostility, insult, or disrespect that may not necessarily target protected groups but still contributes to a harmful communicative environment. Typical linguistic hallmarks include pejorative intensifiers, sarcasm, or moral devaluation without explicit hate content (McGowan, 2019; Hurt, 2021; Barnes, 2025). This distinction is essential for differentiating between toxic and hateful language in computational models. While toxic language may include aggressive expressions, it does not necessarily imply discriminatory intent, as in hate speech.

2.4 Hate Speech (a.k.a., Intolerant Speech)

Although there is no consensus on its formal definition, Fortuna and Nunes (2018) present hate speech as "a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used". This concept is strongly linked to the rhetorical strategies discussed in §2.1, particularly the Figures of Opposition (which includes Passionate Hate and Aversion) and Sanction Speech, as these are often used to construct an "other" and punish non-conformity.

Hate speech is not confined to a single identity or social marker. It encompasses language that insults, intimidates, or harasses people who share characteristics commonly associated with historically marginalized groups—such as race, gender, sexuality, religion, nationality, or socioeconomic status. As Baider (2020) observe, new social

Table 1: Hallmarks Organization.

Macro Hallmarks	Specific Hallmarks	Dangerous Speech	Toxic Speech	Hate Speech	Verbal Aggression	Offensive Language
Dehumanization	Dehumanization	■		■		
Accusation in a Mirror	Accusation in a Mirror	■				
Break of Social Contracts of Integrity or Purity	Threat to Integrity or Purity	■	■			
	Assertion of Attack Against Women and Girls	■				
	Sanction			■	■	■
Questioning In-Group Loyalty	Questioning In-Group Loyalty	■				
Figures of Opposition	Themes and Figures of Opposition			■		
	Passionate Hate and Aversion to the Different ones			■		

groups may also become targets of hatred over time and across contexts. Furthermore, hate speech can intersect multiple identities, reflecting compound discrimination, as noted by Santana and Freitas (2024). Restricting definitions solely to legally protected categories risks overlooking vulnerable populations (Baider, 2020), ultimately undermining victims' fundamental rights to equality.

2.5 Verbal Aggression and Offensive Language

Verbal aggression refers to language used with the intention of insulting, demeaning, intimidating, or otherwise causing harm to an individual or group (Balci and Salah, 2015; Mane et al., 2025). Unlike this, *offensive language* is characterized primarily by its form — such as profanity or taboo expressions — rather than by an explicit intention to harm or a clearly defined target (Liu et al., 2022). While both phenomena may overlap, verbal aggression implies purposeful hostility, whereas offensive language may violate social or cultural norms without necessarily conveying aggression.

This contrast in targeting is significant. While *Verbal Aggression* is typically aimed at a specific individual, ElSherief et al. (2018) point out that *Hate Speech* can take two main forms: *directed*—a personal and explicit attack on an individual based on their identity—or *generalized*, which targets an entire group and often employs extreme vocabulary such as "exterminate" or "kill".

2.6 Cyberbullying

Cyberbullying is conceptually distinct from the other forms of linguistic violence defined in this ontology. While concepts like Hate Speech or Verbal Aggression are defined by their semantic *content* (e.g., attacking a protected group or insulting an individual), Cyberbullying is defined by its *behavioral pattern*.

Smith et al. (2008) formally defines cyberbullying as "an aggressive, intentional act carried out by a group or individual using electronic forms of contact, repeatedly or over time against a victim that cannot easily defend him or herself". Thus, based on this and other definitions (Yao et al., 2019; Salawu et al., 2020), one can note that it is typically defined by three primary attributes: (1) Repetition, as the harmful acts occur repeatedly over time; (2) Targeting, as it is directed at a specific individual (the victim); and (3) Power Imbalance, where the aggressor has some form of social or digital advantage over the victim. Rather than being a distinct *type* of language, Cyberbullying is a pattern of behavior that *uses* other forms of linguistic violence as its tool, such as Verbal Aggression (e.g., repeated insults) or Hate Speech (e.g., repeatedly harassing someone based on their identity).

3 An Ontology for Linguistic Violence

The development of this ontology followed formal design principles, primarily focusing on reuse and consistency (Arp et al., 2015). To guide the modeling process and ensure semantic clarity, we defined a set of Competency Questions (CQs) that the ontology must be able to answer: **CQ1**: Is the linguistic act directed at a specific individual or a protected identity group? **CQ2**: Does the discourse employ specific rhetorical strategies (hallmarks) like dehumanization or mirror accusations? **CQ3**: Does the speech act exhibit a behavioral pattern of repetition or power imbalance? These questions ensure that the classes and object properties defined in our extension remain consistent with the multifaceted nature of linguistic violence.

It is possible to notice a remarkable similarity between some hallmarks. Considering such similarities, we propose that the hallmarks can be organized into an ontology to support automated processing by the application, as well as the definition of rules and inference. For its purpose, we began by defining the major concepts presented before – *Dangerous Speech*, *Toxic Speech*, *Hate Speech*, *Verbal Aggression*, *Offensive Language* and *Cyber-*

bullying – with respect to: its core attributes, its typical target, and the major hallmarks previously specified. Table 2 synthesizes these distinct concepts.

We then searched for existing ontologies specifically related to violent and abusive language, with the aim of reusing them when possible. Among the relevant ontologies are the O-Dang ontology (Stranisci et al., 2022) and the Offensive Language Ontology (OL) (Lewandowska-Tomaszczyk et al., 2021).

The OL was identified as the most suitable for reuse, as several of the concepts we required were already represented in it. In Lewandowska-Tomaszczyk et al. (2021), the authors present an ontological categorization designed for the automated detection of offensive language, developed on the basis of more than 60 available corpus datasets. The vocabulary introduced there has also been employed in subsequent studies by the same authors (Lewandowska-Tomaszczyk et al., 2023b,a).

Extending the OL with the concepts identified in our study, the classes shown in Figure 1 were defined. In this extension, the concepts 'Dangerous Speech', 'Verbal Agression' and 'Cyberbullying' were defined as subclasses of the existing OL class 'Offensive'. The concept 'Toxic Speech' was defined as a subclass of the pre-existing class 'Hate speech'. The concepts 'Offensive Language' and 'Hate Speech' were already represented in the OL.

In addition to these concepts, classes responsible for categorizing the hallmarks were added, with the specific hallmarks defined as individuals. Concepts related to types of target groups (if group or individual) were already defined in the OL, and object properties were created to specify whether the target is an in-group or an out-group.

Regarding the hierarchical modeling in the extended OL ontology (Figure 1), the placement of 'Toxic Speech' as a subclass of 'Hate Speech' is a design choice driven by the reuse constraints of the original OL ontology. While conceptually we treat Toxic Speech as a broader 'community-harm' phenomenon that may lack identity-based intent, the OL framework's structure required this nesting to inherit foundational offensive properties. To maintain conceptual clarity, our framework differentiates them at the instance level through specific attributes: Hate Speech requires a 'Protected Group' target, whereas Toxic Speech is characterized by conversation degradation and 'Sanction Speech' hallmarks.

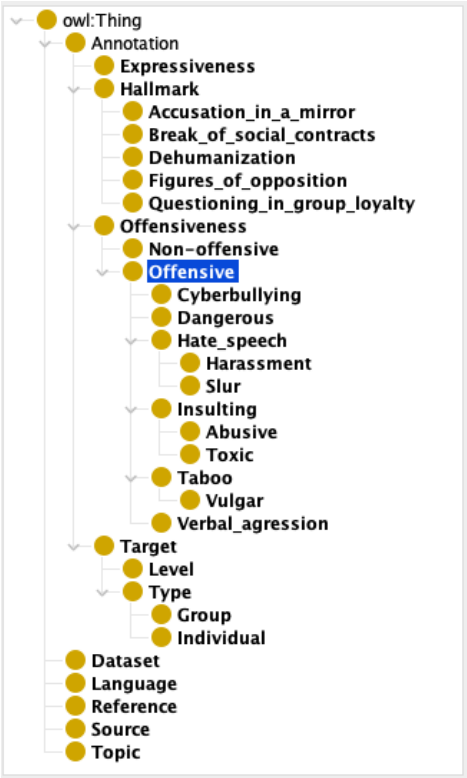

Figure 1: Extended OL Ontology.

The framework uses key differentiation axes, allowing for a granular classification that resolves the ambiguity of overlapping terms. The primary axes are the Central Attribute, which defines the core mechanism of the harm (e.g., whether it is an identity-based attack, an incitement to violence, or a violation of social norms), and the Typical Target, which separates generalized speech against groups from directed speech against individuals.

This multi-axis model provides a clear decision-making path. For example, while Hate Speech and Verbal Aggression may both be offensive, our ontology distinguishes them based on the target: Verbal Aggression is directed at an Individual, whereas Hate Speech is directed at a Protected Group or an individual as a representative of that group. Furthermore, the Associated Macro Hallmarks (derived from Table 1) connect these concepts to their rhetorical strategies. Dangerous Speech, for instance, is strongly associated with incitement hallmarks like Accusation in a Mirror, while Hate Speech is linked to identity-based hallmarks like Figures of Opposition. This structure forms the basis for the computational methodology (§4).

Table 2: Major concepts of Linguistic Violence: Differentiating Concepts by Attribute, Target, and Hallmark.

Concept	**Central Attribute** (The defining feature)	**Typical Target** (The 'ARG1')	**Associated Macro Hallmarks**
Dangerous Speech	Speech that increases the risk of its audience condoning or participating in real-world, large-scale violence.	**Out-Group** (The target of the violence) and **In-Group** (The audience being incited).	• Accusation in a Mirror • Dehumanization • Break of Social Contracts (Threat to Integrity)
Toxic Speech	Broader, "community-harm" speech. Discursive acts that "poison" a conversation, violating norms and inflicting harm, even implicitly.	**Individual** or **General** (the conversation itself).	• (Often lower-level incivility, but can include) Break of Social Contracts (Sanction Speech)
Hate Speech	Attack/incitement against a group based on protected identity characteristics (e.g., race, gender, religion).	**Group (Protected)** or Individual (as a representative of a group).	• Figures of Opposition • Dehumanization • Break of Social Contracts (e.g., Attack on Women/Purity)
Verbal Aggression (or Abusive Language)	A direct, personal attack intended to insult, demean, intimidate, or harm an individual. Corresponds to "insulting words directly addressed".	**Individual** (e.g., "you", "@username").	• Break of Social Contracts (specifically "Sanction Speech") • Questioning In-Group Loyalty (if the individual is seen as a "traitor")
Offensive Language	Speech defined purely by *form*. Use of profanity, vulgarity, or taboo words, often without a specific target or clear intent to harm.	**None / General**.	• None. (This is a key distinction; it's about word choice, not the rhetorical functions listed in Table 1).
Cyberbullying*	A *behavioral pattern* of repeated, directed harassment against a specific individual, often involving a power imbalance.	**Individual** (specific and repeated).	• *Not defined by a specific hallmark, but is a pattern that **uses** Verbal Aggression or Hate Speech as its tool.*

**Cyberbullying is included as a distinct behavioral pattern, rather than a specific speech act, to resolve conceptual ambiguity.*

4 Methodological Approach

Authors of specific sensitive content may intentionally avoid using sensitive phrases to exploit the lack of human reviews with long experience combating hate-detection technologies (He et al., 2020). Recognizing implicitly damaging texts is difficult due to their highly context-sensitive and metaphorical arousal content. In the literature, different strategies have been proposed to identify potential texts that contain hate speech. Some focus mainly on linguistics-based approaches, often analyzing the vocabulary used (lexical analysis). Automated approaches also consider semantic features by applying frames to analyze text content in context. Other approaches rely mainly on machine-learning-based methods, i.e., training a computational model to identify intrinsic characteristics of data previously labeled as hate and non-hate speech to generalize to other domains. However, many of these methods struggle to handle implicit harmful texts, which remain a key challenge for text classification and semantic comprehension. Unfortunately, none of the traditional text representation models employ contextual representation as direct input when encoding the current word, which is required to capture implicit meaning (He et al., 2020).

The proposed methodology draws inspiration from frame semantics and stance detection to model the pragmatics of incitement. Specifically, by isolating the predicate-argument structure through SRL, we operationalize a form of argument mining focused on identifying the speaker's rhetorical positioning. Hallmarks like 'Figures of Opposition' and 'Accusation in a Mirror' are treated as structured indicators of the speaker's stance toward protected groups, allowing the pipeline to differentiate between genuine violence and neutral discourse through semantic framing rather than surface-level keywords.

As previously noted, detecting nuanced linguistic violence is a significant challenge, as meaning is often implicit, contextual, and subjective. Traditional models that rely on lexical analysis or "bag-of-words" approaches struggle to differentiate between, for instance, a general insult (Toxic Speech) and a targeted, identity-based attack (Hate Speech). Although the dangerousness of speech depends significantly on context, which often "cannot be detected and evaluated automatically," it may still be possible to build classifiers that operate by "detecting similarities and patterns" within the harmful speech itself (Benesch, 2023). Our approach, therefore, focuses on identifying these semantic and structural patterns.

To operationalize the ontology proposed (§3), we propose a conceptual NLP pipeline designed to deconstruct a text into its core semantic components. This methodology moves beyond simple

classification by first understanding who is being targeted and what action is being performed, before mapping these components to our framework.

This process, visualized in Figure 2, involves three primary stages: (1) Toxicity detection; (2) Semantic Role & Entity Deconstruction; and (3) the Ontological mapping.

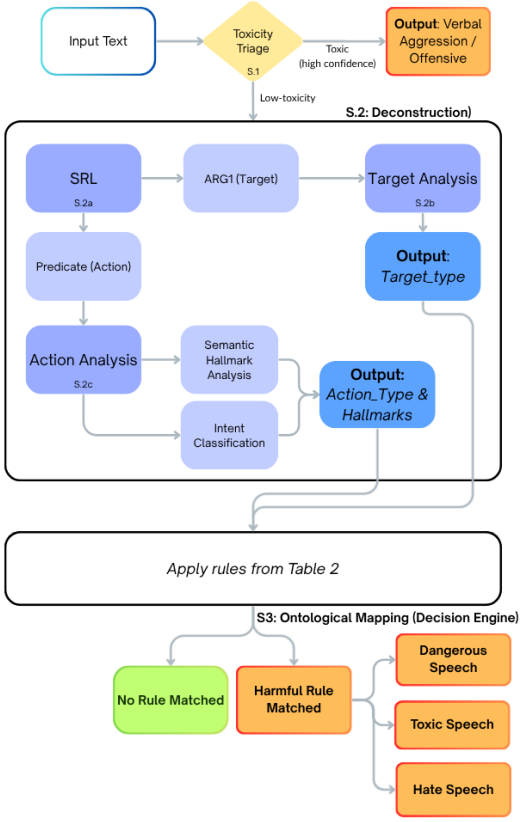

Figure 2: Conceptual NLP Pipeline for Differentiating Linguistic Violence.

4.1 Stage 1: Initial Triage (Toxicity Detection)

The pipeline begins with a high-recall binary toxicity classifier. However, its role is not to filter content, but to route it efficiently.

- If Obviously Toxic (e.g., high confidence score, presence of slurs): The text is likely Verbal Aggression or Offensive Language. It can be classified immediately (or passed to a lightweight version of Stage 3) for efficient processing.
- If Benign or Low-Toxicity (low confidence score): The text exits the simple classification path and is passed to Stage 2 for deep analysis.

For implementation, this stage can be performed by fine-tuning a pre-trained multilingual transformer, such as BERT-base-multilingual-cased (mBERT) (Song et al., 2021; Joshi et al., 2025) or XLM-RoBERTa (Li et al., 2022), on large-scale public toxicity datasets such as OLID-BR (Trajano et al., 2024) or ToxSyn-PT (Brito et al., 2025). Complementary interdisciplinary tools like the Perspective API (Lees et al., 2022) may also be integrated to provide additional toxicity and threat-level signals, enriching the linguistic analysis with socially oriented metrics. Texts receiving high-confidence toxicity scores (e.g., explicit slurs) can be routed to a fast classification stage (e.g., *Offensive Language* or *Verbal Aggression*), while benign or low-toxicity texts are passed to Stage 2 for deeper semantic analysis. This is a critical step, as implicit hate, sarcasm, and complex rhetorical violence (such as *Dangerous Speech*) often masquerade as benign content.

The multi-stage architecture of this pipeline is specifically designed to enable scalability in real-time social media monitoring. By utilizing Stage 1 as a high-recall triage filter, the system can immediately classify content with high toxicity confidence—such as explicit insults or profanity—without the computational overhead of deep semantic analysis. This strategy significantly mitigates the computational bottleneck by reserving the more resource-intensive SRL and NER tasks in Stage 2 for ambiguous, implicit, or low-toxicity cases where granular analysis is indispensable for accurate differentiation.

4.2 Stage 2: Semantic Role & Entity Deconstruction

In addition to lexical and syntactic features, the analysis of semantic roles provides a deeper understanding of who performs and who receives the action expressed in a sentence. Semantic Role Labeling identifies this predicate–argument structure, assigning roles such as `ARG0` (typically the agent or source of the action) and `ARG1` (the patient or target affected by it). For example, in the sentence "Maria insulted João", Maria would be labeled as `ARG0` experiencing the verbal aggression. This distinction is especially relevant for modeling violent or toxic language, since `ARG1` often denotes the target of harmful discourse, making it a key element in distinguishing types of linguistic violence.

This second stage is the core of our methodology, where the speech act is decomposed into complementary NLP tasks, each addressing a distinct analytical dimension. First, the text is processed by

an **SRL** model to identify its predicate–argument structure, isolating the *Predicate* (the action, e.g., "should be attacked") and the ARG1 (the target, e.g., "you" or "those women"). Pre-trained SRL models from libraries like AllenNLP or the Hugging Face Hub—typically based on multilingual BERT—can be adapted for this task. The ARG1 is then passed to a custom **Target Analysis (NER)** model, which classifies the type of target—a key attribute distinguishing concepts in our ontology. Example labels include INDIVIDUAL_REFERENCE (e.g., "you", "@user"), PROTECTED_GROUP (e.g., "immigrants", "women", "Muslims"), and POLITICAL_FIGURE. This model can be fine-tuned from multilingual encoders such as XLM-RoBERTa or mDeBERTa using frameworks like spaCy or transformers.

Finally, the *Predicate* from SRL is analyzed through **Action & Hallmark Analysis** to characterize the intent and nature of the harm. This involves two sub-tasks: (1) an **Intent Classification** model (e.g., fine-tuned mBERT (Jayanth et al., 2024)) identifies the communicative intent, such as *INSULT*, *THREAT*, *VIOLENT_INCITEMENT*, or *CRITICISM*; and (2) **Semantic Hallmark Detection** uses vector semantics to compare the predicate and its complements (e.g., "are rats") against a curated lexicon to detect *Macro Hallmarks* (Table 1). For instance, high similarity between the object (rats") and a *dehumanization lexicon* (e.g., vermin", animals", "insects") flags the *DEHUMANIZATION* hallmark. This task can be implemented using the `sentence-transformers` library and models like *paraphrase-multilingual-MiniLM-L12-v2*.

4.3 Stage 3: Ontological Mapping (Decision Engine)

Finally, the structured outputs from Stage 2 (Target_Type, Action_Type, Hallmark_Detected) are fed into a decision engine. This engine uses the rules defined in our ontology (Table 2) to render a final, granular classification.

This engine also handles cases that are found to be truly benign after deep analysis. If no harmful rules from the ontology are matched (e.g., a text mentions a protected group in a neutral context), the text is confirmed as *Benign*. This ensures that the pipeline correctly distinguishes between harmful rhetoric and non-problematic speech that simple models might misclassify.

5 Application, Discussion, and Limitations

To demonstrate the robustness of our proposed ontology and pipeline, we present a case study analyzing real-world examples. These instances are often misclassified by traditional, 'end-to-end' classifiers that rely on lexical cues rather than semantic deconstruction.

Table 3 demonstrates the framework's capability to deconstruct complex cases. The second example, while easily flagged as Toxic by a baseline model, is categorized by our pipeline. Target_Type is identified as INDIVIDUAL_REFERENCE and the Action_Type as INSULT, leading to the correct classification of Verbal Aggression rather than (identity-based) Hate Speech. Thus, the pipeline identifies cases simple models would miss, such as the example glorifying General Pinochet. Another classifier may see Benign text, but the *Semantic Hallmark Analysis* detects DEHUMANIZATION ("remnant of disgrace") and GLORIFICATION_OF_VIOLENCE. This triggers the ontological rules, correctly mapping the text to both Hate Speech and Dangerous Speech.

The pipeline demonstrates its ability to avoid false positives common to simpler models. In the example "Women's achievements in sports...", a baseline keyword-based model would likely flag the text as problematic simply due to the presence of the term "women." Our pipeline, however, successfully avoids this error. While Stage 2 correctly identifies "women" as a PROTECTED_GROUP, the *Action & Hallmark Analysis* finds non-harmful predicates, specifically CRITICISM ("are underestimated") and PRAISE ("shown exceptional skills"). Crucially, no harmful hallmarks are detected. Consequently, the *Ontological Mapping* (Stage 3) does not match any harmful rules from our ontology (Table 2) and correctly confirms the text as Benign. This ability to deconstruct context and analyze components — rather than just keywords — distinguishes between genuine linguistic violence and neutral discussion about sensitive topics.

A challenge inherent to applying SRL to social media is the informal nature of online discourse, which frequently includes slangs, abbreviations, and non-standard syntax. However, the use of multilingual encoders based on subword tokenization, such as mBERT or XLM-RoBERTa, provides robustness to these variations, allowing the model to capture semantic intent even when faced with

Table 3: Case Study Analysis of Challenging Examples Using the Proposed Framework.

Example Text	**Source**	**Initial Triage (S.1)**	**Semantic Role & Entity Deconstruction (S.2)**	**Ontological Mapping (S.3)**
"[Gays are] what the cows leave behind... maggots [and] flies" (*[Gays são] o que as vacas deixam para trás... larvas [e] moscas*)	2024 Dangerous Speech Election Dataset (Dickinson et al., 2025)	'Toxic' / 'Hate' (High-confidence explicit hate)	• **Target_Type:** 'PROTECTED_GROUP' (Gays) • **Action_Type:** 'DELEGITIMIZATION' (Comparison) • **Hallmark:** 'DEHUMANIZATION' (explicit)	'Hate Speech' & 'Dangerous Speech'
"@user @user @user Hahahahaha you don't understand f***ing anything!!!! That's why your credit is nill! Another left-wing door [insult]!!!!" (*@user @user @user Hahahahaha vc não entende porra nenhuma mesmo!!!! Por isso seu crédito é nulo! Outra porta de esquerda!!!!*)	ToLD-Br (Leite et al., 2020)	'Toxic' (High profanity and direct insult)	• **Target_Type:** 'INDIVIDUAL_REFERENCE' ("vc" / @user) • **Action_Type:** 'INSULT' ("*não entende porra nenhuma*", "*porta de esquerda*") • **Hallmark:** 'Sanction Speech' (Political)	'Verbal Aggression' / 'Toxic Speech'
"I pity you for not having been from the era of the great General Pinochet, who unfortunately let this remnant of disgrace survive." (*Eu tenho pena é de você não ter sido da época do grande General Pinochet que lamentavelmente deixou sobrar este resquício de desgraça.*)	HateBR2.0 (Vargas et al., 2025)	'Benign' or 'Toxic' (Simple models would miss the context)	• **Target_Type:** 'INDIVIDUAL_REFERENCE' ("você") • **Action_Type:** 'INSULT' & 'GLORIFICATION_OF_VIOLENCE' • **Hallmark:** 'DEHUMANIZATION' ("remnant of disgrace"), 'Figures of Opposition' (Pinochet vs. target)	'Hate Speech' (Political) & 'Dangerous Speech'
"Women's achievements in sports are often underestimated, but they have shown exceptional skills and tireless dedication." (*As conquistas das mulheres em esportes muitas vezes são subestimadas, mas elas têm demonstrado habilidades excepcionais e dedicação incansável.*)	ToxSyn-PT (Brito et al., 2025)	'Benign' (Might be flagged by keyword filters)	• **Target_Type:** 'PROTECTED_GROUP' (Women) • **Action_Type:** 'CRITICISM' ("are underestimated") & 'PRAISE' ("shown exceptional skills") • **Hallmark:** 'None'	'Benign' (No harmful rule matched)

out-of-vocabulary terms or orthographic shifts. In cases where SRL produces incomplete frames due to extreme brevity or syntactic fragmentation, the decision engine in Stage 3 serves as a critical safety mechanism: if no structural evidence is sufficient to trigger a specific violence rule from the ontology, the system confirms the text as 'Benign'. This component-based approach prioritizes precision and minimizes the false positives often generated by linguistic noise in traditional end-to-end models.

6 Final Remarks

As stated by Benesch (2023), inciting language exerts a cumulative influence on audiences, shaping perceptions and attitudes over time. Understanding the impact of speech in action is to notice the dependency on contextual elements, such as the speaker, audience, content, and surrounding circumstances. People often use terms that deviate from conventionally accepted definitions to express complex and implied meanings. The speech record differs from the written record and even more from the self-record that takes place online. There is communication in the voice that is not explicit in the text: intonation, way of speaking, emotional tone, a small amount of sarcasm, and irony, which, when represented textually, often escapes existing natural language processing techniques. Regarding hate speech, there is still subjectivity in the sense that there are diverse valid beliefs about what the correct data labels should be (Röttger et al., 2021).

This work presented a conceptual framework and ontology to differentiate between violent, toxic, and hateful language. By structuring these phenomena semantically and organizing them into distinct categories, the ontology aims to support NLP systems in capturing these nuances of linguistic aggression, extending beyond simple lexical cues. Future work includes validating this ontology through corpus annotation and integrating it into explainable NLP models for automated classification. This component-based approach is auspicious for expanding to multilingual contexts, including low-resource languages like Portuguese. Rather than depending on massive, task-specific labeled datasets, our methodology leverages core semantic tasks like SRL and NER, for which foundational, multilingual pre-trained models are increasingly available. This is essential for contexts where textual features and pragmatic dimensions differ significantly from language to language.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/Brazil) – Finance Code 001.

References

Robert Arp, Barry Smith, and Andrew D Spear. 2015. *Building ontologies with basic formal ontology*. Mit Press.

Fabienne Baider. 2020. Pragmatics lost?: Overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society*, 11(2):196–218.

Koray Balci and Albert Ali Salah. 2015. Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. *Comput. Hum. Behav.*, 53(C):517–526.

Michael Randall Barnes. 2025. Complex harms in online speech. *Conversations Online: Explorations in Philosophy of Language*, page 239.

Susan Benesch. 2023. Dangerous speech. In Christian Strippel, Sünje Paasch-Colberg, Martin Emmer, and Joachim Trebbe, editors, *Challenges and perspectives of hate speech research*, volume 12 of *Digital Communication Research*, pages 185–197. Berlin.

Iago Alves Brito, Julia Soares Dollis, Fernanda Bufon Färber, Diogo Fernandes Costa Silva, and Arlindo Rodrigues Galvão Filho. 2025. Toxsyn-pt: A large-scale synthetic dataset for hate speech detection in portuguese. *Preprint*, arXiv:2506.10245.

Judith Butler. 2021. *Discurso de ódio: Uma política do performativo*. Editora Unesp, São Paulo.

Diana Luz Pessoa de Barros. 2014. O discurso intolerante na internet: enunciação e interação.

Robert Allan Dickinson, Cathy Buerger, Tom Cowin, Niko Shahbazian, Alexandra Filindra, Stacey Hunt, and Elizabeth Young. 2025. 2024 Dangerous Speech Election Dataset.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *Preprint*, arXiv:1804.04257.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Guoxiu He, Zhe Gao, Zhuoren Jiang, Yangyang Kang, Changlong Sun, Xiaozhong Liu, and Wei Lu. 2020. Think beyond the word: Understanding the implied textual meaning by digesting context, local, and noise. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2297–2306, New York, NY, USA. Association for Computing Machinery.

Marlon Hurt. 2021. *Pledging to Harm: A linguistic analysis of violent intent in threatening language*. Ph.D. thesis, Aston University.

Md Saroar Jahan and Mourad Oussalah. 2021. A systematic review of hate speech automatic detection using natural language processing. *arXiv preprint*.

K Krishna Jayanth, G Bharathi Mohan, R Prasanna Kumar, and M Rithani. 2024. Intent recognition leveraging xlm-roberta for effective nlu. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 877–882.

Sakshi Joshi, Anindita Mukherjee, Usha A. Jogalekar, Renuka Agrawal, Abhishek Anand, and Santhosh Phanitalpak Gandhala. 2025. Enhancing toxic comment classification with multi-label capabilities: Leveraging bert and roberta models. In *2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T)*, pages 1–6.

Jonathan Leader Maynard and Susan Benesch. 2016. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. *arXiv preprint*.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Barbara Lewandowska-Tomaszczyk, Anna Baczkowska, Olga Dontcheva-Navratilova, Chaya Liebeskind, Giedre Valunaite Oleskeviciene, Slavko Zitnik, Marcin Trojszczak, Renate Povolna, and Linas Selmistraitis. 2023a. Llod schema for simplified offensive language taxonomy in multilingual detection and applications. *Lodz Papers in Pragmatics*, 19:301–324.

Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene, and Slavko Žitnik. 2023b. An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics*, 19(1):7–48.

Barbara Lewandowska-Tomaszczyk, Slavko Zitnik, Anna Baczkowska, Chaya Liebeskind, Jelena Mitrović, and Giedre Valunaite Oleskeviciene. 2021. Lod-connected offensive language ontology and tagset enrichment. In *SALLD-1: Workshop on Sentiment Analysis & Linguistic Linked Data*, pages 135–150. CEUR Workshop Proceedings.

Wenji Li, Anggeng Li, Tianqi Tang, Yue Wang, and Zejian Fang. 2022. Multilingual toxic text classification model based on deep learning. In *2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 726–729.

Junjie Liu, Yong Yang, Xiaochao Fan, Ge Ren, Liang Yang, and Qian Ning. 2022. Offensive-language detection on multi-semantic fusion based on data augmentation. *Applied System Innovation*, 5(1).

Swapnil Sanjaykumar Mane, Suman Kundu, and Rajesh Sharma. 2025. A survey on online aggression: Content detection and behavioral analysis on social media. *ACM Comput. Surv.*, 57(7).

Mary Kathryn McGowan. 2019. *Just words: on speech and hidden harm*. Oxford University Press.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective NLP tasks. *CoRR*, abs/2112.07475.

Semiu Salawu, Yulan He, and Joanna Lumsden. 2020. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1):3–24.

Brenda S. Santana and Larissa A. de Freitas. 2024. Pln em redes sociais. In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edition, book chapter 34. BPLN.

Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

Guizhe Song, Degen Huang, and Zhifeng Xiao. 2021. A study of multilingual toxic text detection approaches under imbalanced sample distribution. *Information*, 12(5).

Marco A. Stranisci, Simona Frenda, Mirko Lai, Oscar Araque, Alessandra T. Cignarella, Valerio Basile, Viviana Patti, and Cristina Bosco. 2022. O-dang! the ontology of dangerous speech messages. *Preprint*, arXiv:2207.10652.

Lynne Tirrell. 2017. Toxic speech: Toward an epidemiology of discursive harm. *philosophical topics*, 45(2):139–162.

Douglas Trajano, Rafael H. Bordini, and Renata Vieira. 2024. Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.

Francielle Vargas, Isabelle Carvalho, Thiago A. S. Pardo, and Fabrício Benevenuto. 2025. Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31(2):435–456.

Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*, pages 3427–3433.