

Supporting the Curation of Educational Texts: Using Computational Methods for Violent Content Screening

Lara da Silva Dias^{1,2,3}, Lucas Larcher^{1,3}, João Augusto Pilato de Castro^{1,2,3},
João Vítor de Castro Martins F. Nogueira^{1,2,3}, Jairo Francisco de Souza^{1,2,3}

¹ LApIC Research Group – Juiz de Fora – MG – Brazil

²Federal University of Juiz de Fora (UFJF) – Juiz de Fora – MG – Brazil

³Center for Public Policy and Education Evaluation (CAEd)
Juiz de Fora – MG – Brazil

{lara.dias, lucas.larcher, joao.castro, joao.nogueira}@fundacaocaed.org.br,
jairo.souza@ufjf.br

Abstract. *Curating texts for educational use is costly but essential for producing learning materials, item tests, and other resources. In addition to aligning with a skill intended to be taught or assessed, the material must be non-harmful to students, especially free of violent content, inappropriate language, or any other form of hate speech. This study analyzes computational methods for automatically detecting and grading violence in Portuguese children's texts. We introduce the Gru, a corpus annotated by experts based on a taxonomy derived from the Brazilian Indicative Rating System. Results reveal that traditional models miss implicit violence, whereas LLMs excel in contextual reasoning.*

1. Introduction

Large-Scale Assessment (LSA) plays a fundamental role in diagnosing and improving educational systems [Clarke and Luna-Bazaldua 2021]. The curation of supporting materials for test items (texts, images, and other stimuli) constitutes a central stage in the item development process, as these materials directly influence the quality, appropriateness, and interpretability of the resulting items [Brennan 2023]. In addition to aligning with the assessed skills, such materials must comply with strict ethical and sensitivity criteria to ensure that students are not exposed to content inappropriate for their age group.

The identification and grading of violent content in candidate texts are critical requirements. The Indicative Rating System [da Justiça e Segurança Pública 2018] in Brazil establishes official guidelines for content classification, defining distinct levels of violence and guiding the suitability of materials for different stages of cognitive and emotional development. Although originally designed for audiovisual media, it provides a structured, standardized set of criteria for assessing the intensity of violence. In this study, these guidelines are adapted to the textual domain, enabling the operationalization of a consistent scale for evaluating violence in educational materials.

Despite its pedagogical importance, the curation and validation of supporting materials require specialized expertise and account for a significant portion of the expert's effort in item development. This reliance on manual review creates a substantial operational bottleneck, as experts often spend a disproportionate amount of time screening

and discarding inappropriate candidate texts that fail to meet sensitivity criteria. Consequently, the speed of item production frequently fails to keep pace with the continuous, high-volume demand for new assessments, thus limiting the scalability of LSA [Haladyna and Rodriguez 2013, Heru et al. 2022]. This challenge is further intensified by the inherent subjectivity of language, as identifying sensitive content requires interpreting contextual nuances, figurative language, and implicit markers of violence, which are often difficult for automated approaches to capture [Hee et al. 2024].

Historically, classical supervised classification methods—such as Support Vector Machines (SVM) and tree-based ensemble models—have shown limited performance in domains that require fine-grained semantic understanding, even when combined with dense vector representations (e.g., Word2Vec) or sentiment analysis techniques [Liu 2012]. Contextual representations derived from Deep Learning architectures, such as BERT, mitigate some of these limitations but still rely primarily on distributional patterns learned during pre-training and do not inherently incorporate complex normative criteria, such as human-defined scales for grading violence [Devlin et al. 2019].

In contrast, recent advances in Natural Language Processing (NLP), driven by the development of Large Language Models (LLMs), offer a promising alternative [Hee et al. 2024, Palla et al. 2025]. These models demonstrate a strong ability to retain long-range context, capture communicative intent, and interpret figurative language—capabilities that are essential for detecting violence in textual data. However, their performance critically depends on prompt design strategies (prompt engineering) and alignment with formal normative frameworks to mitigate algorithmic biases [Palla et al. 2025, Furniturewala et al. 2024].

Given this context, this work investigates the use of LLMs to classify violence in educational texts, evaluating their ability not only to detect its presence but also to assess its severity. The study is guided by the following research questions: **(RQ1)** Are LLMs capable of identifying and grading the intensity of violence in candidate text passages (i.e., stimulus text) for literary and informational texts more effectively than traditional machine learning methods when aligned with the Brazilian Indicative Rating System guidelines?; **(RQ2)** What is the impact of analysis granularity—full-document evaluation *versus* sentence-level evaluation - on the predictive performance of classical and generative classifiers?; **(RQ3)** Is the application of a binary safety threshold, derived from multiclass predictions, a computationally viable strategy to optimize the trade-off between retaining safe texts and reducing false positives?

To address these questions, we propose a classification pipeline centered on LLMs and evaluate it on a violence-intensity scale ranging from 0 to 3. As a comparative baseline, we employ supervised classification approaches—including Logistic Regression, Random Forest, XGBoost, and SVM—applied to contextual embeddings (BERTimbau) and semantic representations (Word2Vec), combined with sentiment-based features.

The main contributions of this work are: **(i)** the construction and release of a gold-standard corpus composed of long-form texts from the educational domain, manually annotated by experts according to multiclass violence criteria; **(ii)** a systematic empirical evaluation comparing state-of-the-art generative AI approaches (LLMs) with classical architectures in the challenging task of detecting both implicit and explicit violence; **(iii)** the

validation of a hybrid curation methodology for educational assessment systems, demonstrating that converting graded classifications into binary blocking thresholds improves operational scalability. By integrating national regulatory guidelines with state-of-the-art NLP techniques, this study aims to provide practical support for developing automated, interpretable, and auditable content moderation mechanisms in educational resources.

2. Related Work

Textual content sensitivity encompasses a broad spectrum of categories. Within the behavioral and ethical domain, notable aspects include bullying—characterized by the endorsement of behaviors that conflict with educational principles [Olweus 1993]—and parental authoritarianism, marked by excessive rigidity and lack of dialogue. Sensitive social issues include prejudice (discrimination based on ethnicity, gender, religion, or body characteristics) and vulnerability, which expose situations of economic, social, or environmental limitations. Other critical dimensions involve individual safety and integrity. Violence, the focus of this study, is defined as the intentional use of force resulting in physical or psychological harm [Krug et al. 2002], while sexual content refers to the use of figurative or literal language to convey ideas of a sexual nature that are inappropriate for a given age group [da Justiça e Segurança Pública 2018]. Additionally, topics such as politics and religion require careful handling to avoid doctrinal bias or ideological tension, as well as the explicit exposure to health and financial issues, which may negatively affect the emotional well-being of developing readers [Haladyna and Rodriguez 2013, Educational Testing Service 2022].

Automatic text classification for detecting harmful and violent content has been widely studied over the past decades, largely driven by the need for content moderation on social media platforms [Schmidt and Wiegand 2017, Fortuna and Nunes 2018]. Historically, the literature shows an evolution in the techniques employed. Early filtering systems, such as Smokey [Spertus 1997] and initial dictionary-based heuristics [Razavi et al. 2010], relied on lexical approaches and static lists of prohibited terms (blocklists). While computationally efficient, these rule-based methods proved fragile in the face of linguistic variation, implicit hostility, and contextual complexity, often resulting in high rates of false positives [Burnap and Williams 2015].

With the advancement of Machine Learning, supervised approaches such as Support Vector Machines (SVM) and Random Forest emerged, often using statistical feature vectors like Term Frequency–Inverse Document Frequency (TF-IDF). Studies such as [de Pelle and Moreira 2017] and [Zampieri et al. 2019] demonstrated the effectiveness of SVM in identifying toxic and offensive comments. More recently, [Battiston 2025] employed a Multinomial Naive Bayes classifier combined with TF-IDF to detect violent news headlines, achieving an accuracy of 86%. [Guide 2022] investigates punitive hate speech on social media and discusses the challenges of identifying such content from linguistic and computational perspectives.

More recently, the state of the art has shifted toward Deep Learning and Transformer-based models (such as BERT and RoBERTa), which enable improved semantic understanding through attention mechanisms and contextual representations. [de Oliveira Trajano 2023] applied a pre-trained BERT model in Portuguese for toxic language detection, highlighting the effectiveness of deep learning approaches over tra-

ditional machine learning models. Additionally, the use of artificial neural networks was validated by [Mantini et al. 2021] for recognizing narratives involving violence against women. In Brazilian Portuguese, machine learning approaches have also been explored for classifying textual typology, demonstrating the feasibility of extracting linguistic features from complex narratives [Barbosa et al. 2022].

A closer examination of the aforementioned approaches reveals two consistent limitations. First, regarding textual corpora, the overwhelming majority of studies focus on short, informal social media texts (e.g., Twitter and Reddit), which lack the syntactic and narrative complexity of educational materials. Second, regarding the classification task, most studies treat violence as a strictly binary problem, and in an educational curation pipeline, a direct binary filter acts as a blunt instrument: it tends to conflate mild narrative tension (such as the necessary conflict of an antagonist in a classic fable) with explicit physical aggression, leading to the unjustified exclusion of rich literary materials.

While binary classification provides a functional baseline for content filtering, adopting a multiclass approach enables the capture of pedagogical nuances. Aligning the classification system with official frameworks, such as the Brazilian Indicative Rating System, makes the automated solution suitable for national educational guidelines, allowing curators to distinguish between narrative tension and explicit aggression. However, this increased granularity inherently escalates the complexity of the classification task.

Distinguishing between subtle levels of violence intensity, particularly in long texts that require resolving long-distance dependencies and handling extended context windows, poses a significant challenge for traditional NLP models. This trade-off highlights a methodological gap, calling for solutions that can navigate the challenges of severity grading while remaining aligned with normative educational standards. Recent explorations into Large Language Models (LLMs) indicate their potential for scaling both the generation of educational items [Nogueira et al. 2025, Chico et al. 2025] and content moderation, provided they are carefully aligned with prompts and constraints to mitigate algorithmic biases and hallucinations [Furniturewala et al. 2024, Palla et al. 2025].

3. Materials and Methods

To validate the effectiveness of the proposed solution, the methodology was structured as follows: dataset construction, definition of the classification system, statistical analysis, and, finally, selection of classification models. Existing Portuguese datasets for sensitive content detection typically focus on social media or news portals, such as ToLD-Br [Leite et al. 2020] (21,000 Twitter posts, crowd-sourced multiclass/multilabel annotation) and OLID-BR [Trajano et al. 2024] (7,943 Twitter/YouTube comments, manual hierarchical annotation), which explore toxicity in environments like Twitter and YouTube, focusing on hate speech and direct offenses. Although HateBR 2.0 [Vargas et al. 2022] offers a balanced, expert-annotated corpus of 7,000 Instagram comments, OffComBR [de Pelle and Moreira 2017] in 1,250 G1 News comments using three human judges. These datasets, while essential for social media moderation, do not adequately capture the narrative complexity, figurative language, and formal structure of literary and informational texts used in educational assessments. This domain distinction justifies the creation of a new dataset tailored to violence detection in educational texts for children – the

Gru dataset¹, which focuses on long-form texts where violence is frequently narrated in a contextualized manner rather than delivered as a direct insult.

To ensure strict control over content quality and avoid algorithmic biases from noisy automated web scraping, the dataset was built through a guided manual collection and curation approach from educational, news, and literary websites. The final corpus comprises 300 long-form texts spanning diverse genres (poems, news articles, short stories, and novel excerpts) and was further segmented into 8,607 sentences. These texts possess an average length of 552 words and a vocabulary of 18,242 unique tokens, reflecting the expected density of stimulus texts for literacy items in large-scale assessments.

To establish the reference classification (gold standard), the curation process was conducted by three experts in large-scale educational assessment in Portuguese. A labeling protocol was defined in which each text was evaluated along two dimensions: (1) binary identification of the presence or absence of violence in texts and (2) definition of the texts and their constituent sentences, which were evaluated strictly on a four-point violence intensity scale (ranging from 0 to 3). The active participation of these experts ensured that the selected texts promote critical reading of contemporary issues without compromising students' emotional well-being in high-stakes assessments. Individual classifications were triangulated, and in cases of disagreement, qualitative discussion rounds were conducted until consensus was reached.

To ensure certain classification accuracy and mitigate hallucinations (false positives) inherent to generative models, the system was parameterized using formal definitions. Violence was defined as the intentional use of force or power resulting in injury or deprivation [Krug et al. 2002]. The taxonomy proposed in [da Silva and Ramos 2023] was adopted, which characterizes the phenomenon based on the nature of the affected asset (physical, psychological, or symbolic), the type of offender, and the context of occurrence. Additionally, tracking indicators were incorporated to refine the detection of subtle psychological aggression [dos Santos Rodrigues 2023].

The normative basis for the evaluation metric was derived from the Brazilian *Indicative Rating Practical Guide* and its supplementary technical document *Content Descriptors* [da Justiça e Segurança Pública 2018, da Justiça e Segurança Pública sd]. By adapting these governmental guidelines into a computational taxonomy, we ensured that the evaluation is grounded in validated pedagogical and psychological parameters rather than arbitrary rules. Consequently, a violence intensity scale ranging from 0 to 3 was defined based on these descriptors, as detailed in Table 1.

Individual expert classifications were triangulated, and qualitative discussion rounds were conducted until consensus was reached. Prior to consensus, inter-annotator agreement at the global-document level was assessed using Fleiss' kappa, the agreement was $\kappa = 0.4407$, with observed agreement $\bar{P} = 0.8600$ and expected agreement $P_e = 0.7497$, indicating moderate agreement before consensus. Pairwise Cohen's kappa values ranged from 0.4144 to 0.4683, with a mean of 0.4443. A critical methodological decision was the granularity of the data. Processing long-form educational texts as single inputs often dilutes short bursts of violence within the broader narrative.

At the document level, based on the maximum violence score present in the

¹<https://github.com/lapic-ufjf/gru-dataset>

Table 1. Violence Intensity Scale

Level	Class	Description
0	None	Text without mention of violent acts, threats, or offensive language.
1	Mild	Verbal/Implicit. Hostile language, mild insults, aggressive irony, or descriptions of tension without physical contact.
2	Moderate	Described Physical Action or Clear Threat. Description of physical altercations (e.g., punches, pushing) without graphic details or explicit threats of severe harm or death.
3	Extreme	Torture/Explicit Death. Graphic descriptions of severe injury, torture, explicit homicide, or intense psychological abuse.

global_level, the final labeled *Gru* dataset contains 62 texts classified as true for violence, 208 texts classified as Level 0, 30 as Level 1, 27 as Level 2, and 35 as Level 3. At the granular level, the sentence distribution reflects the natural imbalance of the educational domain, where safe and neutral contexts predominantly outnumber violent events. Consequently, the sentence-level annotation resulted in 8,299 sentences categorized as Level 0, 94 as Level 1, 119 as Level 2, and 95 as Level 3.

To support research reproducibility while respecting copyright and usage restrictions, a sanitized subset of the consensus dataset will be made publicly available. The full prompts and inference configuration are released with the repository under *llm/*, including the global and sentence-level prompts, exact model identifiers, provider endpoints, decoding parameters, and timeout/retry settings. Each record in this released dataset is structured with specific metadata tags mirroring the internal curation process. These include a unique identifier (*id*), the Portuguese textual fragment (*sentence_pt*) along with its English translation (*sentence_en*), the individual evaluations of the three experts (*expert1*, *expert2*, *expert3*) alongside the consensus score (*final*), and the overall maximum violence level of the source document (*global_level*). Supplementary fields for expert remarks (*comment*) and general pedagogical suitability (*adequacy*) are also provided. To ensure full legal compliance, direct source URLs (*links_news*) and proprietary macro-texts will be omitted from the public release. Following the dataset construction and annotation, we defined the computational approaches for the classification task.

4. Experimental Setup

The study was structured into four experiments. The experimental flow was branched to address a fundamental data volume constraint: all training-dependent models were trained and evaluated exclusively on the granular dataset (8,607 sentences). This cascading architecture enabled not only direct comparison across models but also the evaluation of how text granularity (global vs. sentence-level) and output complexity (binary vs. multiclass) impact model performance. For rigorous comparison, all model metrics were computed on the same holdout sample (1,722 sentences, corresponding to 20% of the sentence-level *Gru* dataset), with the remaining 80% utilized for training under a 5-Fold Cross-Validation scheme. Conversely, Generative LLMs, which operate via zero-shot prompting without requiring weight updates, were evaluated on both this sentence-level holdout and the full texts. The experiments were structured as follows:

Experiment 1 (Binary Classification): The primary objective of this experiment was to verify the capability of the classifiers to accurately separate texts and sentences into violent and non-violent categories, establishing a binary detection baseline. *Macro Approach:* Zero-shot prompted LLM operated on full texts, leveraging full narrative context without requiring training. *Granular Approaches:* Classifiers trained using the partitioned sentence-level data. The granular configuration was divided into three progressive approaches: First, a lexicon-augmented baseline used a Logistic Regression classifier, with the input representation built through early feature concatenation: the continuous bag-of-words vector from Word2Vec was concatenated with a four-dimensional sentiment score vector extracted via VADER. VADER was strictly applied to this baseline because static embeddings often struggle to isolate aggressive polarity without explicit lexicon-based injection. Second, a set of intermediate models comprising SVM, Random Forest, XGBoost, and Logistic Regression was trained using contextual embeddings extracted from BERTimbau. Finally, a fully fine-tuned approach was evaluated. BERTimbau was specifically selected because it represents the state-of-the-art transformer for Brazilian Portuguese, inherently capturing deep semantic and polarity nuances, thus rendering manual feature concatenations redundant for these advanced models.

Experiment 2 (Multiclass Classification): The task was extended to classify violence intensity across the four levels (0–3). *Granular Supervised Approaches:* The supervised models were adapted for multiclass prediction using the 8,607 annotated sentences. *Generative LLM Approaches:* LLMs were evaluated in two distinct settings: evaluating the full macro-text globally and iteratively evaluating it sentence-by-sentence. The primary objective of this dual evaluation was to verify the expectation that forcing LLMs to classify at the granular (sentence) level improves their overall reasoning and decision-making accuracy regarding the text. Furthermore, in the context of educational curation, identifying even a single excerpt containing explicit violence (Levels 2 or 3) is sufficient to exclude the macro-text from use in educational items. Therefore, pinpointing the exact location of the sensitive trigger is indispensable.

Experiment 2.1 (Binary Grouping): To align the multiclass predictions with practical real-world curation pipelines, the outputs were converted into a binary format. Levels 0 and 1 were grouped as “Safe” (0), while levels 2 and 3 were labeled as “Violent” (1). This grouping is justified: while Level 0 is entirely neutral, Level 1 encompasses mild narrative tension and implicit conflict (e.g., a villain’s presence in a fable), which are essential literary elements, penalizing Level 1 would result in the exclusion of educational materials. Conversely, Levels 2 and 3 involve explicit physical aggression and severe harm, which are strictly prohibited in assessment contexts.

Experiment 3 (Control Dataset): To evaluate the applicability of the proposed solution, a negative control dataset was compiled, containing 3,055 validated educational items previously administered in official exams². Because these items underwent rigorous pedagogical review and were approved for large-scale assessments, it is expected that none contain violent content. Therefore, this dataset was submitted to all evaluated models to verify their capability to correctly classify safe operational items, providing a direct measure of their specificity (avoidance of false positives) in a real curation pipeline.

²The items were provided by a Brazilian educational assessment company CAEd Foundation.

5. Results and Discussion

The quantitative analysis was based on Accuracy, Precision, Recall, and F1-Score metrics (macro-averaged for multiclass scenarios). In educational assessments, prioritizing Recall is essential, as failing to detect violent content (a false negative) carries severe psychological risks by exposing students to inappropriate material during high-stakes exams. Concurrently, Precision was monitored to minimize false alarms (false positives), ensuring that pedagogically valuable texts containing only mild, acceptable narrative tension were not unjustly discarded from the item bank. Table 2 shows the results.

Table 2. Models' performance at the global and granular levels.

Exp	Model	Acc	Prec	Rec	F1
	GPT-5.2 Global	0.870	0.618	0.825	0.707
	Claude Sonnet 4.5 Global	0.910	0.813	0.684	0.743
	Gemini 3 Flash Global	0.880	0.624	0.930	0.746
	Sabiá 4 Global	0.913	0.919	0.596	0.723
E1	W2V + VADER	0.857 (± 0.0126)	0.140 (± 0.0178)	0.590 (± 0.1071)	0.226 (± 0.0293)
	Logistic Regression	0.925 (± 0.0063)	0.221 (± 0.0562)	0.443 (± 0.1137)	0.265 (± 0.0744)
	Random Forest	0.962 (± 0.0034)	0.368 (± 0.1484)	0.115 (± 0.0534)	0.175 (± 0.0771)
	SVM	0.954 (± 0.0052)	0.345 (± 0.0623)	0.328 (± 0.0798)	0.336 (± 0.0676)
	XGBoost	0.965 (± 0.0010)	0.556 (± 0.2292)	0.082 (± 0.0488)	0.143 (± 0.0746)
	BERTimbau	0.969 (± 0.0037)	0.557 (± 0.0490)	0.557 (± 0.0253)	0.557 (± 0.0172)
E2	W2V + VADER	0.798 (± 0.0120)	0.298 (± 0.0049)	0.458 (± 0.0284)	0.309 (± 0.0085)
	Logistic Regression	0.918 (± 0.0064)	0.341 (± 0.0061)	0.369 (± 0.0241)	0.349 (± 0.0103)
	Random Forest	0.963 (± 0.0003)	0.241 (± 0.0001)	0.250 (± 0.0000)	0.245 (± 0.0000)
	SVM	0.956 (± 0.0586)	0.442 (± 0.0739)	0.388 (± 0.0853)	0.406 (± 0.0220)
	XGBoost	0.963 (± 0.0003)	0.241 (± 0.0000)	0.250 (± 0.0000)	0.245 (± 0.0000)
	BERTimbau	0.947 (± 0.0045)	0.270 (± 0.0370)	0.337 (± 0.0675)	0.283 (± 0.0373)
	GPT-5.2	0.959	0.365	0.367	0.364
	Claude Sonnet 4.5	0.963	0.575	0.351	0.370
	Gemini 3 Flash	0.951	0.451	0.364	0.358
	Sabiá 4	0.962	0.528	0.282	0.297
	GPT-5.2 Global	0.703	0.617	0.496	0.474
	Claude Sonnet 4.5 Global	0.733	0.534	0.366	0.386
	Gemini 3 Flash Global	0.673	0.580	0.434	0.387
	Sabiá 4 Global	0.711	0.458	0.295	0.289
E2.1	W2V + VADER	0.883	0.275	0.354	0.281
	Logistic Regression	0.950	0.291	0.310	0.298
	Random Forest	0.974	0.244	0.250	0.247
	SVM	0.968	0.335	0.326	0.330
	XGBoost	0.975	0.494	0.256	0.258
	BERTimbau	0.974	0.244	0.250	0.247
	GPT-5.2	0.979	0.405	0.356	0.376
	Claude Sonnet 4.5	0.979	0.442	0.312	0.342
	Gemini 3 Flash	0.978	0.402	0.334	0.357
	Sabiá 4	0.976	0.395	0.300	0.323
	Gemini 3 Flash Global	0.877	0.662	0.823	0.734
	GPT-5.2 Global	0.877	0.658	0.839	0.738
	Claude Sonnet 4.5 Global	0.883	0.909	0.484	0.632
	Sabiá 4 Global	0.847	1.000	0.258	0.410

The first experiment revealed performance disparities associated with model ar-

chitecture and context windows. LLMs, when evaluating full texts, demonstrated strong generalization capabilities. Gemini 3 Flash achieved the highest Recall (0.930) and F1-Score (0.746), indicating suitability for high-safety screening scenarios. In contrast, classical models (e.g., XGBoost and Random Forest), constrained to sentence-level analysis, showed substantial performance degradation, confirming that violence often resides in the construction of scenes rather than isolated words. Only BERTimbau (F1-Score of 0.557) partially mitigated this limitation at the sentence level.

In Experiment 2.1 (tolerating Level 1), performance dynamics shifted significantly. GPT-5.2 (sentence-level) achieved the highest F1-Score (0.738). More importantly, this adjustment exposed a critical vulnerability: a complete collapse toward the majority class. Because safe sentences naturally constitute the vast majority of the granular dataset (over 79%), these models optimized their global loss by predicting the negative class for virtually all instances. Consequently, while they achieved superficially high Accuracy, their Recall plummeted to absolute zero. This accuracy paradox demonstrates their complete inability to detect the rare, higher-intensity violence levels (Levels 2 and 3) without synthetic resampling, highlighting a significant limitation of standard NLP pipelines in heavily imbalanced educational domains.

In multiclass scenarios (Experiment 2), adjacency errors were observed during evaluation, with most misclassifications occurring between neighboring levels (e.g., predicting Level 1 instead of Level 2). This pattern reflects the inherent subjectivity in distinguishing implicit tension from explicit physical violence. Importantly, severe errors (e.g., Level 0 classified as Level 3) were rare, indicating that LLMs mitigate critical hallucinations. In binary classification scenarios (Experiments 1 and 2.1), the analysis focuses on false positives (false alarms) and false negatives (missed detections). Remaining errors often involve texts with symbolic violence, complex metaphors, or historical/journalistic contexts. In such cases, distinguishing between acceptable narrative content and content requiring filtering remains challenging even for human evaluators, reinforcing the importance of the safety threshold defined in this study.

The obtained results challenge the assumption that fragmenting long texts benefits computational processing. While classical models rely on short texts (sentences) to operate statistically through frequency-based vectors, LLMs lose valuable psychological and narrative context due to such fragmentation. Violence in educational and youth-oriented texts does not reside solely in isolated words but in the construction of scenes and meanings. This explains why Gemini achieved a strong Recall in the global context yet lost semantic traction when classifying isolated sentences. A fundamental trade-off in supervised classification is also evident, as discussed: maximizing Recall tends to reduce Precision. The hybrid approach evaluated in Experiment 2.1 (enforcing multiclass grading while applying a post-processing rule that tolerates Level 1) proved to be the most viable strategy for real-world applications. GPT-5.2 and Gemini 3 Flash demonstrated the ability to operate effectively within this balance while preserving interpretability through the extraction of narrative evidence. To determine if the performance differences in Experiment 2.1 were statistically significant, a Cochran's Q test was first conducted, revealing a highly significant overall variation in model accuracy ($Q = 676.80$, $p = 6.6986 \times 10^{-140}$) and rejecting the null hypothesis of equal performance. A subsequent post hoc analysis using McNemar's test with a Bonferroni correction was applied to control for Type I error across

pairwise comparisons (45 pairs), with a significance threshold of 1.1×10^{-3} . Notably, while the Large Language Models (LLMs) diverged significantly from all other traditional models ($p < 10^{-14}$), they did not differ significantly from one another ($p > 6.1 \times 10^{-2}$).

5.1. Qualitative Error Analysis and Model Coverage

Analyzing predictive overlap (consensus coverage) across models reveals how different architectures struggle with explicit lexical triggers versus context-dependent harm. Moreover, this coverage analysis suggests that ensemble methods could have a positive effect on overall accuracy by combining the complementary strengths of these distinct models.

Classical Models at the Sentence Level (Experiment 1): The consensus among all classical models correctly classified 1,366 sentences in the binary task, but only a single instance was actually violent. The 12 sentences where all classical models failed were all false negatives typically involving aggressive actions embedded in playful or gaming contexts sentence like “... *inclusion of a cooperative mode, which allows up to four players to join the brawl*”³ evade detection because the aggressive actions (“brawl/pancadaria”), which traditional models fail to weigh correctly against surrounding neutral vocabulary. Conversely, generated systemic false alarms (predicting safe texts as violent) when evaluating fantasy or environmental narratives laden with action verbs. Sentences such as “*Together they face monsters, level up, and team up with other adventurers.*”⁴ triggered the models’ violence thresholds. The classic algorithms anchored onto isolated words (“face/enfrentam”, “monsters/monstros”) without grasping the figurative intent.

Generative LLMs at the Global Level (Experiments 1 and 2): LLMs reached a correct consensus on 229 texts (29 being violent), demonstrating their ability to leverage broad narrative contexts, committing only 7 consensus errors overall. However, the transition to global multiclass grading (Experiment 2) proved challenging. When forced to assign a severity level to the entire text, LLMs missed 51 violent narratives in consensus (20 Level 1, 8 Level 2, and 23 Level 3). This highlights a critical limitation: forcing an LLM to assign a single holistic label to a long narrative dilutes short bursts of violence, masking explicit physical harm within otherwise educational stories.

Multiclass Granular Complexity (Experiment 2): When all models were predicting exact violence levels at the sentence level, the consensus accurately classified 1,261 sentences—all of which were safe (Level 0). The models unanimously failed on 18 specific sentences, distributing errors across the severity scale. A qualitative review of these “hardest samples” illustrates the semantic complexity of the educational domain: at Level 1, subjective descriptions of psychological terror were universally missed due to the lack of physical keywords; at Level 2, descriptions of systemic abuse or imminent fights, such as “... *people sell him to a circus, where Pinocchio works exhaustively and is mistreated.*”⁵ lacked explicit gore, causing models to underestimate the threat; and at Level 3, explicit lethal actions were paradoxically missed when embedded in classic fables (e.g., “*They decided to kill her and take all that fortune at once.*”⁶ (referring to the Golden Goose) or texts containing objective reviews about video-games, as algorithms treated them as inherently safe children’s literature or neutral tech journalism.

³... *inclusão de um modo cooperativo, que possibilita até quatro jogadores entrarem na pancadaria...*

⁴*Juntos enfrentam monstros, sobem de nível e se aliam a outros aventureiros.*

⁵... *as pessoas o vendem para um circo, onde Pinóquio trabalha exaustivamente e é mal tratado.*

⁶... *resolveram matá-la e pegar aquela fortuna toda de uma vez.*

Binary Grouping (Experiment 2.1): At the granular level, all models reached a correct consensus on 1,440 sentences, all of which were safe. However, they unanimously failed to detect 8 violent sentences; these specific errors highlight the difficulty of modeling complex semantics such as psychological abuse, legislative prevention, and animal welfare advocacy. For example, the sentence “*You’re too sensitive... stop being dramatic and swallow your tears.*”⁷ represents explicit psychological violence but lacks physical trigger words, causing all models to severely downgrade its severity. Similarly, excerpts discussing the prevention of harm, such as animal rights advocacy fundamentally confused the models. At the global level, applying this grouping allowed LLMs to successfully agree on 15 genuinely violent texts, significantly reducing the dilution effect observed in Experiment 2, though 8 global narratives were still missed in consensus.

5.2. Control Dataset: Decoding Pedagogical False Alarms

To validate robustness to false positives, Experiment 3 evaluated a subset of the models on a negative control dataset comprising official, validated safe educational items. Due to the financial costs of processing over 3,000 full texts, GPT-5.2 and Claude Sonnet 4.5 were excluded. The generative evaluation focused on Gemini 3 Flash and Sabiá 4 [Laitz et al. 2026], which provided, respectively, an optimal balance between strong contextual reasoning (as observed in previous experiments) and economic feasibility for large-scale curation pipelines.

Table 3. Specificity metrics on the Control Dataset (Experiment 3).

Models	Correct	False Alarms N2+	Specificity
Gemini 3 Flash	2934	121	96.09%
Logistic Regression	2978	77	97.51%
W2V + VADER	3000	55	98.20%
Sabiá 4	3035	20	99.35%
SVM	3051	4	99.90%
Random Forest, XGBoost, and BERTimbau	3055	0	100.00%

As shown in Table 3, classical models (Random Forest, XGBoost, BERTimbau) achieved 100% specificity (0 false alarms). However, as previously established, this is not indicative of superior semantic understanding; it is a symptom. By predicting “safe” for all inputs, they inherently produce zero false alarms, rendering them effectively useless for active curation. In contrast, Generative LLMs actively attempted to classify the texts. Gemini 3 Flash and Sabiá 4 generated 121 and 20 false alarms, respectively. A deep qualitative review of these misclassified “safe” items answers a fundamental question: *What exactly triggers these models in officially approved educational texts?*

The LLMs did not randomly hallucinate violence; rather, they seemed to accurately map the taxonomy to texts that contain genuine violence. Educational materials frequently employ violent descriptions as functional tools for moral lessons, scientific explanations, or cultural heritage. Consequently, these false positives fall into four distinct categories of normalized violence: (i) Direct Peer Aggression and Moral Lessons (texts

⁷*Você é muito sensível... para de drama e engole o choro.*

addressing bullying often depict explicit physical harm to teach a moral lesson.); (ii) Biological Realism and Scientific Facts (natural sciences frequently use graphic language to describe animal behavior; (iii) Myths, Fables, and Cultural Heritage (Indigenous folklore and classic tales are replete with systemic violence, punishments, and survival threats); (iv) Superheroes and Hyperbolic Narrative (Excerpts describing a villain's traps).

This analysis highlights a critical intersection between NLP taxonomies and educational curation. Many items in legacy assessment banks contain high levels of functional violence that may have been acceptable in older guidelines but conflict with modern taxonomies. Generative models, primarily aligned to sanitize toxic interactions, act as overly cautious censors, exposing the inherent subjectivity of what educators consider "safe". This reinforces our proposed hybrid pipeline: allowing LLMs to extract rich narrative context while enforcing a post-processing threshold that tolerates lower-level tension to prevent depletion of educational item banks.

6. Conclusions

This study addresses a gap by adapting official indicative rating guidelines into a computational taxonomy for violence detection in long-form texts. We introduced the *Gru* dataset and demonstrated that while traditional NLP pipelines struggle with implicit narrative violence, Generative LLMs excel at contextual reasoning. A contribution of this work is the proposal of a parameterization strategy that derives a binary curation decision from granular multiclass analysis, optimizing the balance between pedagogical safety and operational scalability. The proposed pipeline is not intended to exclude or replace manual curation. Instead, it serves as a robust decision-support system that augments human expertise. A unique advantage of Generative LLMs, which was not observed in traditional models, is their ability to provide explicit narrative evidence and justifications for their classifications. By pinpointing specific excerpts and explaining the rationale for a severity grade, LLMs empower curators to focus on high-risk items, significantly reducing the cognitive load of manual review while maintaining essential pedagogical oversight.

This study faces notable limitations. First, dataset construction proved exceptionally challenging; there is a severe scarcity of publicly available texts that mirror the exact structural nuances of educational assessment items while simultaneously containing varying degrees of classified violence. Consequently, achieving class balance proved infeasible, reflecting the real-world domain in which safe texts overwhelmingly dominate. Second, this natural imbalance inherently caused the traditional supervised methods to become biased toward the majority class (safe texts), drastically reducing their capability to detect higher-intensity, rare violent instances. Third, while the current findings highlight a clear advantage of Generative LLMs in contextual reasoning, the results are not yet definitively conclusive; a significantly larger and more diverse corpus is required to ensure robust statistical generalization. Finally, the high computational and financial costs associated with large-scale proprietary LLM inference restricted the scope of this study, preventing the execution of more exhaustive experimental configurations and broader baseline comparisons across all granular levels.

Acknowledgment

This work was supported by the CAEd Foundation.

Artificial Intelligence use disclosure

Tools such as ChatGPT and Grammarly were used to revise text.

References

- Barbosa, G., Batista, H., Miranda, P., Santos, J., Isotani, S., Cordeiro, T., Bittencourt, I. I., and Mello, R. F. (2022). Aprendizagem de máquina para classificação de tipos textuais: Estudo de caso em textos escritos em português brasileiro. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 920–931, Porto Alegre, RS, Brasil. SBC.
- Battiston, T. (2025). Classificação de manchetes de cunho violento por meio de mineração de texto. Trabalho de Conclusão de Curso (Graduação em Engenharia de Controle e Automação) – Universidade Federal de Santa Catarina, Campus Blumenau. Acesso em: 22 abr. 2026.
- Brennan, R. L. (2023). *Educational measurement*. Bloomsbury Publishing PLC.
- Burnap, P. and Williams, M. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making: Machine classification of cyber hate speech. *Policy & Internet*, 7.
- Chico, V., Regino, A., Bonacin, R., and Reis, J. (2025). Llm-based automatic generation of multiple-choice questions with meaningful distractors. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, pages 814–828, Porto Alegre, RS, Brasil. SBC.
- Clarke, M. and Luna-Bazaldua, D. (2021). *Primer on Large-Scale Assessments of Educational Achievement*. World Bank Publications, Washington, DC.
- da Justiça e Segurança Pública, B. M. (2018). Guia prático de classificação indicativa. Acesso em: 22 abr. 2026.
- da Justiça e Segurança Pública, B. M. (s.d.). Descritores de conteúdos. Documento institucional, Ministério da Justiça e Segurança Pública, Brasília.
- da Silva, F. R. and Ramos, E. M. L. S. (2023). Diagrama e taxonomia da tipologia das violências. Relatório técnico, Programa de Pós-Graduação em Segurança Pública, Instituto de Filosofia e Ciências Humanas, Universidade Federal do Pará. PPGSP/IFCH/UFPA.
- de Oliveira Trajano, D. (2023). Detecção de linguagem tóxica aplicada a textos em português. Dissertação de mestrado, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, Brasil. Programa de Pós-Graduação em Ciência da Computação.
- de Pelle, R. and Moreira, V. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, pages 510–519, Porto Alegre, RS, Brasil. SBC.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

- Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- dos Santos Rodrigues, R. C. (2023). Evidências de validade de escala para rastreio da violência contra a pessoa idosa. Master’s thesis, Universidade Federal da Paraíba (UFPB).
- Educational Testing Service (2022). Ets guidelines for fair tests and communications. Technical report, Educational Testing Service (ETS), Princeton, NJ.
- Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).
- Furniturewala, S., Jandial, S., Java, A., Banerjee, P., Shahid, S., Bhatia, S., and Jaidka, K. (2024). Thinking fair and slow: On the efficacy of structured prompts for debiasing language models.
- Guide, B. F. (2022). *Detecção automática de discurso de ódio punitivista em redes sociais*. Tese de doutorado, Universidade de São Paulo, São Paulo, Brasil. Faculdade de Filosofia, Letras e Ciências Humanas.
- Haladyna, T. and Rodriguez, M. (2013). *Developing and Validating Test Items*. Routledge, Abingdon, UK.
- Hee, M. S., Sharma, S., Cao, R., Nandi, P., Nakov, P., Chakraborty, T., and Lee, R. K.-W. (2024). Recent advances in online hate speech moderation: Multimodality and the role of large models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4407–4419, Miami, Florida, USA. Association for Computational Linguistics.
- Heru, S., Hidayah, I., and Kusumawardani, S. (2022). Automatic item generation with reading passages: A systematic literature review. In *8th International Conference on Education and Technology (ICET)*, pages 250–255.
- Krug, E. G., Dahlberg, L. L., Mercy, J. A., Zwi, A. B., and Lozano, R., editors (2002). *World Report on Violence and Health*. World Health Organization, Geneva.
- Laitz, T., Almeida, T. S., Abonizio, H., Junior, R. M., Bonás, G. K., Piau, M., Larcher, C., Pires, R., and Nogueira, R. (2026). Sabiá-4 technical report.
- Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Kam-Fai Wong, K. K. and Wu, H., editors, *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.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, San Rafael, CA.
- Mantini, L. A., de Arajo da Silva, Y. P., and Meireles, M. R. G. (2021). Reconhecimento automático de narrativas violentas contra a mulher utilizando classificadores neurais. *Pista: Periódico Interdisciplinar [Sociedade Tecnologia Ambiente]*, 3(2). Seção Discente – Temática Livre.

- Nogueira, J., Castro, J., Larcher, L., Ferreira, R., Barbosa, B., and Souza, J. (2025). Uma análise de qualidade do uso de grandes modelos de linguagem para geração automática de itens avaliativos em português. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, pages 235–248, Porto Alegre, RS, Brasil. SBC.
- Olweus, D. (1993). *Bullying at school: What we know and what we can do*. Wiley-Blackwell, USA.
- Palla, K., García, J. L. R., Hauff, C., Fabbri, F., Damianou, A., Lindström, H., Taber, D., and Lalmas, M. (2025). Policy-as-prompt: Rethinking content moderation in the age of large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, page 840–854, New York, NY, USA. Association for Computing Machinery.
- Razavi, A., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Artificial Intelligence*, pages 16–27.
- Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In Ku, L.-W. and Li, C.-T., editors, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Spertus, E. (1997). Smokey: automatic recognition of hostile messages. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence, AAAI'97/IAAI'97*, page 1058–1065. AAAI Press.
- Trajano, Douglas, Bordini, H., R., and Vieira, R. (2024). OLID-BR: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.
- Vargas, F., Carvalho, I., de Góes, F. R., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 7174–7183, Marseille, France. European Language Resources Association.
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In Doran, J. B. C. and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.