A Multitask Transformer for Offensive Language Detection and Target Identification in HateBR

Guilherme Silva[1], Pedro Silva[2], Matheus Peixoto[1], Gladston Moreira[2] and Eduardo Luz[2]
[1]Postgraduate Program in Computer Science, Federal University of Ouro Preto, Brazil
[2]Computing Department, Federal University of Ouro Preto, Brazil
{guilherme.lopes, matheus.peixoto}@aluno.ufop.edu.br, {silvap, gladston, eduluz}@ufop.edu.br

Abstract

Hate speech detection is often treated as a binary task, ignoring the hierarchical nature of toxicity, such as severity levels and specific target groups. This work presents a Multitask Learning (MTL) approach for the HateBR dataset, utilizing a shared BERTimbau encoder to simultaneously predict binary offensiveness, ordinal severity, and hate speech targets. Our experiments demonstrate that the MTL architecture outperforms Single-Task baselines on the primary offensive detection task, increasing the Matthews Correlation Coefficient from 0.80 to 0.82. Beyond predictive performance, we show that joint training implicitly enforces hierarchical sanity: the unified model yields a 0% target-inconsistency rate (i.e., no cases where a comment is predicted *Non-offensive* while still assigned a hate target). However, we observe negative transfer in the fine-grained multilabel target task (Micro-F1 drops from 0.59 to 0.42), highlighting a trade-off between logical consistency and target attribution under extreme imbalance.

1 Introduction

The growth of public discourse on social platforms has increased the need for automatic systems that can identify and categorize offensive language beyond a simple binary decision (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). In Brazilian Portuguese, this task is particularly challenging due to linguistic variation, slang, and the irony inherent in internet comments (Vargas et al., 2022). To address this, datasets such as HateBR (Vargas et al., 2022) provide hierarchical annotations, capturing not only whether a comment is offensive but also its severity and the specific target groups.

Prior work in offensive language and hate speech detection has explored multi-stage and multi-task formulations to leverage related supervision signals (e.g., offensiveness/type/target) and to mitigate error propagation in hierarchical settings (Zampieri et al., 2019; Park and Fung, 2017; Dai et al., 2020; Plaza-del Arco et al., 2021; Oliveira et al., 2023). While such approaches are widely studied in English and multilingual benchmarks (Zampieri et al., 2019, 2020; Mandl et al., 2020), HateBR offers a structured benchmark for investigating whether joint learning improves both performance and cross-layer consistency in Brazilian Portuguese.

Traditionally, this multi-level problem is addressed through sequential pipelines of independent classifiers. For instance, a system might first detect offensive content and, conditional on a positive prediction, pass the input to subsequent models to determine severity and hate speech targets (Gandhi et al., 2024). While intuitive, this approach has significant downsides. First, it suffers from error propagation: if the primary binary classifier fails, the downstream tasks are never triggered or receive invalid inputs (Diaz-Garcia and Lopez, 2025). Second, independent training ignores the intrinsic correlations between tasks. For example, the presence of a racial slur generally implies high offensiveness. Finally, deploying multiple large language models (LLMs) is computationally inefficient for real-time moderation (Stickland and Murray, 2019).

In this context, Multitask Learning (MTL) emerges as a promising solution (Caruana, 1997). By training a single model to handle multiple objectives simultaneously, MTL allows tasks to share a common feature representation, acting as a regularizer and potentially improving generalization, especially when training data is limited (Ruder, 2017; Zhang and Yang, 2022).

This work presents a multitask architecture on the HateBR dataset. We propose a model that shares a BERTimbau encoder (Souza et al., 2020) across three distinct heads: binary classification, ordinal severity level, and multilabel hate target. Our main contributions are: (1) demonstrating

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1049–1054

that Multitask Learning acts as an effective regularizer, improving the Matthews Correlation Coefficient (MCC) on the primary task and *observing* lower run-to-run variation across two random seeds; (2) providing empirical evidence that joint optimization enforces hierarchical consistency; and (3) establishing a computationally efficient baseline that reduces the deployment memory footprint via parameter sharing. We also document a clear negative-transfer regime in the multilabel target task, which helps delimit when a unified model is preferable in practice.

The remainder of this paper is organized as follows. Section 2 details the multitask architecture, data preprocessing, and training protocols. Section 3 describes the experimental setup, baselines, and evaluation metrics. Section 4 presents the comparative results, discussing the trade-offs between predictive performance, stability, and hierarchical consistency. Finally, Section 5 concludes the study and outlines directions for future work.

2 Methodology

This section details the experimental framework, covering the data pipeline and architectural design.

2.1 Data Preparation and Preprocessing

The study utilizes the HateBR dataset, sourced from the Hugging Face repository (`ruanchaves/hatebr`). The data schema is normalized to guarantee the presence of four core components: the comment text, the binary offensiveness label, the ordinal severity level, and the nine hate target categories. Missing attributes are handled via a deterministic completion rule implied by the hierarchy or explicit configuration overrides. When a sample is labeled as *Non-offensive* (y_{bin}=0), we set the severity to *None* (y_{level}=0) and the target vector to all zeros ($\mathbf{y}_{target}$=**0**), ensuring that labels remain consistent with HateBR's annotation scheme.

The Text preprocessing includes whitespace trimming and standardization. To mitigate the model's reliance on brittle lexical cues, user mentions, and URLs are optionally masked with special tokens (`<USER>` and `<URL>`). The processed text is stored and aligned with the encoded labels: binary values for offensiveness ($\{0, 1\}$), ordinal integers for severity (0-3), and a one-hot vector for the nine target categories ($\{0, 1\}^9$).

2.2 Model Architecture and Tasks

We employ BERTimbau Base (`neuralmind/-bert-base-portuguese-cased`) as the shared Transformer encoder. The `[CLS]` token embedding from the final hidden layer serves as the unified semantic representation.

In the Multitask configuration, the architecture operates as a hard-parameter sharing network, as illustrated in Figure 1. The tokenized sequences, represented by `input_ids` and `attention_mask`, are processed by the shared BERTimbau encoder. The resulting `[CLS]` vector acts as a universal feature representation, capturing high-level semantic information. This vector is then projected simultaneously into three distinct decision spaces via independent linear heads, which produce unnormalized logits for the following objectives:

1. **Binary Offensiveness:** A 2-logit output head responsible for detecting the presence of offensive content.

2. **Ordinal Level:** A 4-logit output head mapping the severity to specific ordinal classes (0–3).

3. **Hate Targets:** A 9-logit output head identifying specific target groups (e.g., Racism, Homophobia) in a multilabel setting.

In Single-Task baselines, the same encoder architecture is paired with a single dedicated head for the respective task.

2.3 Loss Function

The loss function is adapted to each task: CrossEntropyLoss for the binary and ordinal tasks, and BCEWithLogitsLoss for the multilabel target task. The joint multitask objective is defined as a weighted sum of individual losses:

$$\mathcal{L} = w_{bin}\mathcal{L}_{bin} + w_{level}\mathcal{L}_{level} + w_{target}\mathcal{L}_{target} \quad (1)$$

where $\mathcal{L}$ represents the global joint loss minimized during optimization. The terms $\mathcal{L}_{bin}$ and $\mathcal{L}_{level}$ correspond to the Cross-Entropy losses computed for the binary and ordinal severity heads, respectively, while $\mathcal{L}_{target}$ denotes the Binary Cross-Entropy loss for the multilabel target head. The coefficients w_{bin}, w_{level}, and w_{target} are scalar hyperparameters that control the relative contribution of each task to the total gradient; in our experiments, we set all weights to 1.0 to establish a controlled baseline

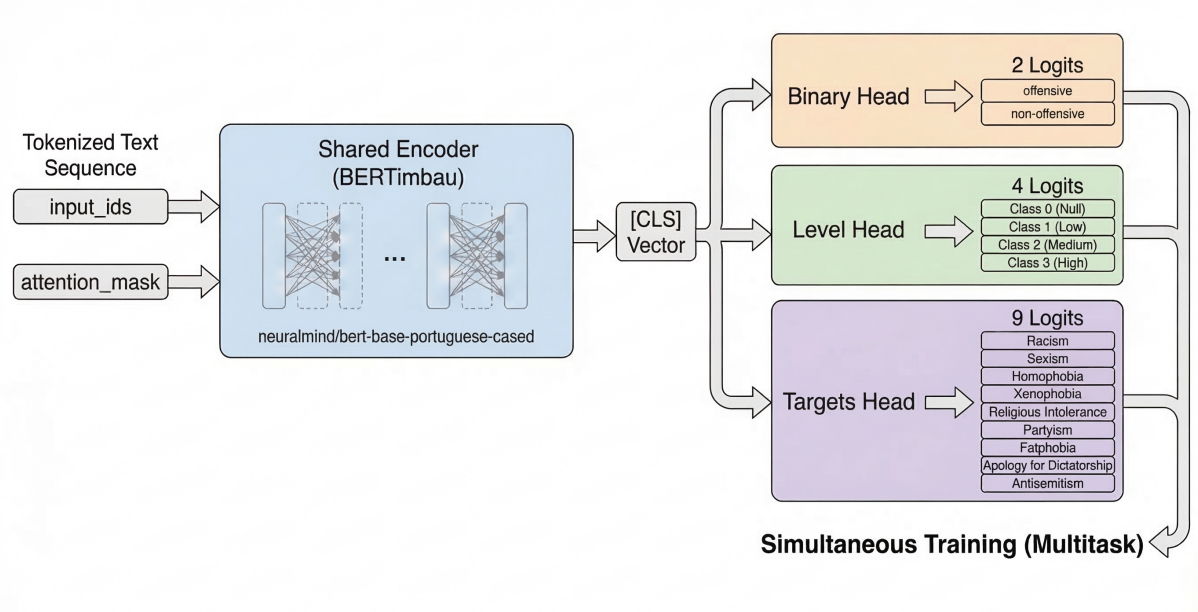

Figure 1: Overview of the Multitask Learning architecture. The shared BERTimbau encoder extracts a context-rich `[CLS]` representation, which is simultaneously fed into three task-specific heads to predict binary offensiveness (2 logits), severity level (4 logits), and hate speech targets (9 logits).

that isolates the effect of hard parameter sharing (without task reweighting). We emphasize that learned or imbalanced weighting may be necessary to mitigate negative transfer in the multilabel target head, but is outside the scope of this paper.

3 Experiments

To validate the proposed approach, we designed a comparative experimental protocol between specialized Single-Task Learning (STL) baselines and the proposed Multitask Learning (MTL) architecture. All experiments utilize the same backbone (BERTimbau Base), seed configuration, and preprocessing pipeline. The implementation of the proposed approach is publicly available at `https://github.com/silvagal/hatebr_multitask_transformer.git`.

3.1 HateBR Dataset

The study uses the HateBR dataset, the first large-scale, expert-annotated dataset of hate speech in Brazilian Portuguese (Vargas et al., 2022). The dataset comprises 7,000 comments collected from Instagram profiles of Brazilian politicians, offering a challenging, realistic scenario characterized by informal language, irony, and context-dependent toxicity.

HateBR has a three-layer hierarchical annotation scheme, validated by three independent experts with high inter-annotator agreement:

1. **Binary Classification (Layer 1):** A fundamental label indicating whether the comment is *Offensive* or *Non-offensive*.

2. **Offensiveness Level (Layer 2):** An ordinal categorization of severity, divided into four classes: *None* (0), *Slightly Offensive* (1), *Moderately Offensive* (2), and *Highly Offensive* (3).

3. **Hate Speech Target (Layer 3):** A multilabel classification identifying the targeted group. It encompasses nine distinct categories: *Xenophobia, Racism, Homophobia, Sexism, Religious Intolerance, Partyism, Apology for Dictatorship, Antisemitism,* and *Fatphobia*.

3.2 Data Splitting Strategy

We established a fixed partition of the 7,000 comments comprising the HateBR dataset. The dataset was divided into three subsets: training (4,480 samples), validation (1,120 samples), and test (1,400 samples). This distribution corresponds to allocating 20% of the total data for final testing, with the remaining development data split between training (80%) and validation (20%). All partitions were generated with a deterministic random seed to preserve the offensive/non-offensive rate across splits (stratified on the binary label). This split is generated once and reused across all runs.

3.3 Tokenization

Inputs are processed using the BERTimbau tokenizer with truncation and padding to a fixed maximum length (default 128 tokens). Data loaders employ shuffling for the training set and deterministic ordering for validation and test sets. The random seeds reported in Section 4 control both model initialization and training-time randomness (including training-set shuffling); validation and test evaluation are deterministic given the fixed split.

3.4 Experimental Setup and Baselines

We train four distinct models:

- **STL-Binary:** A BERT model fine-tuned exclusively for binary offensive detection.
- **STL-Level:** A BERT model fine-tuned exclusively for the 4-class severity classification.
- **STL-Target:** A BERT model fine-tuned exclusively for the 9-class multilabel hate target identification.
- **MTL-Unified:** The proposed model with a shared encoder and three concurrent output heads, optimizing the joint loss $\mathcal{L}_{total}$, presented in Section 2.3.

Hyperparameters were standardized across all runs: 20 epochs. To prevent overfitting, we employ early stopping with a patience of 2 epochs, a batch size of 16, a learning rate of $2e^{-5}$, and a weight decay of 0.01. The experiments were conducted using mixed-precision (FP16) on a single CUDA device. Additionally, models are trained with the AdamW optimizer and a linear learning rate scheduler (warmup ratio of 0.06), and gradient clipping (norm ≤ 1.0).

3.5 Inference and Metrics

The selection criterion depends on the configuration: for single-task models, we monitor the specific task metric (Positive-class F1 for binary, Macro-F1 for level, Micro-F1 for target), whereas for the multitask model, the primary metric is the arithmetic mean of these three scores. Once the checkpoint with the best validation performance is restored, inference for the multilabel target task is performed by applying a sigmoid activation followed by a threshold. While a default threshold of 0.5 is standard, the pipeline includes an adaptive mechanism that fine-tunes this value to maximize Micro-F1 if the model predicts no positive labels during validation, a critical adjustment for highly imbalanced datasets. Importantly, any threshold selection is performed *only* on the validation set and then frozen for test-time evaluation, avoiding leakage. Finally, evaluation metrics include standard F1-scores, alongside Exact Match (Subset Accuracy) to quantify the fraction of incorrect individual labels.

4 Results and Discussion

We evaluate the proposed Multitask Learning architecture against the Single-Task Learning baselines. To probe initialization variability under a fixed split and standardized hyperparameters, all experiments were executed using two distinct random seeds (88 and 89). The reported values represent the aggregated mean and standard deviation across these independent runs. Given the small number of runs, variability-related observations should be interpreted as preliminary.

4.1 Predictive Performance and Stability

Table 1 summarizes the comparative performance. It is important to clarify the structural distinction between the columns: the STL (Baseline) column aggregates results from three independent models, each specialized for the task indicated in the corresponding row. In contrast, the MTL column reports the performance of a single unified model evaluated across all three tasks simultaneously.

Table 1: Performance comparison between Single-Task (STL) and Multitask (MTL) models across seeds 88 and 89.

Task	Metric	STL (Baseline)	MTL
Binary (Offensive)	Accuracy	0.8996 ± 0.0096	$\mathbf{0.9111 \pm 0.0045}$
	F1-Score (Pos)	0.8965 ± 0.0128	$\mathbf{0.9088 \pm 0.0053}$
	MCC	0.8012 ± 0.0167	$\mathbf{0.8232 \pm 0.0086}$
Level (Severity)	Accuracy	0.6443 ± 0.0121	$\mathbf{0.6454 \pm 0.0076}$
	F1-Macro	0.4634 ± 0.0031	$\mathbf{0.4717 \pm 0.0069}$
	Balanced Acc	0.4641 ± 0.0007	$\mathbf{0.4720 \pm 0.0056}$
Target (Hate Type)	Subset Accuracy	$\mathbf{0.9421 \pm 0.0071}$	0.8993 ± 0.0263
	Micro-F1	$\mathbf{0.5923 \pm 0.1011}$	0.4197 ± 0.0639

According to Table 1, the MTL approach demonstrated superior performance in the primary binary classification task, achieving higher Accuracy (91.11% vs. 89.96%) and F1-Score (0.9088 vs. 0.8965). Besides that, the Matthews Correlation Coefficient (MCC) also improved from 0.8012 to 0.8232. In the Level task, MTL also surpassed the baseline in Balanced Accuracy ($+0.8\%$) and

Macro-F1, suggesting that the shared representation helps distinguish severity gradations more effectively than independent models.

Across the two seeds, we observe lower run-to-run variation for Binary Accuracy in the MTL model (± 0.0045) compared to the Single-Task baseline (± 0.0096). While this is consistent with the intuition that auxiliary tasks can stabilize shared representations, we treat this evidence as preliminary given the limited number of runs.

4.2 Hierarchical Consistency

A distinct advantage of our architecture is the intrinsic enforcement of logical consistency across layers, obviating the need for external post-processing rules. We quantified this using the Target Inconsistency Rate, defined as the proportion of samples where the model predicts a specific hate target despite classifying the text as "Non-Offensive". Our experiments revealed a Target Inconsistency of 0.0000 ± 0.0000 (0%) and a negligible Level Inconsistency of 0.0014 ± 0.0010 (0.14%). Across all runs, the multitask model produced zero logical contradictions regarding hate targets, confirming that the shared encoder successfully internalized the hierarchical dependency that a non-offensive classification implies the absence of a hate target. From a deployment perspective, this property is particularly valuable: contradictory outputs ("clean" text paired with "racism") are unacceptable in real moderation pipelines, and MTL reduces such failure modes without additional rules.

4.3 Challenges in Multilabel Targets

According to Table 1, while the core tasks benefited from positive transfer, the multilabel Target task suffered from negative transfer (Micro-F1 0.42 vs 0.59). This is attributed to the extreme class imbalance in HateBR. In this regime, the multilabel loss is dominated by negatives and sparse positives, and the gradients from the high-frequency binary objective can dominate updates to the shared encoder. This may induce representational pressure toward coarse offensiveness cues at the expense of the finer distinctions required for target attribution, especially under equal task weighting. In the joint optimization, the gradients are dominated by the high-frequency binary signals. Consequently, performance on the most frequent target class, *Partyism*, that according to results, dropped from 0.72 (STL) to 0.51 (MTL). These findings delimit a practical trade-off: the unified model improves the primary moderation signal and enforces hierarchical sanity, but robust target attribution likely requires additional balancing strategies (e.g., reweighting or learned task weights) not explored here.

5 Conclusion

This work introduced a unified Multitask Learning architecture for hierarchical hate speech detection in Brazilian Portuguese. Our experiments on the HateBR dataset confirmed that the shared representation acts as an effective regularizer, improving the primary binary classification performance (MCC increased from 0.80 to 0.82) and yielding strictly consistent predictions across hierarchical layers (0% target inconsistency). Qualitatively, the model enforced strict semantic consistency, achieving a 0% rate of hierarchical contradictions without external post-processing. However, we observed negative transfer in the fine-grained target identification task due to extreme class imbalance (Micro-F1 drop from 0.59 to 0.42). We also observe lower run-to-run variation on the primary task across two seeds. Future work will address this limitation by investigating dynamic loss-weighting strategies, such as GradNorm, and data augmentation techniques to better balance learning dynamics between dominant and rare categories.

Acknowledgment

This work was supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq - grants 308400/2022-4, 307151/2022-0), *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil (CAPES - grant 001), and *Fundacão de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG, grants APQ-01768-24). We also thank the UFOP/PPGCC for their support.

References

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. 2020. Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2060–2066, Barcelona (online). International Committee for Computational Linguistics.

Jose A Diaz-Garcia and Julio Amador Diaz Lopez. 2025. A survey on cutting-edge relation extraction techniques based on language models. *Artificial Intelligence Review*, 58(9):287.

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

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8):e13562.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th annual meeting of the forum for information retrieval evaluation*, pages 29–32.

Amanda Oliveira, Thiago Cecote, Pedro Silva, Jadson Gertrudes, Vander Freitas, and Eduardo Luz. 2023. How good is chatgpt for detecting hate speech in portuguese? In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, pages 103–112.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. *arXiv preprint arXiv:2109.10255*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417.

Asa Cooper Stickland and Iain Murray. 2019. BERT and pals: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5986–5995.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Yu Zhang and Qiang Yang. 2022. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609.