Multi-Task Learning for Implicit Hate Detection

Semanur Avşar **Ines Altemir Marinas** **Florian Tanguy**

Group 3

Abstract

This work addresses the challenge of detecting implicit hate speech, a type of harmful online content that escapes traditional detection systems by using subtle language like sarcasm and stereotypes. The issue becomes more difficult due to poor domain generalization across different platforms which limits models effectiveness. We propose a Multi-Task Learning (MTL) framework to equip a BERT-based model with a deeper contextual understanding. By jointly training on the main task and auxiliary tasks of sarcasm, stereotype, and hate subtype classification, our solution is designed to improve both the accuracy of implicit hate detection and its cross-domain robustness.

Keywords: Implicit Hate Speech, Multi-Task Learning, Domain Generalization

1. Introduction

The detection of hate speech, defined by the United Nations as discriminatory language targeting individuals or groups [1], is critical for preventing social division and real-world harm. However, a key limitation in current systems is their poor performance on implicit hate speech. These subtle forms, often using sarcasm, stereotypes, or coded references, are easily missed by models trained primarily on explicit data, limiting their real-world effectiveness [2]. This challenge is compounded by poor domain generalization, where models tend to overfit to specific datasets and fail across platforms [3]. To address these issues, our objective is to develop a multi-task BERT model that detects implicit hate while simultaneously learning its related components: sarcasm detection, stereotype classification, and hate speech subtype identification. We hypothesize that exposing the model to these socially relevant auxiliary tasks will improve both its contextual understanding and cross-domain generalization.

EE-559: *Deep Learning*, 2025.

2. Related Work

Explicit hate speech detection has been extensively studied across a wide range of benchmark datasets, especially in English. Among the most widely used is the dataset by Davidson et al., [4] which categorizes tweets into hate speech, offensive language, or neither, and reports classification F1-score exceeding 90% . Similarly, *OLID* [5] adopts a hierarchical annotation schema that distinguishes between types and targets of offensive speech, supporting more granular classification reaching around 80% F1-score. Furhermore, there is work on merging different datasets such as *MetaHate* [6] to standardize data access and evaluation that can achieve robust in-domain performance and serve as stronger baselines for comparative evaluation reaching around 80% F1-score.

Compared to its explicit counterpart, implicit hate speech detection remains significantly understudied. While explicit hate is often direct and identifiable through lexical cues, implicit hate typically relies on subtle linguistic mechanisms such as sarcasm, stereotypes, which makes it more difficult to detect using traditional supervised learning approaches.

Recent work has attempted to address this gap. [2] introduced the *Latent Hatred* dataset, a benchmark explicitly designed to capture nuanced, context-dependent forms of hate speech. Their mention that models which perform well on explicit hate datasets fail to address implicit hate and create a new dataset for implicit hate detection. Their best models evaluated in the Latent Hatred dataset reached only around 70.4% F1-score for binary classification.

In a similar direction, [7] proposed the *ToxiGen* benchmark, which leverages generative models to synthesize subtly toxic statements that are indistinguishable from benign ones without context. The authors observed that even advanced models like RoBERTa and ToxiGen's own classifier underperformed on these generated samples, again confirming the difficulty of modeling implicit hate.

These studies underscore the limitations of existing hate speech detection systems when applied to more covert and socially nuanced forms of toxicity, and highlight the need for models that go beyond surface-level lexical features.

Furthermore, even though machine learning models for

hate speech detection have achieved high in-domain performance, they often fail to generalize to out-of-domain scenarios and their performance may drop signifantly [3]. This lack of robustness undermines real-world applicability. To address this, [8] explores domain adaptation techniques such as MixUp regularization, curriculum labeling, and adversarial domain adaptation, achieving modest improvements in out-of-domain performance. However, these methods still rely on access to unlabeled target domain data during training and are inherently limited to the adapted domain. As such, they do not solve the broader problem of domain-agnostic generalization, indicating a need for architectures that can robustly infer hate across domains without fine-tuning for each.

3. Method

We implemented a multi-task learning (MTL) architecture built on top of a pre-trained BERT encoder to enhance the detection of implicit hate speech. Our approach was inspired by recent advancements in multi-task learning for hate speech detection, particularly by two key studies.

First, [9] proposed a BERT-based MTL framework where hate speech detection was learned jointly with sentiment, emotion, and target identification. Their work showed that incorporating affective and contextual signals improved the model's performance across related classification tasks, which motivated us to incorporate auxiliary signals that can enrich the semantic understanding of implicit hate.

Second, [10] developed a multi-task architecture aimed specifically at improving cross-domain generalization in hate speech detection. They introduced an auxiliary task of emotion classification and demonstrated consistent improvements in cross-domain evaluations. Their findings showed that emotional cues act as useful inductive biases, helping models generalize across diverse online platforms.

These studies collectively support the idea that enriching the model with social and affective cues leads to better performance and inreased domain generalizability . Our work builds upon this foundation while specifically adapting it to the subtleties of implicit hate.

Our central hypothesis is that co-training on auxiliary tasks that reflect core components of implicit hate such as sarcasm, stereotyping, and fine-grained implicit hate subtypes can improve both classification performance and domain generalization.

Our model consists of a shared transformer backbone followed by four task-specific classification heads:

- A main head for classifying into *Explicit Hate*, *Implicit Hate*, and *Non-Hate*: 3-class version of the *Latent Hatred* dataset [2] is used for this purpose.
- Auxiliary heads for:
  - Fine-grained implicit hate subtype classification: Fine-grained version of the *Latent Hatred* dataset [2] that annotates subtypes like stereotype, incitement, grievance is used for this purpose,
  - Sarcasm detection: *iSarcasmEval* [11], which is a dataset capturing intended sarcasm in tweets is used for this purpose. It helps the model identify sarcastic expressions that often mask hate speech.
  - Stereotypical bias detection: *StereoSet* [12] which measures stereotypical bias across domains (e.g., gender, race) is used for this purpose. It helps the model identify social cues.

We are using Latent Hatred dataset because it also contains fine-grain labels.

Furthermore, in order to test out of domain generalizability, we have used ToxiGen [7] which is a large-scale adversarial dataset for implicit toxicity detection.

4. Validation

Experimental Setup. Our experiments use a bert-base-uncased model trained with the AdamW optimizer. Following the methodology in the *Latent Hatred* paper [2], we use a 60%-20%-20% stratified data split and evaluate performance with the macro F1-score to account for class imbalance. Texts underwent standard preprocessing, such as normalization and special token replacement (e.g., for URLs and mentions, etc.), before tokenization. Our baseline is a single-task model, against which we compare our multi-task models trained using stochastic sampling of batches with a weighted loss for each task.

A direct comparison to the results in the *Latent Hatred* paper [2] is challenging due to differences in task formulation. While the source paper reports binary classification metrics (implicit vs. non-hate), their dataset contains three classes, and their handling of the 'explicit hate' category is unspecified. To provide a more transparent and granular evaluation, our work retains all three classes and reports macro-F1 scores on this more complex ternary task. This methodological choice explains the difference between our baseline results and those in the literature. The results we give is the results evaluated on the test set of Latent Hatred dataset.

Impact of Auxiliary Task Weighting. We first investigated the effect of varying the loss weights for the auxiliary tasks. As shown in Table 1, we tested weights of 0.5, 1.0, and 2.0. While a weight of 2.0 yielded the highest accuracy (75.4%), a weight of 1.0 ($\mathbf{BERT_{1.0}}$) achieved the

best F1-score (61.3%). This highlights a classic trade-off where higher auxiliary weights may improve classification on the majority "Non-Hate" class but can slightly reduce the model's sensitivity to minority classes. Based on its balanced F1-score, the model with $w = 1.0$ (referred to as **BERT_{all}**) was chosen for out-of-domain evaluation.

Model	P	R	F	Acc
$\text{BERT}_{0.5}$	64.6	58.9	60.9	74.9
$\text{BERT}_{1.0}$	62.1	**60.9**	**61.3**	74.2
$\text{BERT}_{2.0}$	**66.5**	58.4	60.6	**75.4**
$\text{BERT}_{\text{baseline}}$	60.3	60.7	60.5	73.2

Table 1. Classification performance metrics for ternary classification (Non-Hate, Implicit Hate, and Explicit Hate) for MTL models with different auxiliary task weights. **$\text{BERT}_{\text{baseline}}$** is the single-task model. For example, **$\text{BERT}_{\text{2.0}}$** has $w = 1$ for the main task and $w = 2$ for each auxiliary task.

Ablation Study. To understand the contribution of each auxiliary task, we conducted an ablation study by removing auxiliary tasks. The results, presented in Table 2, are revealing. The model trained with only the fine-grained implicit hate labels (**$\text{BERT}_{\text{fine-grain-label}}$**) achieved the highest F1-score (63.1%), a 4.3% improvement over the baseline. This strongly suggests that providing highly relevant, domain-specific sub-categories is the most effective way to improve in-domain performance. Conversely, training with only sarcasm and stereotype detection performed slightly below the baseline, indicating a potential for negative transfer where the features for general-purpose sarcasm are not perfectly aligned with its use in implicit hate speech.

Model	P	R	F	Acc
BERT_{all}	62.1	60.9	61.3	74.2
$\text{BERT}_{\text{fine-grain-label}}$	63.4	**62.8**	**63.1**	**75.7**
$\text{BERT}_{\text{sarcams+stereotype}}$	**64.4**	57.1	59.4	74.3
$\text{BERT}_{\text{baseline}}$	60.3	60.7	60.5	73.2

Table 2. Classification performance metrics for ternary classification (Non-Hate, Implicit Hate and Explicit Hate) for MTL model with only certain auxiliary tasks.

Out-of-Domain Generalization. The core test of our hypothesis was evaluating the model's ability to generalize. We evaluated our best balanced model, **BERT_{all}**, on the unseen *ToxiGen* dataset. As shown in Table 3, the MTL model outperformed the single-task baseline, with a 1.2% increase in F1-score and a robust 4.0% increase in accuracy. This improvement confirms that exposure to related social concepts through MTL enhances the model's ability to generalize, but was below our expectations likely due to the limited scope of implicit hate categories considered.

Model	F1 (%)	Accuracy (%)
$\text{BERT}_{\text{baseline}}$	62.7	60.9
BERT_{all}	**63.9 (+1.2↑)**	**64.9 (+4.0↑)**

Table 3. Classification performance metrics on out-of-domain dataset Toxigen. **BERT_{all}** was chosen as the MTL with optimized weights for the auxiliary tasks.

Error Analysis. To assess whether our model was learning the intended nuances, we analyzed its misclassifications of stereotypical and ironic implicit hate from the test set. Table 4 shows that the MTL framework improved performance in this area. The **$\text{BERT}_{\text{0.5}}$** model reduced the misclassification of stereotypical hate as "Non-Hate" by 6.5% and cut the confusion rate for ironic hate by 7%. This is the strongest evidence that our approach works as intended: it successfully teaches the model to recognize the subtle linguistic mechanisms of implicit hate, preventing harmful content from being dismissed as benign.

Model	Confusion on Stereotypical (%)	Confusion on Irony (%)
$\text{BERT}_{\text{baseline}}$	24.9	10.9
$\text{BERT}_{0.5}$	**18.4**	**3.9**
$\text{BERT}_{1.0}$	19.0	7.8
$\text{BERT}_{2.0}$	27.9	9.3

Table 4. Percentage of stereotypical-implicit-hate and irony-implicit-hate misclassified as non-hate, for different models. A reduction of 6.5% and 7% respectively on each task.

Limitations. Despite these positive outcomes, our work has limitations. Our model's understanding is constrained by the specific auxiliary tasks of sarcasm and stereotypes; it is not equipped to handle other implicit forms like coded language. The datasets used do not cover the entire diversity of domains and cultural contexts, which may still limit the model's generalization. These factors underscore that while our MTL approach is a significant step forward, implicit hate detection remains a complex and open challenge.

Future Work. Building on the limitations, this work could be extended by expanding auxiliary tasks by incorporating additional categories of implicit hate speech from broader contextual domains.

5. Conclusion

Our multi-task BERT model improves implicit hate detection by jointly learning sarcasm, stereotypes, and hate subtypes. While gains in domain generalization were modest, the model excelled at recognizing hate hidden in irony and stereotypes. This confirms that learning social context builds more reliable detection systems than single-task approaches and advances efforts to create safer online spaces.

References

[1] United Nations, "What is hate speech?," n.d.

[2] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang, "Latent hatred: A benchmark for understanding implicit hate speech," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 345–363, 2021.

[3] T. Wullach, A. Adler, and E. Minkov, "Towards hate speech detection at large via deep generative modeling," in *IEEE Internet Computing*, pp. 48–56, 2021.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pp. 512–515, 2017.

[5] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proceedings of NAACL*, 2019.

[6] P. Piot, P. Martín-Rodilla, and J. Parapar, "Metahate: A dataset for unifying efforts on hate speech detection," in *Proceedings of the 18th International AAAI Conference on Web and Social Media (ICWSM)*, 2024.

[7] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 940–955, 2022.

[8] F. Ludwig, K. Dolos, T. Zesch, and E. Hobley, "Improving generalization of hate speech detection systems to novel target groups via domain adaptation," in *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pp. 29–39, 2022.

[9] F. M. Plaza-del-Arco, M. D. Molina-Gonzalez, L. A. Ureña-Lopez, and M. T. Martin-Valdivia, "A multi-task learning approach to hate speech detection leveraging sentiment analysis," pp. 1–1, 2021.

[10] S. Y. Hong and S. Gauch, "Improving cross-domain hate speech generalizability with emotion knowledge," *arXiv preprint arXiv:2311.14865*, 2023.

[11] I. A. Farha, S. V. Oprea, S. R. Wilson, and W. Magdy, "Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in english and arabic," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022.

[12] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5356–5371, 2021.