Multi-Task Learning for Implicit Hate Detection

Inés Altemir Marinas, Florian Tanguy, Semanur Avşar
Group 3

Problem definition

Online hate speech has evolved beyond explicit slurs to sophisticated implicit forms using sarcasm, stereotyping, and coded language. Current detection systems, designed primarily for explicit hate, fail to capture these subtle expressions that rely on contextual understanding and linguistic nuance. Existing approaches suffer from critical limitations including reliance on surface-level lexical patterns, single-task optimization that ignores hate speech's connections to broader affective phenomena, and poor cross-domain generalizability.

To address these challenges, we propose a Multi-Task Learning approach using a BERT-based model trained on implicit hate detection alongside auxiliary tasks of sarcasm and stereotype detection.

Key Related Works

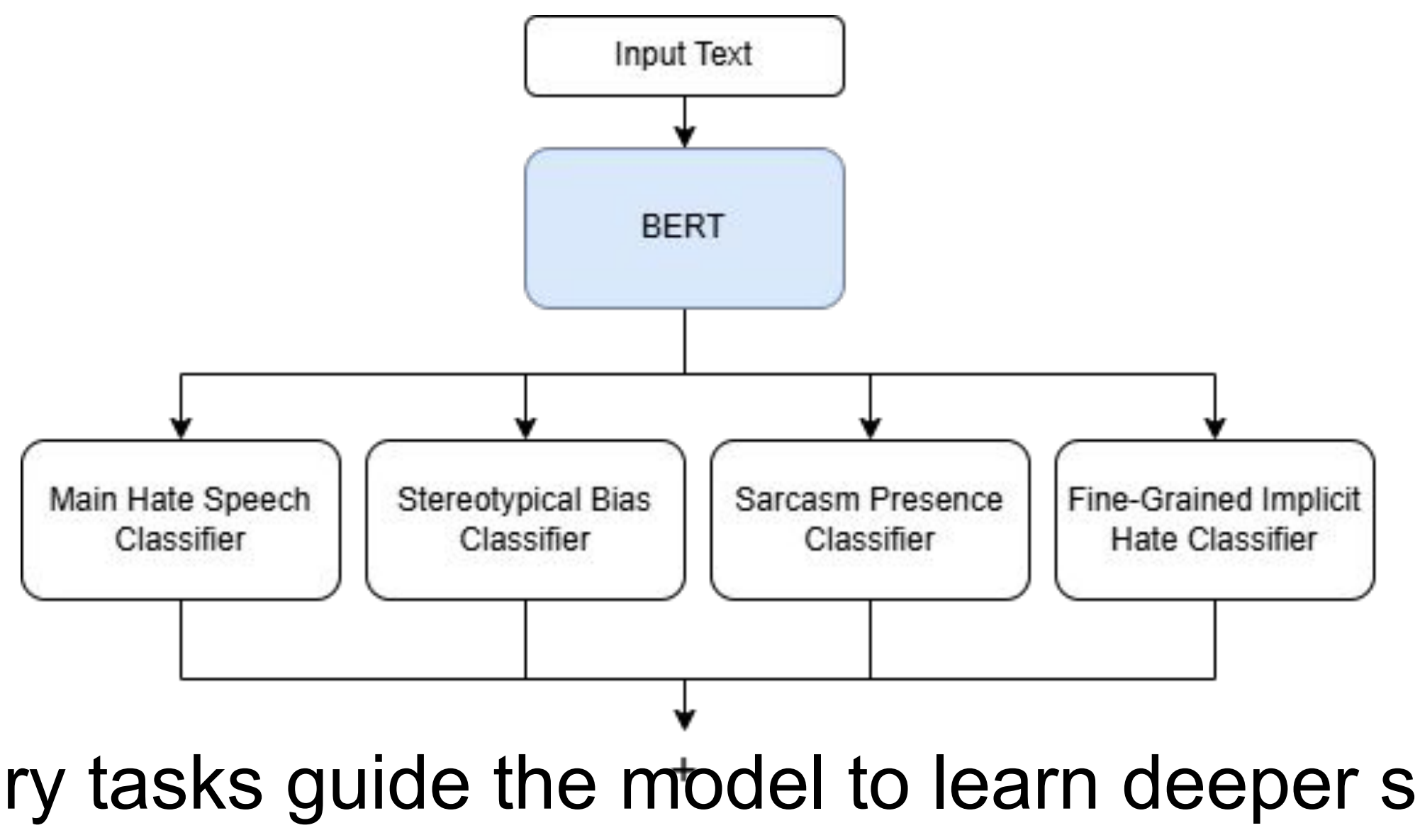
Identifying *implicit* hate speech, which relies on subtle sarcasm and stereotypes, is a key challenge [1]. While multi-task learning (MTL) benefits general hate speech using affective cues [2] and emotion knowledge aids generalizability [3], a dedicated MTL approach co-learning the intrinsic components of *implicit* hate itself remains underexplored. The current literature lacks an MTL framework specifically leveraging sarcasm, stereotype detection, and fine-grained implicit labels as auxiliary tasks to directly improve implicit hate classification. Our work targets this by using these specific signals to deepen model understanding of subtle biases.

Method

We implement a multi-task BERT-based architecture that jointly learns from:

- Main Task: Implicit hate speech classification
- Auxiliary Tasks: Sarcasm detection, stereotype detection, and fine-grained implicit hate categorization

This setup leverages a shared transformer encoder and multiple task-specific classification heads.


Input Text
BERT
Main Hate Speech Classifier
Stereotypical Bias Classifier
Sarcasm Presence Classifier
Fine-Grained Implicit Hate Classifier

The auxiliary tasks guide the model to learn deeper semantic and social cues, thereby enhancing both performance and domain generalizability.

Dataset(s)

- **Latent Hatred**: 3-class version: Implicit Hate, Explicit Hate, and Non-Hate (used as the main task)
  - Fine-grained version: Detailed implicit hate subtypes such as stereotype, grievance, etc. (used as an auxiliary task)
- **iSarcasmEval** Used as an auxiliary task to help the model capture sarcasm
- **StereoSet** Used as an auxiliary task to help the model capture social bias cues
- **ToxiGen** Used only at evaluation time to assess how well the trained model generalizes to out-of-domain hate content

References

[1] 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*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 345–363. doi: 10.18653/v1/2021.emnlp-main.29.

[2]F. M. Plaza-del-Arco, S. Halat, S. Padó, and R. Klinger, “Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language,” arXiv:2109.10255 [cs.CL], 2022. [Online]. Available: https://arxiv.org/abs/2109.10255

[3]S. Y. Hong and S. Gauch, “Improving Cross-Domain Hate Speech Generalizability with Emotion Knowledge,” arXiv:2311.14865 [cs.CL], 2023. [Online]. Available: https://arxiv.org/abs/2311.14865

Results

Multi-Task Learning with Varying Auxiliary Task Weights

Model	P	R	F	Acc
BERT$_{0.5}$	64.6	58.9	60.9	74.9
BERT$_{1.0}$	62.1	**60.9**	**61.3**	74.2
BERT$_{2.0}$	**66.5**	58.4	60.6	**75.4**
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 model with different auxiliary task weights. ***BERT$_{\text{baseline}}$*** *is the single-task model. For example* ***BERT$_{2.0}$*** *has w=1 for principal task and w=2 for each aux task*

Impact of removing auxiliary tasks

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

Domain Generalization Performance on ToxiGen

Model	F	Acc
BERT$_{\text{baseline}}$	62.7	60.9
BERT$_{\text{all}}$	63.9 **(+1.2↑)**	64.9 **(+4.0↑)**

Table 3.Classification performance metrics on out-of-domain dataset Toxigen. BERT$_{\text{all}}$ was chosen as the MTL with optimized weights for the auxiliary tasks

Implicit Hate Confusion with Non-Hate

Model	Confusion on Stereotypical	Confusion on Irony
BERT$_{\text{baseline}}$	24.9 %	10.9 %
BERT$_{0.5}$	**18.4 %**	**3.9%**
BERT$_{1.0}$	19.0 %	7.8 %
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 26% and 64% respectively on each task

- MTL obtains a 1.3% increase in F1 score for the best combination of weights with all auxiliary tasks with respect to the baseline
- MTL with only fine-grained labels as auxiliary task obtains a 4.3% increase in F1 score compared to the baseline, which aligns with expectations as it directly captures the specific content of each implicit hate instance in the dataset.
- MTL with weights=0.5 for all auxiliary task is most successful at reducing the misclassification of stereotypical-implicit Hate and irony-implicit Hate with Non-Hate
- Out-of-domain performance improved as well but was below our expectations likely due to the limited scope of implicit hate categories considered.

Limitations

- Our model focuses on specific forms of implicit hate. The omission of other forms of implicit hate and the lack of skills such as compositional reasoning and discourse relation are not covered in this study may limit the performance increase.
- The datasets used do not cover the entire diversity of domains and cultural contexts, which may limit the model's generalization

Conclusion

While in-domain performance showed only modest improvement, the gains in out-of-domain generalization were more apparent. Most notably, MTL significantly reduced misclassification rates on stereotypical and ironic content, showing that auxiliary tasks targeting specific forms of implicit hate help the model better recognize those nuances.

EE-559: Deep Learning, 2025