AmpleHate: Amplifying the Attention for Versatile Implicit Hate Detection

Yejin Lee, Joonghyuk Hahn, Hyeseon Ahn and Yo-Sub Han*
Yonsei University, Seoul, Republic of Korea,
{ssgyejin,greghahn,hsan,emmous}@yonsei.ac.kr

Abstract

Implicit hate speech involves subtle and indirect expressions of prejudice or hostility toward a group. Detecting it is challenging because it relies on nuanced context and implication rather than explicit offensive language. Current approaches rely on contrastive learning, which is shown to be effective on distinguishing hate and non-hate sentences. Humans, however, detect implicit hate speech by first identifying specific *targets* within the text and subsequently interpreting how these targets relate to their surrounding context. Motivated by this reasoning process, we propose AmpleHate, a novel approach designed to mirror human inference for implicit hate detection. AmpleHate identifies explicit targets using a pretrained Named Entity Recognition model and captures implicit target information via [CLS] tokens. It computes attention-based relationships between explicit, implicit targets and sentence context and then, directly injects these relational vectors into the final sentence representation. This amplifies the critical signals of target-context relations for determining implicit hate. Experiments demonstrate that AmpleHate achieves state-of-the-art performance, outperforming contrastive learning baselines by an average of 82.14% and achieves faster convergence. Qualitative analyses further reveal that attention patterns produced by AmpleHate closely align with human judgement, underscoring its interpretability and robustness. Our code is publicly available at: https://github.com/leeyejin1231/AmpleHate

1 Introduction

Warning: *this paper contains content that may be offensive or upsetting*.

Internet and social media platforms have profoundly reshaped contemporary communication, enabling rapid information dissemination and interpersonal interactions. Yet, alongside the benefits, these platforms also enable the spread of hate speech, becoming a serious social issue. Recent studies indicate that approximately 30% of young individuals have experienced cyberbullying (Kansok-Dusche et al., 2023). This highlights the urgent need to detect and control hate speech.

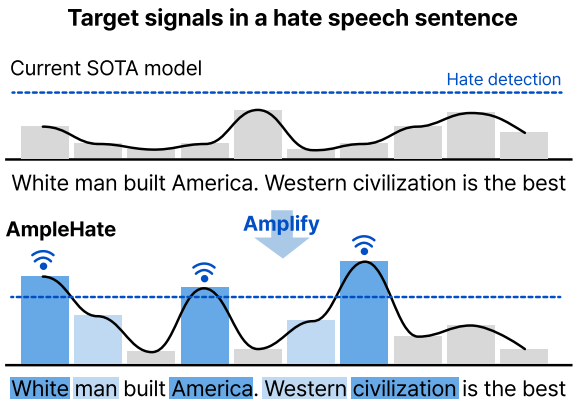

Figure 1: AmpleHate effectively detects implicit hate sentences by amplifying the target signals of hate-related tokens in the context of implicit hateness.

For explicit hate speech detection, traditional text classification approaches such that learn global sentence representations or keyword features are frequently used to distinguish hate from non-hate (Gao et al., 2017; Saleem et al., 2017). These approaches have been successful as explicit hate speech, with its direct and unambiguous characteristics, contains clear expressions such as offensive words or insults. Yet, implicit hate speech hides its hostility in latent context rather than explicit forms. Specifically, it conveys hate through context, sarcasm, indirect expressions, or culturally specific references (Wiegand et al., 2021; ElSherief et al., 2021). Consequently, standard keyword-based or holistic embedding methods often struggle to accurately detect implicit forms of hate.

*Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28851–28863
November 4-9, 2025

Recent studies have advanced implicit hate speech detection by employing contrastive learning frameworks, such as InfoNCE-based objectives (Oord et al., 2018; Kim et al., 2024). These methods lead semantically hateful sentences to have similar representations and push non-hateful sentences apart (Kim et al., 2022; Ahn et al., 2024). Although these techniques are effective in implicit hate speech detection, they operate on holistic sentence-level embeddings and overlooks the internal interactions within the sentences.

In contrast, humans identify implicit hate through a different and inherently more structured reasoning process: first identifying specific *targets* (e.g., sensitive entities, which are demographic groups such as immigrants or Muslims, LGBTQ+ individuals, or cultural referents) within a sentence and subsequently interpreting how the context frames or characterizes these targets to infer hateful intent. Motivated by this insight, we propose **AmpleHate**, a novel method explicitly designed to mirror the human inferential process of identifying implicit hate speech detection.

AmpleHate first identifies target entities (explicit targets) within sentences using a pre-trained Named Entity Recognition (NER) model. We then leverage the sentence-level embedding through the Transformer (Vaswani et al., 2017) encoder's [CLS] token as an implicit representation of global contextual information (implicit targets). AmpleHate computes attention-based relational interactions between explicit, implicit targets and sentence contexts. quantifying the influence each target has on the sentence's hateful meaning. These relational vectors are then *directly injected* into the final sentence embedding, amplifying target-context signals and steering the classifier to base its decisions on explicit target-context relationships, similar to human reasoning.

One might question the effectiveness of direct injection due to its susceptibility to potential noises. However, by limiting the injection specifically to the relationships between explicit, implicit targets and context, our method effectively enhances the learning of implicit hate speech. As a result, AmpleHate achieves optimal performance earlier compared to existing contrastive learning approaches.

Through extensive experiments on implicit hate speech benchmarks, AmpleHate achieves state-of-the-art performance, outperforming conventional contrastive learning baselines by 6.87%p in average. Additionally, Figure 1 and our qualitative analyses confirm that AmpleHate produces attention patterns closely aligned with human judgement behaviors, highlighting its interpretaiblity and robustness.

2 Related Work

2.1 Implicit Hate Speech

Implicit hate speech targets protected groups and is conveyed through subtle or indirect language rather than explicit slurs (GNET, 2024). Unlike explicit hate, which can often be caught via keyword matching (Warner and Hirschberg, 2012; Davidson et al., 2017) or simple toxicity filters (Waseem and Hovy, 2016; Nobata et al., 2016), implicit hate relies on metaphor or context-dependent cues that evade straightforward lexical methods. This poses a significant challenge for automated implicit hate speech detection. Recent works have primarily focused on two trajectories to address this challenge: (1) contrastive representation learning (Kim et al., 2023; Ahn et al., 2024; Kim et al., 2024); and (2) data sampling/selection strategies (Ocampo et al., 2023; Kim et al., 2025; Kim and Lee, 2025). Meanwhile, these approaches depend on the availability of corpora specifically annotated for implicit hate speech, such as Sap et al. (2020); Mathew et al. (2021) and ElSherief et al. (2021).

However, most traditional studies treat the protected target merely as a class label and thus fail to capture group-specific cues or coded language. This oversight motivates the next subsection, which surveys target-aware modeling techniques that explicitly integrate target information.

2.2 Target-Aware Modeling and Embeddings

Recent methods not only identify the target (Jafari et al., 2024) but also explicitly model target identity by using multi-task approach (Chiril et al., 2022). They jointly learn to detect hatefulness and to predict the specific target category or identity group, leveraging shared representations and affective lexical to transfer knowledge across topics and improve generalization. Adversarial training methods such as Xia et al. (2020) removed spurious group cues to strip dialectal markers from embeddings and reduce false positives. Chen et al. (2024) used a hypernetwork conditioned on target embeddings to generate filters that eliminate biased features. These approaches achieved significant gains in both accuracy and fairness. In addition to those target-aware methods, we draw inspiration from recent advances

in embedding-space control and attention dynamics. Han et al. (2024) applied a low-rank linear transform to embeddings to steer model behavior with minimal overhead and Barbero et al. (2025) showed that attention sinks prevent over-mixing in deep transformers. Motivated by their findings, we inject a lightweight, target-aware attention module that both steers the model toward group-specific signals and dedicates an extra attention head to capture subtle, implicit target cues.

3 AmpleHate: Amplify the Attention

We use the standard Transformer (Vaswani et al., 2017) self-attention mechanism in AmpleHate. The multi-head attention calculation enables the model to represent sentence-level context more effectively. AmpleHate consists of three core steps: 1) Target Identification, 2) Relation Computation, and 3) Direction Injection. Figure 2 illustrates the overall procedure of AmpleHate.

3.1 Target Identification

Let the input sentence X be tokenized as

$$X = [x_0, x_1, \ldots, x_n], \quad x_0 = [\text{CLS}],$$

and let a Transformer encoder produce contextual embeddings

$$H = [h_0, h_1, \ldots, h_n] \in \mathbb{R}^{(n+1)\times d},$$

where $h_i \in \mathbb{R}^d$, d is the size of hidden dimension, and h_0 is the [CLS] embedding. We begin by selecting the tokens that will serve as our *explicit targets*.

In the implicit hate speech dataset, the critical entities are mainly groups, organizations and local (Khurana et al., 2025). Thus, we apply a pre-trained NER tagger to each token and retain only those with labels in $\{\text{ORG}, \text{NORP}, \text{GPE}, \text{LOC}, \text{EVENT}\}$, yielding a binary mask $m_i \in 0, 1$ where $1 \leq i \leq n$. ORG represents organizations, NORP represents nationalities, religious, and political groups, GPE represents countries, cities, and states, LOC represents non-GPE locations such as mountain ranges or bodies of water, and EVENT represents named events such as wars, sports events, or disasters. Let

$$I_{exp} = \{i \mid m_i = 1\}.$$

The explicit target embeddings is then,

$$H_{exp} = [h_i]_{i \in I_{exp}} \in \mathbb{R}^{|I_{exp}|\times d}.$$

These embeddings H_{exp} serve as our explicit targets. On the other hand, we use the embedding h_0 of the [CLS] token as our *implicit token* H_{Imp}. This is supported by Barbero et al. (2025) showing that initial tokens such as $[CLS]$ or $< bos >$ naturally attract significant attention and serve as stable anchors in transformer models, helping to regulate information.

3.2 Relation Computation

Rather than simply boosting attention weights during training–which can fail to reflect the weights directly to the final decision–AmpleHate aims to capture the fine-grained influence of explicit targets H_{exp} and implicit targets H_{imp} on the context of the whole sentence X. We achieve this by applying a standard attention mechanism over the [CLS] embedding and the target embeddings, producing a relational vector $r \in \mathbb{R}^d$ that highlights the most relevant target-context interactions.

First, we form our attention inputs by letting the [CLS] embedding h_0 act as both the query and the value, and the target embeddings serve as keys:

$$Q = h_0, \; K = H_{tgt}, \; V = h_0,$$

where H_{tgt} is either H_{exp} or H_{imp}. We then compute a score for each target $h_t \in H_{tgt}$ using scaled dot-product attention:

$$r_{tgt} = \sum_{k \in H_{tgt}} \text{Softmax}(\frac{Qk^T}{\sqrt{d}})V.$$

This dynamically assigns higher weights to targets whose embeddings align more closely with the sentence-level context. As we compute when $H_{tgt} = H_{exp}$ and $H_{tgt} = H_{imp}$,

$$r = r_{exp} + r_{imp}.$$

Note that when we compute the attention for an implicit target, the equation is the same as computing the self-attention of the [CLS] embedding.

The relational vector r thus encodes how each explicit and implicit target interacts with the sentence as a whole by incorporating a learned self-adjustment of the global context of the sentence.

3.3 Direct Injection

In order to explicitly amplify and reflect the target-context relation we have computed in Section 3.2, we inject the relation vector r directly into the output, [CLS] embedding. This *direct injection* makes

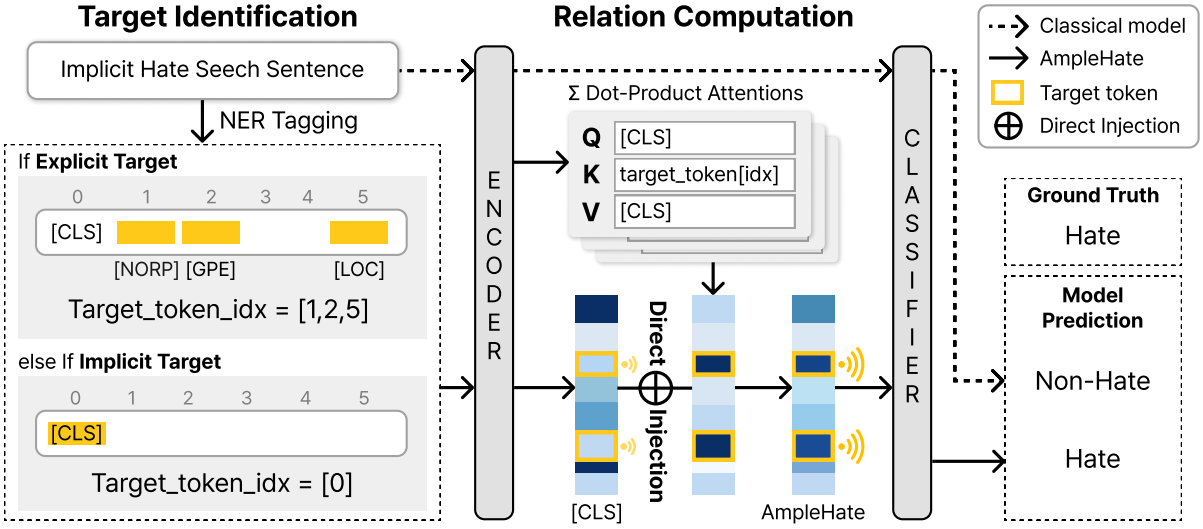

Figure 2: Overall procedure of AmpleHate.

sure that the output contains the signals of target relations. The updated output z is then passed to the classification head, ensuring that z guides the right prediction $\hat{y}$:

$$z = h_0 + \lambda \cdot r, \quad \hat{y} = \text{Softmax}(z),$$

where λ controls the degree of amplification for the injected signal.

This direct injection appropriately amplifies the signals of target-context relations for implicit hate speech detection, yielding two key benefits:

Stronger decision ques. By injecting r into h_0 right before classification, AmpleHate amplifies target-context interactions directly in the final logits. This ensures that the AmpleHate's predictions rely on these amplified cues rather than only on diffuse, indirect attention patterns.

Selective noise reduction. Since r is built only from identified targets, injection amplifies meaningful relations while filtering out irrelevant signals. This effectively reduces potential noises and guarantees the faster convergence.

3.4 Implementation Details

We identify the explicit targets by tagging named entities. Specifically, we apply the pre-trained NER taggers 'dbmdz/bert-large-cased-finetuned-conll03-english' and 'dbmdz/bert-base-cased-finetuned-conll03-english'. Also, the degree λ of amplification in Section 3.3 is in $\{0.5, 0.75, 1.0, 1.25, 1.5\}$.

4 Experiments and Analysis

4.1 Datasets

IHC (ElSherief et al., 2021) is designed to support implicit hate speech classification, featuring about 22k tweets annotated not only with hate labels but also with textual descriptions that reveal underlying hateful implications.

SBIC (Sap et al., 2020) offers a large-scale resource for studying social bias in language, pairing social media posts with structured frames that annotate offensiveness, speaker intent, and group targets.

Dynahate (Vidgen et al., 2021) introduced a challenging hate speech dataset built via a human-and-model-in-the-loop process, incorporating perturbed and implicitly hateful examples to enhance model generalization and robustness.

Hateval (Basile et al., 2019) contains about 19k Twitter posts annotated for binary hate speech specifically targeting immigrants or women, with additional tags for aggressiveness and target scope.

Toxigen (Hartvigsen et al., 2022) comprises approximately 274k machine-generated statements, evenly split between toxic and non-toxic language concerning 13 minority groups.

White (de Gibert et al., 2018) is a hate speech dataset collected from a White Supremacy Forum. Each sentence is manually annotated for the presence or absence of hate speech according to specific annotation guidelines.

Models	Datasets							Average
	IHC	SBIC	DYNA	Hateval	Toxigen	White	Ethos	
BERT	77.70	83.80	78.80	81.11	90.06	44.78	70.67	75.27
SharedCon	78.50	**84.30**	79.10	80.24	91.21	46.15	69.05	75.50
LAHN	78.40	83.98	79.64	80.42	90.42	47.85	75.26	76.56
AmpleHate (*bert-base-ner*)	81.46	83.79	81.39	80.56	90.74	71.96	**78.61**	81.21
AmpleHate (*bert-large-ner*)	**81.94**	84.03	**81.51**	**82.07**	**93.21**	**75.17**	77.06	**82.14**

Table 1: Experimental results of baselines and ours for each dataset. *bert-base-ner* means using a fine-tuned bert-base model, while *bert-large-ner* means using a fine-tuned bert-large model for NER tagging. The performance score (macro-F1) is the average of three runs with different random seeds.

Models	Test Datasets							Average
	IHC	SBIC	DYNA	Hateval	Toxigen	White	Ethos	
BERT	78.53	80.78	**81.06**	85.73	92.61	58.13	67.92	77.82
SharedCon	78.98	80.77	79.31	86.81	93.33	55.91	69.23	77.76
LAHN	77.40	80.34	79.64	84.49	90.48	59.09	68.34	77.11
AmpleHate (*ours*)	**86.14**	**82.03**	79.92	**87.17**	**97.04**	**71.41**	**71.30**	**82.18**

Table 2: Experimental results of baselines and ours for the whole dataset together as 'combined' in Table 5. We combine all datasets and treat them as a single dataset to see the general performance. We use *bert-large-ner* model for NER tagging. The performance score (macro-F1) is the average of three runs with different random seeds for the combined dataset.

ETHOS (Mollas et al., 2022) offers 998 YouTube/Reddit comments labeled for hate speech presence or absence. Table 5 in Appendix A reports the data statistics.

4.2 Baselines

BERT (Devlin et al., 2019) is a transformer-based language model pre-trained on large corpora using masked language modeling and next sentence prediction. Its bidirectional encoding enables effective contextual understanding for downstream classification tasks.

SharedCon (Ahn et al., 2024) is a SOTA method for the implicit hate speech detection task. The approach clusters sentence embeddings by label, uses each cluster centroid as an anchor, and then fine-tunes the model with a supervised contrastive loss.

LAHN (Kim et al., 2024) integrates momentum contrastive learning with label-aware hard-negative sampling, selecting the most similar opposite-label sampling for each anchor and training with a joint supervised-contrastive and cross-entropy loss.

4.3 Experimental Setup

We use the BERT-base model (Devlin et al., 2019) with the 100M parameter size as our encoder and train on an NVIDIA RTX 4090 GPU. The decision threshold ranges from 0.05 to 0.95 in steps of 0.05. We evaluate model performance using the macro-F1 and report averaged scores over three independent runs with different random seeds. We use the AdamW optimizer with a fixed learning rate of 2e-5, a batch size of 16, and 6 training epochs.

4.4 Experimental Results

We begin with experiments on individual datasets, where each model is trained and evaluated on the same dataset. We then train a single model on the combined dataset and evaluate it across individual test sets, demonstrating the versatility of our method.

Individual Dataset. Table 1 shows that AmpleHate consistently outperforms nearly all baseline models, with particularly strong gains on datasets with implicit hate (e.g., +27.32%p on White). On average, it improves over BERT by 6.87%p and over the strongest baseline (LAHN) by 5.58%p.

These results confirm that identifying explicit and implicit targets, coupled with amplification of target-context relation signals by direct injection, yields clear and consistent improvements over contrastive learning approaches.

Combined Dataset. In this setup, the model is trained on the combined set of all seven datasets

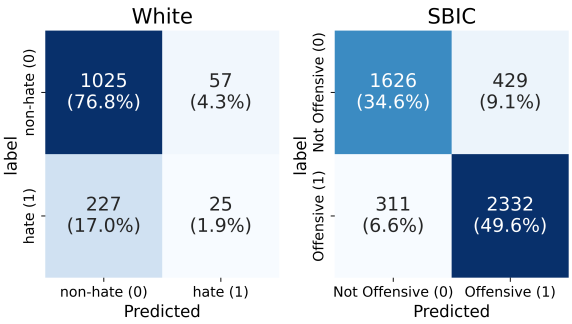

Figure 3: Confusion matrices for the White (left) and SBIC (right) datasets, showing counts and percentages of true vs. predicted labels for 'non-hate/not-offensive' and 'hate/offensive' classes.

and then evaluated independently on each test set.

Table 2 shows that AmpleHate achieves the highest macro-F1 on six out of seven datasets, gaining approximately 5%p gains on average for all baselines.

These findings indicate that AmpleHate maintains its effectiveness even without dataset-specific fine-tuning, making it suitable for practical deployment. By incorporating target-focused attention during training, the model retains robustness across varying expressions of implicit hate speech.

4.5 Case Analysis

Analysis on White Dataset Characteristics In the individual dataset experiments, AmpleHate achieves the highest relative improvement on the White dataset, which motivates a detailed examination of its linguistic and structural characteristics. Figure 3 shows the confusion matrix on White (left), where AmpleHate demonstrates strong discrimination between hate and non-hate instances, despite the skewed label distribution. The model maintains high precision on non-hate samples while successfully identifying a portion of subtle hate cases.

Several factors likely contribute to this result. First, White contains a high proportion of implicit hate speech, where hateful intent emerges through context rather than direct slurs. Many examples are structured in a way that requires identifying the target entity before the hatefulness becomes clear. This aligns well with the intended role of AmpleHate–to highlight target mentions and capture context-dependent signals.

Second, explicit references to target groups frequently appear in the data. While standard models may treat these as ordinary nouns, AmpleHate interprets them as meaningful cues and allocates attention accordingly.

These properties make the White dataset particularly well-matched to the goal behind AmpleHate, which focuses on detecting subtle and indirect hate speech through target-aware contextual understanding.

Error Analysis on SBIC In Table 1, AmpleHate shows only a marginal improvement over the baseline on the SBIC dataset. We analyze the SBIC dataset to understand this experimental result.

Hate speech datasets usually involve explicit attacks against specific identity groups. In contrast, offensive language datasets label content as offensive regardless of whether the language targets a group. SBIC adopts this broader definition, which creates a fundamental distinction from the other datasets in the study.

Figure 3 shows the confusion matrix for SBIC (right). We extract 740 misclassified examples and analyze their content. Most errors fall into the false positive category. Many of these cases contain highly offensive expressions, but the original label marks them as non-offensive. For example, the sentence "in fucking mexico bitches" includes profanity and a strong negative tone, but it carries a non-offensive label.

We randomly sample 100 of the misclassified examples and review them manually. Among them, 51% appear to contain labeling inconsistencies. This high noise ratio suggests that SBIC includes annotation artifacts that differ from hate speech corpora. These inconsistencies likely contribute to the limited performance gains of AmpleHate on SBIC.

5 Analysis

5.1 Impact of Targets

AmpleHate distinguishes between random targets, explicit targets, identified via NER tagging, and implicit targets represented by the [CLS] token, which captures global context. We conduct an ablation study to understand the contribution of each target type. The experiment includes fine-tuning the vanilla BERT baseline without any target signals, with randomly assigned 20% targets, with only the implicit targets, with only explicit targets, and with both implicit and explicit targets combined. Since explicit targets were tagged for approximately 20% of the data on average, we introduced the Random 20% condition to serve as a comparable baseline.

Target	Datasets							Avg. (F1)
	IHC F1 (step)	SBIC F1 (step)	DYNA F1 (step)	Hateval F1 (step)	Toxigen F1 (step)	White F1 (step)	Ethos F1 (step)	
No Target	77.70 (1146)	83.80 (2016)	78.80 (2316)	81.11 (1953)	90.06 (1147)	44.78 (**360**)	70.67 (160)	75.27
Random Target	77.00 (**450**)	83.95 (**750**)	79.99 (1950)	79.44 (400)	89.59 (**300**)	59.46 (500)	67.54 (**50**)	76.71
Implicit Target	78.76 (1200)	83.63 (2167)	80.54 (**647**)	81.70 (2120)	90.52 (647)	59.00 (413)	75.57 (160)	78.53
Explicit Target	77.92 (910)	83.93 (2150)	80.98 (907)	81.33 (1827)	90.55 (907)	59.69 (387)	72.33 (120)	78.10
Implicit + Explicit	**81.93** (1246)	**84.03** (2166)	**81.51** (2350)	**82.07** (**1647**)	**93.21** (933)	**75.16** (380)	**77.06** (127)	**82.13**

Table 3: Ablation study results showing macro-F1 scores and stopping steps for five target configurations (No Target, Random Target, Implicit Target, Explicit Target, Implicit + Explicit) across seven datasets. We use *bert-large-ner* tagger to explicit tagging. Stopping epochs are averaged over three runs and then rounded to the nearest integer.

Dataset	Highlighted Implicit Hate Speech Sentence
IHC	even the white nationalist movement is ran by jews.
SBIC	Yo hoe is like a tick so she hunting for a blood
DYNA	tell chicks to stop dressing like street walkers
Hateval	Arrest them and DEPORT THEM ALL!!!! # BuildThatWall
Toxigen	all black people are suspected of trying to pull something
White	The German men sound so sexy.
Ethos	Too bad women don't know how to kill themselves

Table 4: Visualization of token-level attention weights derived from the [CLS] token for each dataset example. Highlighted tokens are those that the AmpleHate model considers more important than a standard BERT baseline, in the implicit-target scenario where no explicit target token is provided.

Table 3 presents both the macro F1-scores and the number of convergence steps across these settings.

Interestingly, our results reveal that incorporating only the implicit target yields greater improvements than relying solely on explicit targets. In particular, using the implicit targets shows 3.26%p improvements while using the explicit targets shows 2.83%p on average, compared to fine-tuning BERT with no signals. This finding suggests that, in the case of implicit hate speech–where the intent is often in latent form and not tied to named entities–contextual representation via the [CLS] token is particularly powerful for target-context relation computation. Nevertheless, models leveraging explicit targets also outperform the baseline without any target signals, confirming that entity-specific cues add valuable information for detecting implicit hate. Crucially, when both implicit and explicit targets are integrated, AmpleHate achieves the highest performance across all settings.

In addition to the macro-F1 performance, we observe the training efficiency. Including target signals generally achieves faster convergence steps than fine-tuning vanilla BERT with no signals. On average, using only implicit or only explicit targets reduces the number of steps for the convergence by approximately 20%. However, when both target types are used together, the reduction in convergence steps is more modest with approximately 3% compared to the vanilla BERT. This small tradeoff is caused by a substantial gain in detection performance, achieving 6.86%p improvement.

These results suggest that while incorporating implicit or explicit targets individually leads to faster convergence but slight performance gains, combining both enables AmpleHate to achieve the best performance along with the competitive convergence speed.

5.2 Comparison of Convergence Step

We compare the convergence steps of AmpleHate with the state-of-the-art model LAHN to assess the training efficiency and optimization stability of our approach. Figure 4 presents the comparison on the Ethos dataset. Figure 4 (a) shows the validation F1-score each training step, where the star markers indicate the step at which each model achieves its highest validation performance. AmpleHate reaches its peak F1-score at an earlier step than LAHN, achieving comparable or better perfor-

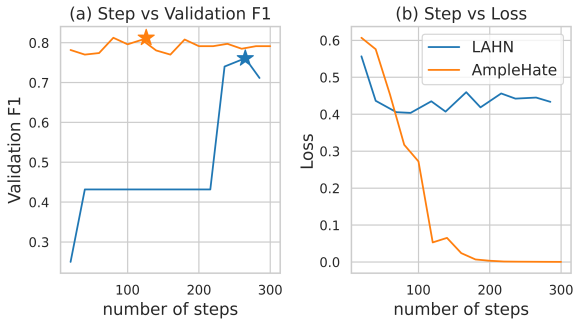

Figure 4: Comparison of AmpleHate and LAHN on the Ethos datasets. (a) Validation F1-score over training steps, with start indicating the stopping step. (b) Training loss over steps. AmpleHate converges faster and maintains lower loss than LAHN. Appendix B provides further details.

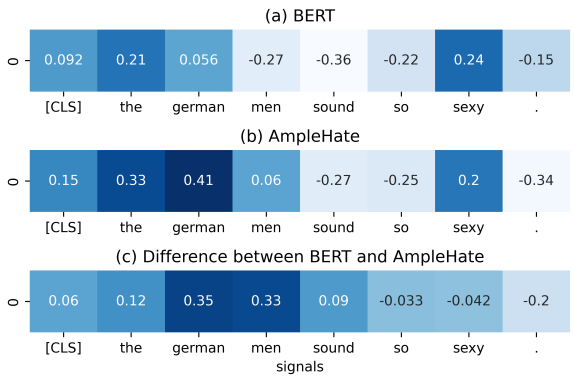

Figure 5: Token-level target signals between the [CLS] representation and each token in an implicit hate speech sentence from the White dataset. (a) shows BERT's signals, (b) shows signals from AmpleHate trained on combined datasets and (c) highlights the token-level difference.

mance with significantly fewer training steps. Figure 4 (b) represents the loss curves over the same training steps, where AmpleHate consistently maintains a lower loss throughout training. It demonstrates that AmpleHate is more stable and effective for optimization.

This early convergence and stable training behavior indicate that AmpleHate is more computationally efficient and less prone to overfitting compared to LAHN. Similar patterns appear across other datasets, with detailed results provided in Appendix B. These findings highlight the strength of AmpleHate in delivering high performance with rapid and stable convergence, making it a highly efficient solution for implicit hate speech detection.

5.3 Token-level Attention with Implicit Tokens

AmpleHate assigns the [CLS] token as a target token when no explicit target entity appears in a sentence. This strategy raises a question: does [CLS] effectively serve as an implicit focus point in such cases? To answer this question, we examine how strongly the [CLS] representation attends to each token in sentences labeled as implicit hate speech.

Figure 5 visualizes token-level target signals with [CLS] for three cases: (a) BERT, (b) AmpleHate, (c) the difference between them. We use AmpleHate trained on the combined_datasets setting described in Section 4.4. BERT represents attention broadly, without clear focus. In contrast, AmpleHate places greater emphasis on the token "germen". This token plays a central role in interpreting the hateful intent of the sentence. Even without target supervision, AmpleHate identifies "germen" as a core indicator of implicit hate, suggesting that its target-aware attention mechanism successfully captures subtle targets. This observation implies that computing token-level relation with [CLS] provides sufficient signal for detecting implicit hate, even in the absence of explicit target tokens.

Table 4 extends the analysis across all seven datasets. In each example, we highlight tokens where AmpleHate assigns significantly higher attention than BERT. These implicit targets often include stereotype triggers. The results demonstrate that AmpleHate consistently prioritizes tokens humans are also likely to attend to when interpreting implicit hate.

This analysis confirms that using [CLS] as the target token does not disrupt the model's ability. Instead, it enables AmpleHate to maintain robust attention over hate-relevant context, even without explicit target supervision.

6 Conclusion

We introduced AmpleHate, an approach that mirrors human reasoning for implicit hate detection by focusing on target-context interaction, in contrast to conventional contrastive learning approaches. AmpleHate first identifies explicit and implicit targets, then computes attention-based relation vectors, and finally injects the signals into the output embeddings. This targeted injection amplifies target-context relations, which specifically contain relevant signals for implicit hate, suppressing noise. AmpleHate achieves the state-of-the-art F1 scores on seven implicit hate speech datasets both in in-

dividual and combined dataset train settings. This empirically shows the generalizability of AmpleHate. Furthermore, AmpleHate presents faster convergence than the contrastive-learning baselines. While AmpleHate depends on an external NER tagger to locate explicit targets, we aim to explore richer target detection and relation injection strategies as well as applications to other tasks that hinge on subtle target-context reasoning. Overall, AmpleHate demonstrates that encoding and injecting target-context relations can outperform conventional contrastive learning methods, offering a more robust and efficient approach for implicit hate speech detection.

Limitations

AmpleHate relies on the accuracy of the external NER tagger for explicit target identification; errors or insufficient entity recognition may limit performance, especially for context-specific or newly emerging hate expressions with latent implications. In addition, the approach of AmpleHate on modeling implicit target centers on the [CLS] token, which may not always capture the full range of implicit cues, particularly in longer or more complex sentences.

Moreover, while AmpleHate demonstrates strong performance in both macro-F1 scores and the convergence rate, there remains the possibility that AmpleHate may not perform the best on forms of implicit hate which are not well represented in current benchmarks or in new domains.

We plan to address these limitations in future work by developing more adaptive and context-aware target strategies and extending our approach to tasks that involve more complex or subtle forms of latent hate.

Ethical Consideration

Minimizing Annotator Harm Conventional hate speech detection approaches rely on manual annotations, which can expose annotators to distressing or harmful sentences. AmpleHate reduces this exposure by leveraging implicit contextual signals, thus lowering the dependence on explicit and manual labels. This contributes to a more ethical data collection and annotation pipeline by minimizing the mental burden on human annotators (Vidgen et al., 2019).

Contextual Awareness AmpleHate models explicit and implicit target information, enabling it to recognize diverse and nuanced forms of hate speech. By capturing latent patterns through its design, AmpleHate supports more context-sensitivity and fair detection, reducing the risk of overfitting to majority groups or missing contextually-dependent harms.

Risks of Potential Misuse AmpleHate could be misused–such as by adversarial users attempting to bypass detection or abuse the method to generate more sophisticated hate speech. The deployment of AmpleHate should be accompanied by careful monitoring and critical assessment of model outputs to address these risks.

Acknowledgments

This research was supported by the NRF grant (RS-2025-0222262) and the AI Graduate School Program (RS-2020-II201361) funded by the Korean government (MSIT).

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. Sharedcon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics, ACL*, pages 10444–10455.

Jean-Thomas Baillargeon and Luc Lamontagne. 2024. SMARTR: A framework for early detection using survival analysis of longitudinal texts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 36–41.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Razvan Pascanu, and 1 others. 2025. Why do llms attend to the first token? *CoRR*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *SemEval@NAACL-HLT*, pages 54–63.

Yves Bestgen. 2022. Please, don't forget the difference and the confidence interval when seeking for the state-of-the-art status. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC*, pages 5956–5962.

Tong Chen, Danny Wang, Xurong Liang, Marten Risius, Gianluca Demartini, and Hongzhi Yin. 2024. Hate speech detection with generalizable target-aware fairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD*, pages 365–375.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, 14(1):322–352.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515.

Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, NAACL-HLT*, pages 4171–4186.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 345–363.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *IJCNLP(1)*, pages 774–782. Asian Federation of Natural Language Processing.

Serge Gladkoff, Lifeng Han, and Goran Nenadic. 2023. Student's t-distribution: On measuring the inter-rater reliability when the observations are scarce. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, RANLP*, pages 419–428.

GNET. 2024. "substitution": Extremists' new form of implicit hate speech to avoid detection. Accessed: 2024-06-24.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek F. Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. 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 (Volume 1: Long Papers), ACL*, pages 3309–3326.

Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. Target span detection for implicit harmful content. In *ICTIR*, pages 117–122.

Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, Anke Zeißig, Lisanne Seemann-Herz, Sebastian Wachs, and Ludwig Bilz. 2023. A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, violence, & abuse*.

Katherine A. Keith and Brendan O'Connor. 2018. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 4575–4585.

Urja Khurana, Eric T. Nalisnick, and Antske Fokkens. 2025. Defverify: Do hate speech models reflect their dataset's definition? In *Proceedings of the 31st International Conference on Computational Linguistics, COLING*, pages 4341–4358.

Do Kyung Kim, Hyeseon Ahn, Youngwook Kim, and Yo-Sub Han. 2025. Analyzing offensive language dataset insights from training dynamics and human agreement level. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING*, pages 9780–9792.

Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware hard negative sampling strategies with momentum contrastive learning for implicit hate speech detection. In *Findings of the Association for Computational Linguistics, ACL*, pages 16177–16188.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING*, pages 6667–6679.

Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. Conprompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In *EMNLP (Findings)*, pages 10964–10980.

Yumin Kim and Hwanhee Lee. 2025. Selective demonstration retrieval for improved implicit hate speech detection. In *arXiv preprint arXiv:2504.12082*.

Wei Li, Zhen Huang, Xinmei Tian, Le Lu, Houqiang Li, Xu Shen, and Jieping Ye. 2024. Interpretable composition attribution enhancement for visio-linguistic compositional understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 14616–14632.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI*, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Griorios Tsoumakas. 2022. Ehos: a multi-labeled hate speech detection dataset. In *Complex & Intelligent Systems*.

Chikashi Nobata, Joel R. Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *WWW*, pages 145–153.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics, ACL*, pages 2758–2772.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *CoRR*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 5477–5490.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 1667–1682.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language - what does it actually look like and why are we not getting there? In *NAACL-HLT*, pages 576–587.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *SocialNLP@ACL*, pages 7–14.

A Data Statistic

We evaluate AmpleHate on seven publicly available hate speech datasets covering both explicit and implicit targets. Table 5 summarizes the number of examples in each train, validation, and test split. Note that the combined corpus's test set (marked with †) is held out from our main evaluations.

Dataset	Train set	Valid set	Test set
IHC	14,932	1,867	1,867
SBIC	35,504	4,673	4,698
DYNA	33,004	4,125	4,126
Hateval	10,384	1,298	1,298
Toxigen	5,420	678	678
White	10,668	1,334	1,334
Ethos	798	100	100
Combined	110,710	14,075	14,101†

Table 5: The statistical information of datasets in our experiments. The test split of the combined dataset (denoted with †) is excluded from AmpleHate evaluations.

B Convergence Efficiency

We compare the convergence behavior of AmpleHate and LAHN, a state-of-the-art contrastive learning baseline across all datasets, presented in Table 6. Here, convergence step refers to the validation step at which early stopping selects the best model. Lower convergence steps indicate faster model training and more efficient progression toward optimal performance.

Acrross six out of seven benchmarks, AmpleHate consistently reaches peak performance with substantially lower training steps than LAHN. For instance, AmpleHate converges more than three times faster on the White dataset, and achieves approximately $\times$ 1.5 speedup for the other datasets. Notably, AmpleHate not only converges faster but also achieves higher macro-F1 scores than LAHN reported in Table 1, demonstrating that its efficiency does not degrade its performance. The only exception is Hateval, where LAHN converges marginally faster. However, as shown in Table 1 AmpleHate outperforms LAHN by approximately 2%p in macro-F1, confirming that this tradeoff is reasonable.

The primary factor of AmpleHate's rapid convergence is its direct injection of target-context relational signals. By amplifying the most relevant cues of implicit hate by using targets, AmpleHate enables the model to focus on the critical features earlier in training, reducing unnecessary parameter updated and accelerating optimization.

Overall, these findings highlight that AmpleHate is both accurate as well as highly efficient, requiring fewer training steps and converging more rapidly than contrastive learning methods. This strengthens AmpleHate's suitability for implicit hate speech detection.

C Statistical Significance Analysis

We evaluate the performance of AmpleHate across three random seeds on eight implicit hate speech detection datasets. Table 7 reports the results including the average and standard deviation for each dataset. We compute a 95% confidence interval (CI) using the t-distribution which is widely used to analyze statistical significance (Keith and O'Connor, 2018; Bestgen, 2022; Gladkoff et al., 2023; Baillargeon and Lamontagne, 2024; Li et al., 2024). Since we conducted 3 runs in total, our degrees of freedom (df) is 2 and the interval is calculated as follows:

$$\text{CI} = \bar{x} \pm t_{(\frac{\alpha}{2}, df)} \cdot \frac{s}{\sqrt{n}} = \bar{x} \pm t_{(0.025,2)} \cdot \frac{s}{\sqrt{n}},$$

where $\bar{x}$ is the average score, s is the standard deviation, n is the number of runs, and $t_{0.025,2} \approx 4.303$ is the t-statistics for 95% confidence with two degrees of freedom.

The results demonstrate that AmpleHate achieves highly consistent performance across multiple runs. For instance, the CI for SBIC and IHC is [83.71, 84.35] and [81.10, 82.78], respectively. These findings underscore the reliability and robustness of AmpleHate.

Models	Datasets						
	IHC	SBIC	DYNA	Hateval	Toxigen	White	Ethos
LAHN	1683	3969	2962	1358	1592	1286	256
AmpleHate *ours*	980	2250	2100	1646	933	380	126
$\text{Speedup} = \frac{\text{LAHN steps}}{\text{AmpleHate steps}}$	$\times 1.72$	$\times 1.76$	$\times 1.41$	$\times 0.83$	$\times 1.71$	$\times 3.38$	$\times 2.03$

Table 6: Convergence results of AmpleHate and LAHN. For each dataset, we report the early stopping step at which the best validation performance is achieved, as well as the relative speedup ($\times$) of AmpleHate over LAHN. A higher speedup indicates that AmpleHate converges in fewer steps, improving the training efficiency.

Dataset	Run 1	Run 2	Run 3	AVG. $\pm$ STD.	CI
IHC	82.31	81.63	81.87	81.94 ± 0.34	[81.10, 82.78]
SBIC	84.05	84.15	83.90	84.03 ± 0.13	[83.71, 84.35]
DYNA	81.44	81.77	81.32	81.51 ± 0.23	[80.94, 82.08]
Hateval	82.14	82.02	82.04	82.07 ± 0.06	[81.92, 82.22]
Toxigen	93.83	92.60	93.21	93.21 ± 0.62	[91.67, 94.75]
White	74.62	75.88	75.00	75.17 ± 0.65	[73.56, 76.78]
Ethos	79.14	77.29	74.76	77.06 ± 2.20	[71.59, 82.53]
AVG.	82.50	82.19	81.73	82.14 ± 0.39	[81.17, 83.11]

Table 7: Experimental results of AmpleHate for each dataset. Each row reports the results for a given dataset, including all three runs, the average (AVG.), standard deviation (STD.), and the 95% confidence interval (CI) using the t-distribution.