Aligning Implied Statements for Implicit Hate Speech Generalizability with Context-Bounded Semi-hard Negative Mining

Anonymous ACL submission

Abstract

Classifying implicit hate speech remains a challenge, intent is often masked through insinuation and context rather than explicit slurs. Prior supervised contrastive approaches improve in-domain detection but can overfit surface cues and struggle to transfer across datasets. We propose IMPSH, a triplet-based framework that aligns posts with implied statements when available and uses context-bounded semi-hard negatives to focus learning on near confusions. We also examine AUGSH, which forms positives via data augmentation. In controlled evaluations on IHC, SBIC, and DYNAHATE with BERT and HATEBERT, IMPSH is a viable alternative to standard supervised contrastive baselines and often improves cross-domain performance under matched preprocessing and tuning budgets. Representation analysis using alignment and uniformity indicates tighter positive pairs with balanced global spread, and qualitative nearest-neighbor case studies illustrate typical false negatives under domain shift. These results demonstrate that aligning posts with their implied statements via context-bounded mining provides a more stable, bijective-like mapping to related insinuations, overcoming the volatility inherent in traditional clustering-based representation learning.[1]

Content Warning

The content of this paper may contain offensive, harmful, or distressing language, including examples of hate speech and discriminatory expressions. These materials are included solely for research purposes and do not reflect the views of the authors. Reader discretion is advised.

1 Introduction

Classifying implicit hate remains challenging because hateful intent is often expressed indirectly through sarcasm, euphemisms, rhetorical questions, and other pragmatic cues, so surface text features are often insufficient (ElSherief et al., 2021; Sap et al., 2020; Kim et al., 2022; Zhang et al., 2024). A second challenge is semantic overlap with non-hateful content: posts that target the same group can share topical content despite different labels, which blurs the decision boundary and weakens supervision (Sap et al., 2020; ElSherief et al., 2021). When this boundary is weak, standard training can learn dataset-specific shortcuts tied to topic and group-linked surface cues, reducing cross-dataset generalization and transfer to related abusive-language settings (Nejadgholi and Kiritchenko, 2020; Röttger et al., 2022; Sap et al., 2019).

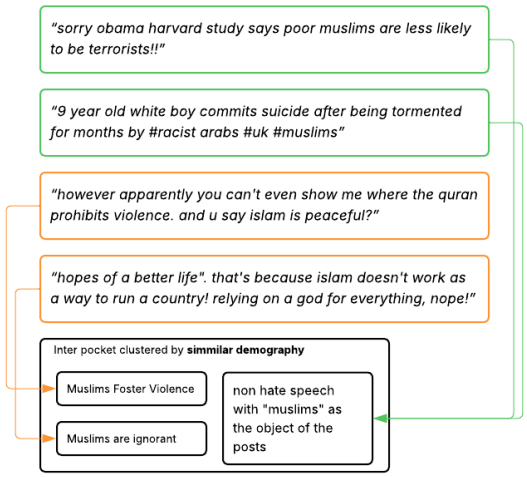

Figure 1: Posts that share similar implied targets form tight yet fragmented clusters; small wording changes disperse points even within the same demographic group (ElSherief et al., 2021).

Figure 1 illustrates this issue for examples aligned to the same target demographic. Despite different labels, many instances share topic and group mentions and can lie close in representa-

[1] Code will be released under the MIT license upon acceptance.

tion space, so accurate detection requires inferring intent and implied meaning, not just matching surface wording. To address semantic overlap, prior work uses supervised contrastive learning (SCL), most notably IMPCON. IMPCON pulls each post toward its human-annotated implied statement as a positive pair (Khosla et al., 2020; Gunel et al., 2021; Kim et al., 2022). However, standard SCL treats most other in-batch examples as negatives; in implicit hate datasets, many near-neighbors differ only slightly, so repelling them can introduce false negatives and hurt generalization (Huynh et al., 2020; Kalantidis et al., 2020; Wang et al., 2019). Ahn et al. (2024) propose SHAREDCON, which forms positives from shared semantics among same-label posts, reducing reliance on implied-statement annotations (Ahn et al., 2024), yet it still inherits the same fragile negative treatment when many examples are topically similar.

Motivated by this remaining fragility on the negative side, we instead target *negative selection*. Through the lens of alignment and uniformity (Wang and Isola, 2020), SCL-style training that aligns positives while repelling all other in-batch instances as negatives (Kim et al., 2022; Ahn et al., 2024) can yield tight, topic-driven clusters and weaker global coverage of the embedding space, which can hurt transfer. This effect is amplified by false negatives and class collisions that distort local neighborhoods (Chuang et al., 2020). To mitigate clustering driven purely by shared topics and target mentions, we explicitly separate semantically close instances with opposing labels, using a margin-based triplet objective with semi-hard negative mining (Schroff et al., 2015; Hermans et al., 2017; Wu et al., 2017; Musgrave et al., 2020; Xuan et al., 2020; Robinson et al., 2021).

In summary, we make the following contributions:

- We propose a triplet-based framework for implicit hate detection that uses *context-bounded semi-hard negative mining* to avoid repelling all in-batch negatives, while keeping a standard cross-entropy objective for classification.
- We introduce two variants, IMPSH and AUGSH. IMPSH uses post-implied-statement positives when available, while AUGSH uses augmentation-based positives for all instances to isolate the role of implication.
- We evaluate on IHC, SBIC, and DYNAHATE with BERT and HATEBERT under matched tokenization and tuning budgets, and analyze representation structure with alignment and uniformity scores alongside qualitative neighbor and embedding visualizations.

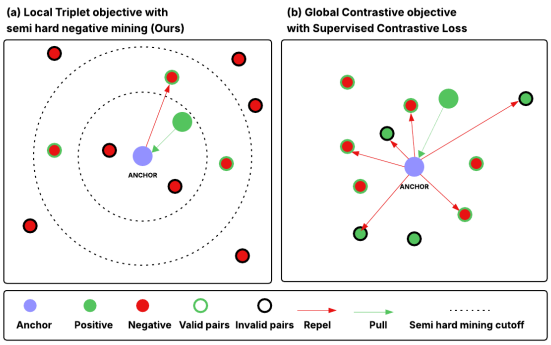

Figure 2: Triplet vs. SCL. Triplet updates use a margin-violating, confusable negative; SCL repels all non-positives in the batch.

2 Related Work

Early hate speech detection relied on lexical cues (Waseem and Hovy, 2016; Davidson et al., 2017), which often fail for implicit hate expressed via sarcasm, euphemisms, and other pragmatic cues (Badjatiya et al., 2017; Golbeck et al., 2017). More recent benchmarks such as IHC and SBIC, and cross-domain evaluation settings like DYNAHATE, shifted attention to semantic modeling, but robust cross-dataset generalization remains difficult (Sap et al., 2020; ElSherief et al., 2021; Vidgen et al., 2021; Kim et al., 2022; Ramponi and Tonelli, 2022).

To improve generalization, recent work applies supervised contrastive learning (SCL) to implicit hate. Kim et al. (2022) propose IMPCON, which uses the human-annotated implied statement as a positive and pulls post-implication pairs together. Ahn et al. (2024) propose SHAREDCON, which removes reliance on implied statements by mining shared semantics among same-label instances (e.g., clustering) and constructing positives from label-consistent neighborhoods.

2.1 Triplet Objectives and Negative Selection

Triplet losses optimize relative comparisons by pulling an anchor toward a positive and pushing it away from a negative by a margin (Schroff et al., 2015; Hermans et al., 2017). In contrast, SCL pulls each anchor toward all same-label examples in the

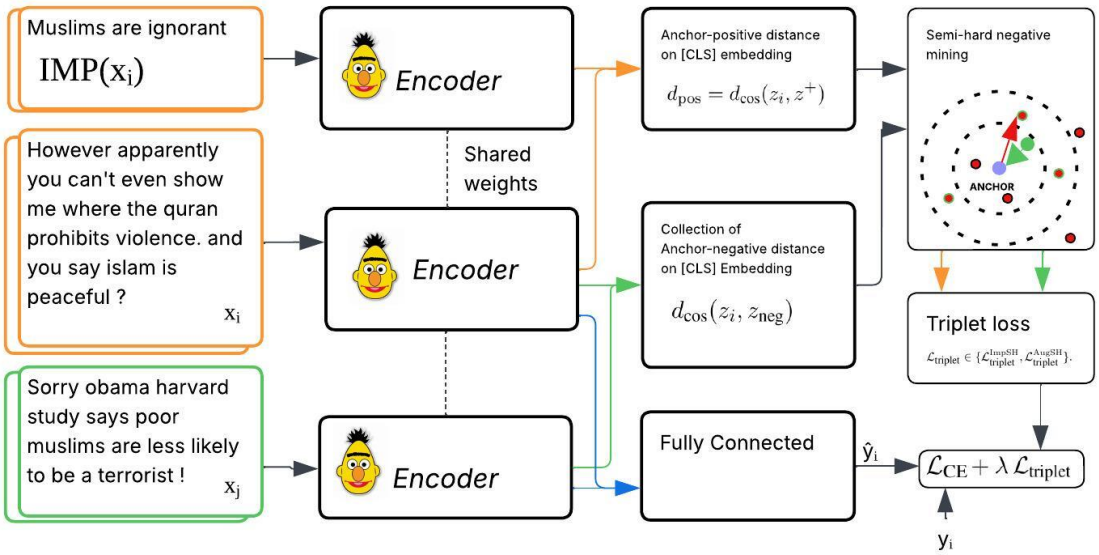

Figure 3: Training overview for our triplet framework. Left: positive formation (ImpSH uses a post–implication pair; AugSH uses augmentation-based positives when implication is absent). Right: semi-hard negative mining from the minibatch relative to the chosen anchor–positive pair. The encoder is updated by the triplet loss $\mathcal{L}^{\text{ImpSH}}_{\text{triplet}}$ or $\mathcal{L}^{\text{AugSH}}_{\text{triplet}}$ (Eqs. 5 and 6), alongside a standard classification loss $\mathcal{L}_{\text{CE}}$.

minibatch and pushes it away from different-label examples (Khosla et al., 2020; Gunel et al., 2021; Liao, 2021). In datasets with heavy topical overlap, treating most in-batch items as negatives can create false negatives and distort local neighborhoods, which can hurt transfer (Huynh et al., 2020; Kalantidis et al., 2020; Wang et al., 2019; Wang and Isola, 2020).

Negative selection is therefore critical. Semi-hard mining focuses on negatives that violate the margin but are not extreme outliers, concentrating updates on realistic near-misses (Schroff et al., 2015; Wu et al., 2017; Musgrave et al., 2020; Xuan et al., 2020; Robinson et al., 2021). For implicit hate, where examples can share topic and target cues while differing in implied intent, context-bounded semi-hard selection helps the model separate intent-sensitive cases without over-repelling broadly similar content.

3 Methodology

Figure 3 summarizes training. We jointly optimize a standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ and a triplet loss, mining semi-hard negatives within each minibatch (Eqs. 5-6). For each anchor-positive pair, we select an opposite-label negative that is farther than the positive but still within the margin band, focusing updates on near-confusable cases.

The standard cross-entropy loss is:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]. \quad (1)$$

We use cosine distance:

$$d_{\cos}(z_i, z_j) = 1 - \frac{z_i \cdot z_j}{\|z_i\|\,\|z_j\|}. \quad (2)$$

We define one positive per anchor. For hateful x_i, $z_i^+ = h(\text{IMP}(x_i))$; for non-hate x_i, $z_i^+ = h(\text{AUG}(x_i))$, where $h(\cdot)$ is the encoder, $\text{IMP}(\cdot)$ returns the implied statement, and $\text{AUG}(\cdot)$ applies synonym augmentation. The anchor-positive distance is:

$$d_{\text{pos}} = d_{\cos}(z_i, z_i^+). \quad (3)$$

To ensure the model separates negatives that are truly close to the positive, we select a context-bounded semi-hard negative from the opposite-label pool $\mathcal{N}$:

$$z_{\text{neg}} = \underset{\substack{z_j \in \mathcal{N} \\ d_{\text{pos}} < d_{\cos}(z_i, z_j) < d_{\text{pos}} + \alpha}}{\arg\min} d_{\cos}(z_i, z_j). \quad (4)$$

The triplet objectives specialize by the choice of z^+:

$$\mathcal{L}_{\text{triplet}}^{\text{ImpSH}} = \frac{1}{N}\sum_{i=1}^{N} \text{ReLU}\big[d_{\text{pos}} - d_{\cos}(z_i, z_{\text{neg}}) + \alpha\big] \quad \text{with } z^+ = h(\text{IMP}(x_i)) \tag{5}$$

$$\mathcal{L}_{\text{triplet}}^{\text{AugSH}} = \frac{1}{N}\sum_{i=1}^{N} \text{ReLU}\big[d_{\text{pos}} - d_{\cos}(z_i, z_{\text{neg}}) + \alpha\big] \quad \text{with } z^+ = h(\text{AUG}(x_i)) \tag{6}$$

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\, \mathcal{L}_{\text{triplet}}, \mathcal{L}_{\text{triplet}} \in \{\mathcal{L}_{\text{triplet}}^{\text{ImpSH}}, \mathcal{L}_{\text{triplet}}^{\text{AugSH}}\}. \tag{7}$$

For ImpSH 5 method we use implied statement if available, otherwise we use augmentation. And for AugSH we use augmentation for both label.

4 Experiments

4.1 Dataset

Following Kim et al. (2022); Ahn et al. (2024), we treat implicit hate detection as a binary classification task. We evaluate on three datasets shown in Table 1. IHC (ElSherief et al., 2021) contains tweets labeled for hate, with annotations for target groups and implied hateful meanings. SBIC (Sap et al., 2020) contains Reddit posts annotated for social bias and offensiveness, and it also provides free-text implications. DYNAHATE (Vidgen et al., 2021) is a dynamically generated hate speech dataset created through human-model-in-the-loop collection, with diverse and challenging hate examples. For SBIC and DYNAHATE, we map all abusive categories (explicit/implicit/offensive) to the *hate* label for binary evaluation.

Dataset	Total	Non-hate	Hate
IHC	18,666	13,206	5,460
SBIC	44,875	21,276	23,599
DYNAHATE	33,006	17,804	15,202

Table 1: Datasets used for training and cross-dataset evaluation, including class distribution. The *hate* label covers offensive, explicit, and implicit hate.

4.2 Implementation Details

We fine-tune BERT (Devlin et al., 2019) and HateBERT (Caselli et al., 2021) as sentence encoders. Following Kim et al. (2022), we fix the learning-rate-to-batch-size ratio, using a batch size of 8 and a learning rate of 2×10^{-5} with AdamW. Models are trained for 6 epochs with dropout 0.1 on RTX 3050 GPU (4GB). For the metric-learning objective, we tune $\lambda = 0.25$ and $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ using dev set performance. Experiments use 4 random seeds $\{0, 1, 2, 3\}$. Synonym substitution is applied via NLPAug[2] as in Kim et al. (2022).

4.3 Baselines

We compare against four objectives:

- **CE.** Standard fine-tuning with cross-entropy only.
- **CE + SCL** (Gunel et al., 2021). Cross-entropy plus supervised contrastive loss.
- **CE + ImpCon** (Kim et al., 2022). Cross-entropy plus implication-aware supervised contrastive loss.
- **CE + SharedCon** (Ahn et al., 2024). Cross-entropy plus shared-semantics supervised contrastive loss.

All baselines use the same backbone, preprocessing, and optimization setup as our method.

5 Results and Analysis

5.1 F1 Score result

Table 2 reports macro-F1 (%) averaged over four seeds described in Section 4.2 with both in-domain and cross-dataset evaluation. Since the goal of IMPSH is to handle semantic overlap during representation learning for generalization, we compare IMPSH primarily on cross-dataset transfer and we use AUGSH as an ablation that keeps augmentation based multi-view training but removes implied-statement supervision. Prior work finds that in-domain and out-of-distribution performance can be strongly correlated but can also become inversely related on real benchmarks, so selecting models only by in-domain performance can miss the best cross-dataset model (Miller et al., 2021; Teney et al., 2023). With BERT trained on IHC, IMPSH is best for transfer to DYNAHATE but it does not surpass the strongest baseline for transfer to SBIC and it also does not surpass the strongest baseline for in-domain evaluation on IHC, where SHAREDCON

[2] https://nlpaug.readthedocs.io/en/latest/augmenter/word/synonym.html

Model	Objective	IHC→SBIC	IHC→DYNAHATE	IHC→IHC
BERT	CE	56.9%	53.3%	77.7%
BERT	CE+SCL (Gunel et al., 2021)	59.8%	52.2%	77.7%
BERT	CE+ImpCon (Kim et al., 2022)	60.8%	57.9%	78.0%
BERT	CE+SharedCon (Ahn et al., 2024)	**65.2%**	59.1%	**78.4%**
BERT	CE+AugSH	59.2%	54.8%	77.8%
BERT	CE+ImpSH	61.4%	**59.2%**	78.3%
HateBERT	CE	58.8%	54.8%	76.5%
HateBERT	CE+SCL (Gunel et al., 2021)	55.9%	52.7%	76.9%
HateBERT	CE+ImpCon (Kim et al., 2022)	63.9%	59.5%	**77.4%**
HateBERT	CE+SharedCon (Ahn et al., 2024)	63.5%	57.7%	77.1%
HateBERT	CE+AugSH	58.2%	54.8%	77.2%
HateBERT	CE+ImpSH	**65.0%**	**60.8%**	76.4%
		SBIC→IHC	SBIC→DYNAHATE	SBIC→SBIC
BERT	CE	59.7%	60.3%	83.7%
BERT	CE+SCL (Gunel et al., 2021)	59.4%	60.9%	83.7%
BERT	CE+ImpCon (Kim et al., 2022)	61.4%	61.2%	83.8%
BERT	CE+SharedCon (Ahn et al., 2024)	**62.5%**	**62.0%**	83.8%
BERT	CE+AugSH	58.4%	60.6%	82.3%
BERT	CE+ImpSH	61.8%	**62.0%**	**84.1%**
HateBERT	CE	59.0%	60.2%	84.1%
HateBERT	CE+SCL (Gunel et al., 2021)	59.5%	59.3%	84.3%
HateBERT	CE+ImpCon (Kim et al., 2022)	60.1%	60.4%	**84.8%**
HateBERT	CE+SharedCon (Ahn et al., 2024)	59.3%	59.4%	84.5%
HateBERT	CE+AugSH	60.0%	**60.6%**	84.4%
HateBERT	CE+ImpSH	**60.6%**	60.4%	84.4%

Table 2: F1-scores for in-domain and cross-dataset evaluations. Models are trained on IHC or SBIC and tested on the dataset indicated by the arrow. Bold marks the best objective *within each encoder* for a given source→target; ties are both bolded. See Appendix A for details.

remains best. In both cases IMPSH still surpasses the next strongest baseline which is IMPCON, and it remains above AUGSH across all three evaluations, which supports that implied-statement supervision adds signal beyond augmentation alone. While SHAREDCON shows high scores in specific settings, it relies on clustering-based shared semantics which is often volatile and sensitive to batch composition. In contrast, IMPSH maintains more consistent alignment by mining negatives specifically relative to the context of the implied statement, focusing on separating negatives that are semantically close to the positive. With HateBERT trained on IHC, IMPSH is strongest on both transfer targets but it does not surpass the strongest in-domain baseline, where IMPCON is best and AUGSH is second, and IMPSH falls below both. This suggests a transfer versus in-domain tension in this configuration, where objectives that best separate the source dataset can differ from objectives that best preserve meaning under cross-dataset shift, while the transfer gap between IMPSH and AUGSH still supports benefits beyond augmentation on both targets.

With BERT trained on SBIC, IMPSH does not surpass the strongest baseline for transfer to IHC, where SHAREDCON remains best, but it surpasses the next strongest baseline which is IMPCON and it also stays above AUGSH. For transfer to DYNAHATE, IMPSH matches the strongest baseline which is SHAREDCON and it stays above the next strongest baseline which is IMPCON. For in-domain evaluation on SBIC, IMPSH is the strongest objective and it surpasses the next strongest baselines, IMPCON and SHAREDCON, and it remains above AUGSH, which again supports that implied-statement supervision contributes beyond augmentation.

With HateBERT trained on SBIC, IMPSH is best for transfer to IHC and it surpasses the next strongest baseline which is IMPCON while remain-

ing above AUGSH. For transfer to DYNAHATE, IMPSH does not surpass the strongest baseline, where AUGSH is best, and IMPSH matches the next strongest baseline which is IMPCON. This is the one setting where the ablation suggests that multi-view regularization accounts for most of the gain and implied-statement supervision adds limited additional benefit. For in-domain evaluation on SBIC with HateBERT, IMPSH does not surpass the strongest baseline, where IMPCON is best, and it also does not surpass the next strongest baseline, where SHAREDCON is second, while it matches AUGSH, which is consistent with implied-statement supervision being less helpful for maximizing in-domain fit on this source.

5.2 Alignment and uniformity

Following standard practice for representation evaluation (Wang and Isola, 2020), we assess **Alignment** and **Uniformity** on our best encoder, HATEBERT (Caselli et al., 2021). We L2-normalize embeddings so that $\|f(x)\|_2 = 1$. Alignment measures how close a positive pair lands in the embedding space, and Uniformity measures how evenly the full set of normalized embeddings spreads on the hypersphere, which helps detect representation collapse. Lower is better for both metrics, and for Uniformity this typically appears as more negative values (Wang and Isola, 2020). We use $r = 2$ for Alignment and report a global score by averaging per-class values, and we use $t = 2$ for Uniformity.

This analysis supports our hypothesis because improved cross-domain generalization should appear as tighter positive neighborhoods without sacrificing global spread. Table 3 shows that IMPSH achieves the best Alignment in four of six train→test settings, specifically IHC→IHC, IHC→SBIC, SBIC→SBIC, and SBIC→IHC. The remaining two cases are DYNAHATE evaluations, where SHAREDCON attains lower Alignment, which is consistent with its clustering-based objective that pulls together shared semantics and can be effective under perturbation-heavy shifts (Ahn et al., 2024; Vidgen et al., 2021). For Uniformity, IMPSH is best in three of six settings, including both transfers to DYNAHATE, while IMPCON remains strongest on several in-domain or near-domain evaluations. Overall, IMPSH tends to improve local compactness under transfer while maintaining a competitive global spread, and this trend is qualitatively consistent with our t-SNE inspection.

	Dataset	ImpCon	SharedCon	ImpSH (Ours)
	Trained on IHC			
Align. →	IHC	1.925	1.785	**1.741**
	SBIC	1.736	1.663	**1.659**
	DYNAHATE	1.683	**1.423**	1.754
	Trained on SBIC			
	SBIC	1.590	1.738	**1.457**
	IHC	1.910	1.689	**1.386**
	DYNAHATE	1.870	**1.222**	1.395
	Trained on IHC			
Unif. →	IHC	**-3.471**	-2.695	-3.373
	SBIC	**-3.104**	-2.593	-2.925
	DYNAHATE	-2.613	-2.322	**-3.351**
	Trained on SBIC			
	SBIC	-2.550	-2.803	**-2.891**
	IHC	**-3.170**	-2.431	-2.657
	DYNAHATE	-2.620	-1.766	**-2.765**

Table 3: Averaged Alignment and Uniformity scores across 4 seeds comparison across methods trained on HateBERT Encoder. Lower is better (↓). Best results in bold.

5.3 Qualitative Representation Analysis

We further probe the embedding space with t-SNE projections that provide a qualitative view of how classes and target groups organize in the learned representation. We report a class-colored view and a target-colored view, where target labels serve as a rough proxy for shared topical content. We train on IHC as the source domain, visualize the IHC in-domain test split, and then visualize transfer to SBIC and DYNAHATE. On IHC, both IMPCON and SHAREDCON show target-consistent clusters in the target view (Fig. 4d-f). In the class view, IMPCON shows more class interleaving within these clusters, while SHAREDCON and IMPSH exhibit a clearer large-scale separation between HATE and NON-HATE (Fig. 4a-c). IMPSH appears to preserve target-coherent neighborhoods while reducing cross-class overlap in several regions of the map.

Under transfer to SBIC, class mixing increases for all methods, which is expected under a larger shift, but IMPSH shows comparatively less overlap in the class view while still maintaining target-coherent neighborhoods in the target view (Fig. 4g-l). On DYNAHATE, all projections become more fragmented and mixed, yet IMPSH retains a weak but visible global class structure compared to the baselines (Fig. 4m-o). These observations are qualitative and can vary with t-SNE settings, but they

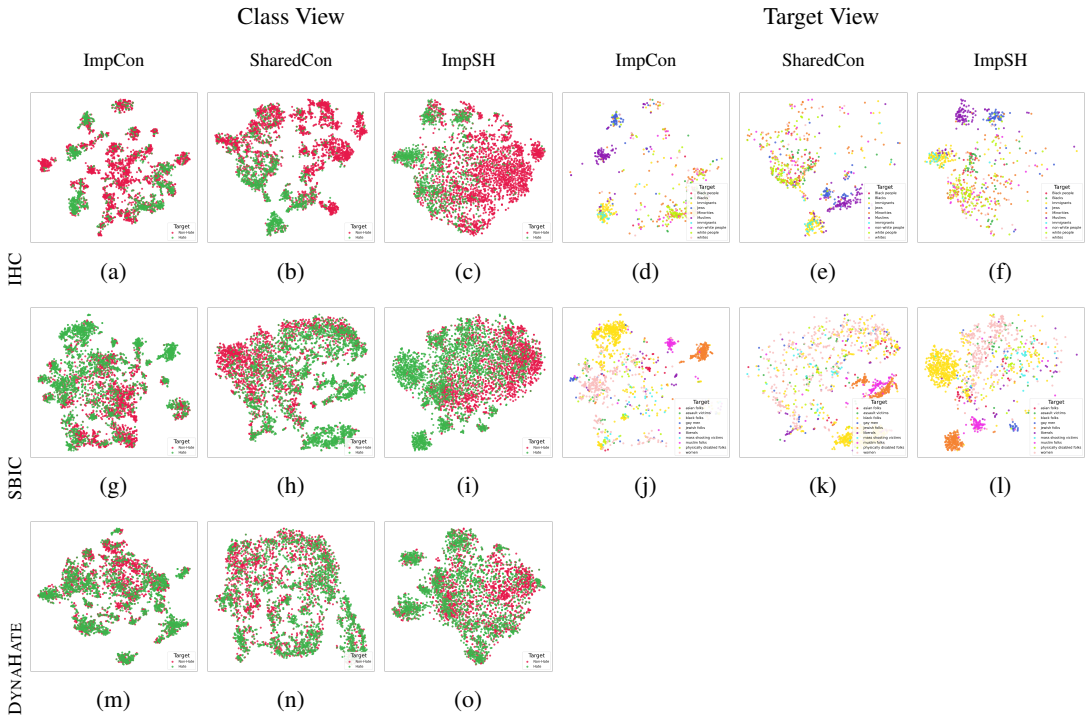

Figure 4: t-SNE visualization of test embeddings. Rows: IHC (in-domain), SBIC and DynaHate (cross-domain). Left: class-colored; right: target-centric clusters. ImpCon fragments class boundaries into target-centric clusters. SharedCon shows partial separation. ImpSH maintains global Hate/Non-Hate separation across all domains.

are broadly consistent with the Alignment and Uniformity trends in Table 3.

6 Case Study

Table 4 lists nearest neighbours from ImpSH with cosine similarity (lower is closer). True positives show that the model clusters coherent hate themes: on SBIC, misogynistic posts group together; on DynaHate, homophobic contexts align tightly. The false negatives expose three failure modes. First, topic bias: in SBIC the anchor is a non-hate "Stable Genius" joke, yet its neighbours are political Trump comments that match topic but not label, indicating target-centric proximity can override stance. Second, stance reversal: in DynaHate the anchor attacks immigrants while a neighbour condemns slurs, so opposite polarity is pulled close by shared lexical cues. Third, shallow lexical ties: "Rats are disgusting" is retrieved via short, high-overlap phrasing despite being off-topic. Overall, ImpSH groups most hate correctly at the sentence level, but residual errors arise from topic alignment dominating stance, brief or unequal contexts, and lexical shortcuts.

7 Conclusion

We proposed ImpSH and AugSH, two objectives that combine semantic priors with semi-hard negative mining to support transfer under domain shift in implicit hate detection. Across BERT and HateBERT, ImpSH is consistently competitive on cross-dataset evaluation and it is strongest in several transfer settings, most clearly when training on IHC with HateBERT. When training on SBIC, the best objective depends on the target and encoder, and clustering based baselines such as SharedCon or the augmentation ablation AugSH can be stronger in some cases. The geometry analysis supports this interpretation, where alignment and uniformity trends and qualitative visualizations suggest clearer Hate versus Non-Hate structure and less target specific fragmentation compared with prior contrastive objectives. Overall, our findings suggest that ImpSH facilitates superior separation in cross-data scenarios because it maps the input post somewhat bijectively to its underlying insinuation. By carefully mining negatives within the context of these implied statements, we mitigate the volatility of unsupervised clustering and achieve a more robust alignment of intent.

Table 4: **Alignment case study** (k=2). Rows show true-positives (TP) and false-negatives (FN) for label hate (Offensive in SBIC, Hate in DYNAHATE). NN-i / Sim-i give the i-th neighbour and cosine similarity (lower = closer).

Dataset	Class	Outc	Post	NN-1	Sim-1	NN-2	Sim-2
SBIC	off	TP	...femcunt can't last a week without sex, wants a superChad...	...Women should be euthanized after menopause...	0.68	...femoids give me disgusted looks in public...	0.66
SBIC	off	FN	...'Stable Genius' kept Trump rent-free in NPR cat ladies' heads...	...Trump never lets evidence stop his accusations...	0.46	...Trump didn't give me 100 % of what I wanted on Jan 20...	0.45
DYNAHATE	Hate	TP	...Comcast router 69 fags aren't as hidden as they think...	...Met girl via app; angry hood, Portland feels third-world...	0.69	...Bearded guy at the bar makes me wary of being here...	0.67
DYNAHATE	Hate	FN	...holy sheeeeit that's racist, immigrant rats...	...'immigrants are disgusting' shouldn't be feed	0.51	Rats are disgusting	0.50

8 Limitations and broader impact

Our method inherits limitations from triplet learning with semi-hard mining. It uses a fixed margin, so training can be sensitive to the margin choice and to batch composition. Small batches can produce too few informative triplets and larger batches increase compute and memory due to within-batch pairwise distance computation. Empirically, gains over strong contrastive baselines are modest and not consistent across all transfer directions, and prior objectives can remain stronger for in-domain fitting, so we position IMPSH as a complementary objective rather than a replacement that dominates prior work (Kim et al., 2022; Ahn et al., 2024). A further limitation is dependence on implication supervision because implied statements define positive pairs, which limits applicability when implication annotations are missing or noisy (Kim et al., 2022). A practical extension is to replace the fixed margin with an adaptive margin scheme and to reduce reliance on implied statements by constructing positives through label-aware neighborhoods that cluster shared semantics within each label (Ahn et al., 2024), while still sampling negatives from nearby opposing clusters to keep the objective focused on borderline errors.

9 Ethics statement

This work proposes a training objective for hate related text classification. We evaluate on public datasets that contain hateful and abusive content, so readers may face secondary exposure. We include a content warning and we minimize qualitative examples. When examples are necessary, we keep excerpts short and we redact slurs and identity terms when this does not change the linguistic phenomenon under discussion.

Our objective uses semi-hard negative mining around borderline pairs. This focuses learning on difficult cases, but it can also amplify the impact of annotation noise and spurious cues such as identity markers that correlate with labels. For this reason, we emphasize cross-dataset evaluation and we recommend targeted error analysis that checks false positives for benign identity mentions and false negatives for implicit hate. Where metadata permits, we recommend reporting group-wise error rates and documenting dataset populations and labeling choices.

We do not present trained checkpoints as a ready to deploy moderation tool. Hate detection models can be misused for surveillance or censorship and they can produce harmful errors in deployment. If models or code are released, we recommend documenting intended use, failure modes, and evaluation conditions using established responsible NLP guidance and model documentation practices.

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 2024*, pages 10444–10455.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse*

and Harms (WOAH 2021), pages 17–25, Online. Association for Computational Linguistics.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, 1, pages 512–515.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, and 1 others. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations (ICLR)*.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hung Huynh, Vittorio Murino, and Tinne Tuytelaars. 2020. Boosting contrastive self-supervised learning with false negatives. *arXiv preprint arXiv:2011.11765*.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Yejin Kim, Soojin Park, and Young-Sook Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Danqi Liao. 2021. Sentence embeddings using supervised contrastive learning. *arXiv preprint arXiv:2104.09540*.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pages 7721–7735. PMLR.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Computer Vision – ECCV 2020*, pages 681–699, Cham. Springer International Publishing.

Isar Nejadgholi and Svetlana Kiritchenko. 2020. On cross-dataset generalization in automatic detection of online abuse. *arXiv preprint arXiv:2010.07414*.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040. Association for Computational Linguistics.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *ICLR*.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Leo 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*, pages 5477–5490, Online. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. 2023. Id and ood performance are sometimes inversely correlated on real-world datasets. *Advances in Neural Information Processing Systems*, 36:71703–71722.

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 (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5022–5030.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. 2017. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848.

Hong Xuan, Abby Stylianou, and Robert Pless. 2020. Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2474–2482.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.

A Seed Report

Table 5 reports macro-F1 across four random seeds for in-domain and cross-domain evaluation. When trained on IHC, BERT (α=0.3) achieves an in-domain mean macro-F1 of 0.783 (std 0.0025). Cross-domain evaluation yields 0.614 (std 0.0119) on SBIC and 0.592 (std 0.0036) on DYNAHATE. Under the same training set, HATEBERT (α=0.4) achieves 0.764 (std 0.0049) in-domain, and 0.650 (std 0.0148) on SBIC and 0.608 (std 0.0035) on DYNAHATE in cross-domain evaluation.When trained on SBIC, BERT (α=0.4) achieves an in-domain mean macro-F1 of 0.841 (std 0.0022), with cross-domain results of 0.618 (std 0.0195) on IHC and 0.620 (std 0.0024) on DYNAHATE. HATEBERT (α=0.2) achieves 0.844 (std 0.0035) in-domain, with cross-domain results of 0.606 (std 0.0243) on IHC and 0.604 (std 0.0180) on DYNAHATE. Overall, the results are stable across seeds. The largest variance appears in the SBIC→IHC transfer setting, suggesting higher sensitivity to initialization in that direction.

B Sampling Strategies Ablation

To determine the optimal mining strategy for our triplet objective, we evaluate four sampling rules. These rules govern how negative examples are selected for each anchor-positive pair.

Random Select a same-class post at random. This ignores the implied statement for defining negatives.

Hard Margin Use the implied statement as positive. Always pick the hardest negative that violates the margin.

Semi-Hard Use the implied statement as positive. Choose a negative inside the margin but farther from the anchor than the positive.

Semi-Hard + Fallback Attempt semi-hard first. If no suitable negative is found in the batch, fall back to the hard margin rule.

Our initial hypothesis was that the **Hard Margin** strategy would be most effective. By forcing the model to contend with the most confusing negative samples relative to a post's implied meaning, we expected it to learn to separate similar insinuations effectively.

However, the results in Table 6 point decisively to the superiority of **Semi-Hard** sampling for generalization. When trained on IHC, it achieves the highest F1 score on both cross-domain datasets. This strength is mirrored when training on SBIC, where Semi-Hard again secures the best performance on the IHC cross-domain task plus ties for the best score on DYNAHATE. Its robust performance, including tying for the best in-domain

Table 5: Macro-F1 across four seeds on IMPSH. Seed/Mean are shown in %, while Std is reported in raw F1 (0-1).

Model	α	Train	Test	Seed 0	Seed 1	Seed 2	Seed 3	Mean	Std
BERT	0.3	IHC	IHC	78.7%	78.3%	78.1%	78.3%	78.3%	0.0025
			SBIC	62.5%	61.3%	62.0%	59.7%	61.4%	0.0119
			DYNAHATE	59.5%	59.2%	58.7%	59.4%	59.2%	0.0036
HateBERT	0.4	IHC	IHC	76.0%	76.2%	77.1%	76.2%	76.4%	0.0049
			SBIC	64.9%	63.0%	66.5%	65.5%	65.0%	0.0148
			DYNAHATE	60.9%	60.5%	61.2%	60.5%	60.8%	0.0035
BERT	0.4	SBIC	SBIC	84.0%	84.4%	83.8%	84.0%	84.1%	0.0022
			IHC	63.6%	59.2%	61.7%	62.9%	61.8%	0.0195
			DYNAHATE	62.2%	61.8%	62.2%	61.9%	62.0%	0.0024
HateBERT	0.2	SBIC	SBIC	84.7%	84.5%	84.5%	83.9%	84.4%	0.0035
			IHC	62.7%	58.4%	58.6%	62.8%	60.6%	0.0243
			DYNAHATE	62.9%	59.0%	59.2%	60.5%	60.4%	0.0180

score, contrasts sharply with the volatility of other methods. Random sampling may perform well in-domain but fails on transfer, while Hard Margin is inconsistent across different setups.

This empirical outcome is consistent with the original analysis of triplet loss by (Schroff et al., 2015). They argue that focusing exclusively on the hardest negatives can yield unstable gradients plus lead to poor local minima early in training. The semi-hard approach provides more stable learning signals, balancing model improvement without risking collapse. Given its superior and consistent performance on cross-domain evaluation, we adopt the **Semi-Hard** sampling strategy for all main experiments.

Table 6: Macro F1 for each mining rule. Best score in each column is bold.

Mining Strategy	IHC	SBIC	DYNAHATE
Trained on IHC			
Hard Margin	0.785	0.584	0.565
Random	**0.789**	0.593	0.545
Semi-Hard	0.787	**0.625**	**0.595**
Semi-Hard + Fallback	0.777	0.619	0.585
Trained on SBIC			
Mining Strategy	SBIC	IHC	DYNAHATE
Hard Margin	**0**.840	0.618	0.602
Random	0.814	0.610	0.602
Semi-Hard	**0.847**	**0.627**	**0.629**
Semi-Hard + Fallback	0.841	0.598	0.618

Implementation Details of Ablation We evaluate both BERT and HateBERT encoders on models trained with either IHC or SBIC. For the semi-hard triplet mining strategy, we swept the margin parameter $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and selected the best-performing configuration per model-dataset pair based on validation performance. The optimal margins were: $\alpha = 0.3$ for BERT trained on IHC, $\alpha = 0.4$ for HateBERT trained on IHC, $\alpha = 0.4$ for BERT trained on SBIC, and $\alpha = 0.2$ for HateBERT trained on SBIC. we replicate only on seed 0.

C Alpha Study

We conducted an α hyperparameter study using the same training setup and optimization settings described in Section 4.2. Each configuration was evaluated by sweeping $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and performance was measured using macro-F1 averaged over multiple random seeds. The results for both in-domain and cross-domain evaluations are reported in Table 7. For each training setup, we select the α value that yields the best overall performance, which is highlighted in bold.

D Representation quality through class alignment and uniformity

To quantitatively evaluate the structure of the learned embedding space, we use the **Alignment** and **Uniformity** metrics proposed by Wang and Isola (2020). These metrics provide a formal way to measure two desirable properties of a representation: that similar samples should be mapped

Table 7: Alpha study on IMPSH: macro-F1 (mean over seeds) for in-domain and cross-domain evaluation across different α values. Bold indicates the selected α per training setup (BERT+IHC: α=0.3, HateBERT+IHC: α=0.4, BERT+SBIC: α=0.4, HateBERT+SBIC: α=0.2).

Model	**Train**	**Test**	α				
			0.1	0.2	0.3	0.4	0.5
BERT	IHC	IHC (in)	0.774	0.778	**0.783**	0.779	0.780
		SBIC	0.580	0.609	**0.614**	0.598	0.618
		DYNAHATE	0.563	0.577	**0.592**	0.586	0.584
HateBERT	IHC	IHC (in)	0.767	0.762	0.761	**0.764**	0.769
		SBIC	0.611	0.634	0.637	**0.650**	0.645
		DYNAHATE	0.566	0.589	0.606	**0.608**	0.606
BERT	SBIC	SBIC (in)	0.839	0.835	0.839	**0.841**	0.839
		IHC	0.602	0.609	0.612	**0.618**	0.613
		DYNAHATE	0.611	0.609	0.610	**0.620**	0.613
HateBERT	SBIC	SBIC (in)	0.842	**0.844**	0.844	0.845	0.842
		IHC	0.600	**0.606**	0.604	0.603	0.594
		DYNAHATE	0.605	**0.604**	0.605	0.600	0.600

to nearby embeddings (intra-class compactness) and that dissimilar samples should be spread out evenly (inter-class separation). Embeddings are L2-normalized before computing distances.

Alignment

Alignment measures the expected distance between embeddings of positive pairs. In our context, a positive pair consists of two samples belonging to the same class. A lower alignment score indicates that samples within the same class are more tightly clustered. It is defined as:

$$\mathcal{L}_{\text{align}}(f; r) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} [\|f(x) - f(y)\|_2^r] \quad (8)$$

where f is the encoder, $(x, y) \sim p_{\text{pos}}$ denotes a pair of samples drawn from the positive pair distribution (i.e., having the same class label), and $\alpha > 0$ is a parameter (typically set to 2).

Uniformity

Uniformity measures how well the embeddings of negative pairs are distributed across the embedding space. A lower uniformity score indicates that dissimilar samples are spread further apart and more evenly, maximizing the entropy of the representation. It is defined as:

$$\mathcal{L}_{\text{uniform}}(f; t) \triangleq \log \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{neg}}} [e^{-t\|f(x)-f(y)\|_2^2}] \quad (9)$$

where $(x, y) \sim p_{\text{neg}}$ denotes a pair of samples drawn from the negative pair distribution (i.e., having different class labels), and $t > 0$ is a parameter (typically set to 2).

Together, these two metrics provide a comprehensive evaluation of the global structure of the embedding space, quantifying both the cohesion within classes and the separation between them.