Submitted to *INFORMS Journal on Computing*
manuscript

Online Supplement: Semantic Aggregated Adversarial Training Framework for Hate Speech Detection

Appendix A Theoretical Analysis

In this section, we present a theoretical analysis of the effectiveness of semantic aggregation in imbalanced adversarial training. We first illustrate the limitations in reweighting-based adversarial training through robust risk analysis, where minority samples are assigned larger weights. Standard adversarial training seeks to minimize the worst-case loss over the set of allowable perturbations:

$$R_{\text{adv}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\max_{\delta\in S}\ell(f_\theta(x+\delta), y)\Big], \tag{A.1}$$

where δ belongs to the admissible perturbation set S, which represents the allowed perturbations, typically bounded by a p-norm such that $\|\delta\|_p \le \epsilon$. We present a theoretical analysis using Rademacher complexity to quantify the impact of both sample weights and adversarial perturbations δ on the generalization error. Specifically, for a function class $\mathcal{F}$ and a dataset $\{x_1, \ldots, x_N\}$, the empirical Rademacher complexity is defined as: (Yin et al. 2019):

$$\hat{\mathcal{R}}_N(\mathcal{F}) = \mathbb{E}_\rho\Big[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\rho_i f(x_i)\Big], \tag{A.2}$$

where $\rho_i \in \{-1, +1\}$ are independent Rademacher variables. Rademacher complexity measures the ability of $\mathcal{F}$ to fit random noise. For adversarial training, the function class under perturbation becomes:

$$\mathcal{F}_{\text{adv}} = \Big\{x \mapsto \max_{\delta\in S} f_\theta(x+\delta) : f_\theta \in \mathcal{F}\Big\}, \tag{A.3}$$

with sample-specific weights ω_i, the empirical risk can be denoted as:

$$\hat{R}^{\text{imbal}}_{\text{adv}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\omega_i \max_{\delta_i\in S}\ell(f_\theta(x_i+\delta_i), y_i), \tag{A.4}$$

the corresponding weighted Rademacher complexity becomes:

$$\hat{\mathcal{R}}_N(\mathcal{F}^{\text{imbal}}_{\text{adv}}) = \mathbb{E}_\rho\Big[\sup_{f_\theta\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\rho_i\omega_i\max_{\delta_i\in S} f_\theta(x_i+\delta_i)\Big]. \tag{A.5}$$

Assuming f_θ is L-Lipschitz and bounded by B (Pauli et al. 2021), for each $\delta_i \in S$, we have:

$$|f_\theta(x_i+\delta_i)| \le |f_\theta(x_i)| + L\|\delta_i\|_p \le |f_\theta(x_i)| + L\epsilon. \tag{A.6}$$

applying Khintchine inequality gives an upper bound for the weighted complexity (Bartlett and Mendelson 2002):

$$\hat{\mathcal{R}}_N(\mathcal{F}^{\text{imbal}}_{\text{adv}}) \le \frac{\sqrt{\sum_i \omega_i^2}}{N}(B + L\epsilon). \tag{A.7}$$

Then, for balanced weights $\omega_i \approx 1$, the empirical Rademacher complexity scales as $\hat{\mathcal{R}}_N \sim B/\sqrt{N}$. In contrast, for imbalanced weights with $\omega_i \gg 1$ for minority samples, it scales as $\hat{\mathcal{R}}_N \sim \frac{\sqrt{\sum_i \omega_i^2}}{N}(B + L\epsilon) \gg B/\sqrt{N}$. Moreover, standard Rademacher theory provides an upper bound on the gap between the expected adversarial risk and the empirical risk (Bartlett and Mendelson 2002):

$$R_{\text{adv}}(\theta) - \hat{R}_{\text{adv}}^{\text{imbal}}(\theta) \leq 2\hat{\mathcal{R}}_N(\mathcal{F}_{\text{adv}}^{\text{imbal}}) + O\Big(\sqrt{\frac{\log(1/\delta)}{N}}\Big). \tag{A.8}$$

Therefore, the combination of large sample weights and a substantial perturbation range ϵ can inflate the Rademacher complexity, increasing the upper bound on the robust generalization error and widening the generalization gap. As a result, imbalanced adversarial training is especially susceptible to severe overfitting. To mitigate the overfitting problem, we attempt to optimize the upper bound of robust generation risk.

To provide a theoretical interpretation of semantic aggregation, following established approaches (Dobriban et al. 2023), we consider a binary classification problem with imbalanced classes:

$$y \sim \begin{cases} 0 & \text{with probability } p_0 = K/(K+1) \\ 1 & \text{with probability } p_1 = 1/(K+1) \end{cases}, \quad x \sim \begin{cases} \mathcal{N}(\mu_0, \sigma_0^2 I) & \text{if } y = 0 \\ \mathcal{N}(\mu_1, \sigma_1^2 I) & \text{if } y = 1 \end{cases}, \tag{A.9}$$

where μ_y and σ_y denote the mean and standard deviation of the features for each class, with $p_0 \gg p_1$ reflecting the class imbalance. Let $h = g_\theta(x)$ denote the feature mapping, and $f_\theta(h) = \text{sign}(\langle w, h\rangle + b)$ defines a linear classifier.

Let $\Delta := \mu_1 - \mu_0$ denotes the inter-class displacement, with $d := \|\Delta\|$ representing the inter-class distance. We decompose the classifier weight vector as $w = \alpha u + w_\perp$, where $u = \Delta/d$ and $\alpha = \langle w, u\rangle$. The effective class margin is defined as $\gamma = \langle w, \mu_1 - \mu_0\rangle = \alpha d$, which is bounded by $\gamma \leq \|w\| d$ according to the Cauchy–Schwarz inequality. For each class y, we define the class-wise deviation as:

$$\epsilon_h := h_y - \mu_y, \quad h_y + \delta = \mu_y + \epsilon_h + \delta, \tag{A.10}$$

and the projection mean as $m_y = \langle w, \mu_y\rangle$. Let the linear score be $a = \langle w, h\rangle$, its conditional second moment can be decomposed as:

$$\mathbb{E}[a^2 \mid y] = m_y^2 + \|w\|^2 \sigma_y^2, \tag{A.11}$$

where σ_y^2 is the intra-class variance along w. Finally, the adversarial error probability for each class y can be expressed as follows:

$$\text{AdvError}_y = \Pr(f_\theta(h_y + \delta) \neq y) = \Pr(\langle w, \epsilon_h + \delta\rangle \leq -\gamma_y), \tag{A.12}$$

where $\gamma_y := \langle w, \mu_y\rangle + b$ denotes the class-wise margin. Define $X := \langle w, \epsilon_h\rangle$ with $\mathbb{E}[X] = 0$ and $\text{Var}(X) = \|w\|^2 \sigma_y^2$. Since $|\langle w, \delta\rangle| \leq \|w\|\epsilon$, the error probability can be bounded using Chebyshev's inequality (Xu and Mannor 2012):

$$\text{AdvError}_y \leq \Pr(|X| \geq \gamma_y - \|w\|\epsilon) \leq \frac{\|w\|^2 \sigma_y^2}{(\gamma_y - \|w\|\epsilon)^2}. \tag{A.13}$$

Similarly, the natural–adversarial gap for the minority class is given by:

$$\mathrm{Gap}_1 := \mathrm{AdvError}_1 - \mathrm{NatError}_1 \le \|w\|^2 \sigma_1^2 \left(\frac{1}{(\gamma - \|w\|\epsilon)^2} - \frac{1}{\gamma^2} \right). \tag{A.14}$$

Considering the magnitude constraint on ϵ, we have $\gamma_y > |w|\epsilon$, so increasing the class margin effectively reduces the upper bound on the error.

The weighted empirical Rademacher complexity in Equation A.5 of the adversarial hypothesis class can be further bounded as:

$$\hat{\mathcal{R}}_N(\mathcal{F}_{\mathrm{adv}}) \lesssim \frac{1}{N} \sqrt{\sum_{i=1}^{N} w_i^2 \, \mathbb{E}[a_i^2]} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} w_i^2 \big(m_{y_i}^2 + \|w\|^2 \sigma_{y_i}^2 \big)}. \tag{A.15}$$

Thus, class variance is positively correlated with both the error probability upper bound and the Rademacher complexity. Optimizing this bound can effectively reduce the robust risks in imbalanced adversarial training. Building on this insight, we propose Semantic Aggregation, which mitigates overfitting by increasing inter-class separation and compressing class feature distributions, thereby reducing both adversarial errors and the generalization gap:

$$\|\mu_1 - \mu_0\| \to \|\mu_1 - \mu_0\|^{\mathrm{SA}} > \|\mu_1 - \mu_0\| \quad \Rightarrow \quad \gamma^{\mathrm{SA}} > \gamma, \sigma_y \to \sigma_y^{\mathrm{SA}} < \sigma_y. \tag{A.16}$$

Appendix B Supplementary Experimental Results

This section provides additional experimental results that supplement the experiments reported in the main text. Table B.1 shows the robust performance of benchmark methods under character-level attacks. Table B.2 to Table B.5 present the robust performance of eight benchmark methods under TextBugger, TextFooler, Hotflip and BAE attack. As a side note, SAAT performs better against character-level attacks, such as TextBugger and HotFlip, than against word-level attacks like TextFooler and BAE. These character-level attacks are often more effective, leading to significant degradation in detection performance. In addition, Table B.6 illustrates the bias and explainability performance of benchmark methods on HateXplain dataset. Table B.7 reports mean and standard deviation scores across varying model initializations and dataset partitions. Table B.9 presents the robust performance of the benchmark adversarial training methods. In Table B.10, we report the performance of benchmark aggregation methods compared with SAAT and its variant without semantic aggregation. Figure B.1 shows the robust performance of benchmark methods under black-box attacks with varied sample sizes. Figure B.2 presents the regular and robust performance of SAAT under varied hyper-parameters.

Moreover, we conduct a corpus-level comparative analysis between Toxigen and three other imbalanced datasets (HSOL, CLID, and Latent Hatred) using three categories of indicators: syntactic complexity, lexical overlap, and adversarial phrasing. Syntactic complexity is measured by *td* (Tree Depth), *asl* (Average Sentence Length), and *awl* (Average Word Length), capturing sentence and word structural variation via dependency parsing (Wang and Buschmeier 2024). Lexical

Table B.1 Robust Performance Under White-Box Attacks at Character Level

Methods	Wassem							OLID						
	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$
Tw-Star	0.4223	0.4394	0.5662	0.4948	0.4098	0.6127	0.4259	0.3009	0.0440	0.0699	0.0540	0.2551	0.4449	0.6228
CNN_LSTM	0.5615	0.5476	0.7021	0.6153	0.5521	0.7396	0.2695	0.4887	0.1856	0.1433	0.1617	0.2745	0.5166	0.6725
CNN_GRU	0.499	0.4991	0.5823	0.5375	0.4955	0.6878	0.3417	0.4831	0.1662	0.1539	0.1598	0.2128	0.3690	0.7122
ToxDectRoBERTa	0.5203	0.5132	0.7804	0.6192	0.4852	0.7777	0.3067	0.5144	0.2019	0.1688	0.1839	0.3506	0.6255	0.5666
BERT-HateXplain	0.4835	0.4878	0.6016	0.5388	0.4781	0.6894	0.3264	0.5279	0.2144	0.2049	0.2095	0.3042	0.4985	0.6188
HateBERT	0.5956	0.5841	0.6615	0.6204	0.5933	0.8185	0.2372	0.5189	0.2067	0.1994	0.2030	0.3144	0.6090	0.5838
fBERT	0.5617	0.5435	0.7522	0.6303	0.5435	0.8157	0.268	0.5387	0.2203	0.2276	0.2239	0.3688	0.6499	0.5328
SAAT(Ours)	**0.5982**	**0.5932**	**0.8713**	**0.7058**	**0.7382**	**0.8982**	**0.2119**	**0.7466**	**0.6728**	**0.6948**	**0.6836**	**0.5899**	**0.7204**	**0.2697**
Methods	**HSOL**							**de Gibert**						
	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$
Tw-Star	0.4532	0.1042	0.1218	0.1123	0.1797	0.4112	0.5546	0.2384	0.1188	0.3635	0.1791	0.1985	0.4196	0.506
CNN_LSTM	0.4665	0.1213	0.1285	0.1248	0.1476	0.399	0.5415	0.364	0.1284	0.3833	0.1924	0.2775	0.5843	0.484
CNN_GRU	0.4956	0.1816	0.1709	0.1761	0.1657	0.4056	0.5066	0.3453	0.1303	0.411	0.1979	0.2622	0.5177	0.384
ToxDectRoBERTa	0.4881	0.1645	0.1542	0.1592	0.3268	0.6441	0.5125	0.6647	0.169	0.3243	0.2222	0.504	0.7466	0.392
BERT-HateXplain	0.4695	0.1344	0.1161	0.1246	0.3227	0.6488	0.5344	0.5124	0.1405	0.4857	0.2180	0.4316	0.614	0.578
HateBERT	0.459	0.1559	0.1481	0.1519	0.3713	0.691	0.5122	0.5762	0.1887	0.4412	0.2643	0.5241	0.7374	0.414
fBERT	0.4662	0.1673	0.1396	0.1522	0.3519	0.6616	0.5223	0.6823	0.1438	0.3818	0.2089	0.505	0.7497	0.376
SAAT(Ours)	**0.7945**	**0.7348**	**0.7155**	**0.725**	**0.8123**	**0.8013**	**0.1458**	**0.8121**	**0.5499**	**0.4458**	**0.4924**	**0.6785**	**0.8346**	**0.15**
Methods	**GHC**							**HateXplain**						
	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$
Tw-Star	0.3459	0.1145	0.4858	0.1853	0.257	0.452	0.6176	0.2408	0.1667	0.1346	0.1489	0.2307	0.4103	0.6883
CNN_LSTM	0.4581	0.1121	0.5792	0.1878	0.3184	0.5723	0.4892	0.3375	0.2445	0.156	0.1905	0.3145	0.5237	0.5798
CNN_GRU	0.4757	0.1232	0.5583	0.2019	0.3268	0.5945	**0.423**	0.3108	0.1935	0.1259	0.1525	0.2842	0.4984	0.6086
ToxDectRoBERTa	0.5016	0.1167	**0.566**	0.1935	0.4156	0.6877	0.536	0.4822	0.4583	0.2244	0.3013	0.3423	0.6333	0.411
BERT-HateXplain	0.491	0.1226	0.5614	0.2013	0.4139	0.6047	0.6032	0.295	0.2644	0.2307	0.2464	0.292	0.4714	0.6358
HateBERT	**0.6028**	0.1699	0.5556	0.2602	**0.494**	**0.7277**	0.48	0.3584	0.3214	0.2719	0.2946	0.3458	0.5377	0.5783
fBERT	0.518	0.1006	0.4035	0.1603	0.4111	0.6106	0.5542	0.3816	0.3286	0.2354	0.2706	0.3657	0.6066	0.8503
SAAT(Ours)	0.4515	**0.1971**	0.4867	**0.2806**	0.4673	0.657	0.595	**0.695**	**0.6338**	**0.3214**	**0.4265**	**0.3932**	**0.6726**	**0.267**
Methods	**Latent Hatred**							**Toxigen**						
	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$
Tw-Star	0.4995	0.4902	0.0503	0.0907	0.3725	0.5933	**0.0477**	0.1566	0.1918	0.214	0.2023	0.1532	0.2726	0.7926
CNN_LSTM	0.3106	0.3201	0.338	0.3288	0.3095	0.4759	0.5434	0.206	0.2024	0.2	0.2012	0.206	0.3512	0.7216
CNN_GRU	0.3114	0.348	0.3655	0.357	0.5707	0.6735	0.2162	0.1653	0.1735	0.178	0.1757	0.1649	0.2811	0.7806
ToxDectRoBERTa	0.2811	0.1309	0.2293	0.1667	0.2664	0.4421	0.6585	0.3154	0.3612	0.472	0.4219	0.2509	0.416	0.5677
BERT-HateXplain	0.3425	0.2163	0.3609	0.2705	0.3356	0.5452	0.521	0.2009	0.2656	0.347	0.3009	0.184	0.3533	0.7484
HateBERT	0.3104	0.0977	0.1221	0.1085	0.2729	0.4602	0.5811	0.3315	0.3534	0.413	0.3809	0.2757	0.4608	0.6024
fBERT	0.2886	0.1515	0.302	0.2018	0.2796	0.4831	0.616	0.1726	0.2014	0.226	0.2122	0.1679	0.2978	0.7821
SAAT(Ours)	**0.5503**	**0.5297**	**0.4273**	**0.473**	**0.6355**	**0.6685**	0.2144	**0.3356**	**0.3893**	**0.483**	**0.4311**	**0.2925**	**0.5042**	**0.3467**

Note: * refers to the robust performance under adversarial attacks.

Table B.2 Performance Under TextBugger

Methods	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$
	Wassem							**OLID**						
Tw-Star	0.0961	0.019	0.016	0.0174	0.0902	0.1769	0.8708	0.1721	0.0089	0.0062	0.0073	0.1486	0.2739	0.6876
CNN_LSTM	0.1324	0.0082	0.006	0.0068	0.1164	0.2075	0.8371	0.4647	0.152	0.1121	0.1290	0.2404	0.4816	0.5598
CNN_GRU	0.1194	0.0128	0.01	0.0112	0.1084	0.1945	0.842	0.4413	0.1013	0.1252	0.1120	0.2244	0.4595	0.6076
ToxDectRoBERTa	0.1306	0.0698	0.06	0.0645	0.1257	0.232	0.8385	**0.5304**	**0.1625**	0.1124	0.1329	0.3079	**0.7412**	**0.5576**
BERT-HateXplain	0.1257	0	0	0	0.1111	0.2265	0.8418	0.455	0.1004	0.1135	0.1065	0.2032	0.4639	0.6331
HateBERT	0.3308	0.0526	0.02	0.029	0.2588	0.5931	0.5951	0.515	0.1488	**0.152**	**0.1504**	**0.3497**	0.6687	0.5828
fBERT	0.2322	0.0179	0.01	0.0128	0.1908	0.4383	0.7013	0.5035	0.118	0.1437	0.1296	0.3395	0.6497	0.5685
SAAT(Ours)	**0.4367**	**0.2133**	**0.169**	**0.1886**	**0.4575**	**0.6861**	**0.5592**	0.4977	0.1392	0.1173	0.1273	0.2863	0.5733	0.5855
Methods	**HSOL**							**de Gibert**						
Tw-Star	0.5763	0.1065	0.1225	0.1139	0.3656	0.692	0.5305	0.1216	0.1411	0.2976	0.1139	0.1071	0.1985	0.63
CNN_LSTM	0.4173	0.1033	0.1442	0.1204	0.2944	0.5368	0.5478	0.3493	0.1457	0.3044	0.1204	0.2587	0.5384	0.646
CNN_GRU	0.4326	0.1128	0.1478	0.1279	0.302	0.555	0.5198	0.3356	0.1477	0.3061	0.1279	0.2509	0.5009	0.5699
ToxDectRoBERTa	0.4425	0.1788	0.2556	0.2104	0.3056	0.7601	0.5125	0.6246	0.1844	0.3268	0.2104	0.3842	0.646	0.5973
BERT-HateXplain	0.4518	0.1847	0.2675	0.2185	0.3155	0.7547	0.5071	0.3881	0.191	0.3455	0.2185	0.2795	0.4173	0.5708
HateBERT	0.475	0.1845	0.2566	0.2147	0.322	0.7811	0.5066	0.6204	0.1913	0.3552	0.2147	0.3827	0.6559	0.5442
fBERT	0.5981	0.1932	0.279	0.2283	**0.3742**	**0.7815**	0.4896	0.5962	0.2065	0.404	0.2283	0.378	0.6381	0.5392
SAAT(Ours)	**0.605**	**0.3705**	**0.3458**	**0.3578**	0.3681	0.6395	**0.3324**	**0.7399**	**0.4871**	**0.3949**	**0.4093**	**0.6182**	**0.7994**	**0.2145**
Methods	**GHC**							**HateXplain**						
Tw-Star	0.2635	0.106	0.3844	0.1662	0.2086	0.3392	0.6536	0.0726	0.0246	0.0045	0.0043	0.0677	0.1356	0.9085
CNN_LSTM	0.4595	0.1145	0.4755	0.1846	0.3148	0.5678	0.5692	0.216	0.1069	0.1024	0.1046	0.1791	0.3299	0.7276
CNN_GRU	0.473	0.116	0.4866	0.1873	0.3211	0.5896	0.553	0.1675	0.1122	0.1221	0.1169	0.1831	0.2643	0.7913
ToxDectRoBERTa	**0.6304**	**0.5562**	**0.5014**	**0.5274**	**0.4278**	**0.6274**	**0.5326**	0.3104	0.2476	0.2422	0.2449	0.2766	0.5398	0.5894
BERT-HateXplain	0.294	0.1523	0.4888	0.2322	0.2272	0.3169	0.5741	0.1124	0.1125	0.1123	0.1411	0.181	0.197	0.8706
HateBERT	0.4766	0.1342	0.5011	0.2117	0.3225	0.5053	0.5837	0.1657	0.1278	0.1162	0.1124	0.1932	0.2745	0.8129
fBERT	0.3868	0.139	0.4204	0.2089	0.2815	0.425	0.5777	0.1252	0.1113	0.1031	0.1078	0.1233	0.2297	0.8503
SAAT(Ours)	0.5742	0.3552	0.2688	0.3059	0.3533	0.526	0.5718	**0.5705**	**0.3548**	**0.3741**	**0.3642**	**0.4595**	**0.7152**	**0.56**
Methods	**Latent Hatred**							**Toxigen**						
Tw-Star	**0.4083**	0.1849	0.054	0.0836	0.3232	0.521	**0.4124**	0.0542	0.061	0.062	0.0615	0.0539	0.0901	0.9276
CNN_LSTM	0.2234	0.2505	0.2728	0.2635	0.2206	0.3694	0.6735	0.082	0.082	0.083	0.082	0.082	0.1407	0.8923
CNN_GRU	0.2663	0.1514	0.116	0.1313	0.4537	0.3873	0.752	0.062	0.0723	0.074	0.0731	0.0619	0.1056	0.916
ToxDectRoBERTa	0.2307	0.1788	0.1223	0.1452	0.1901	0.3028	0.7118	0.1954	0.2161	0.255	0.2339	0.1447	0.311	0.7462
BERT-HateXplain	0.2655	0.1021	0.1824	0.1309	0.2199	0.3602	0.6823	0.1224	0.1935	0.224	0.2076	0.1071	0.2104	0.8462
HateBERT	0.2267	0.1156	0.1436	0.1281	0.1843	0.3083	0.7018	0.1651	0.2087	0.247	0.2262	0.1603	0.2969	0.7843
fBERT	0.1984	0.1327	0.0625	0.085	0.1764	0.2753	0.7381	0.1322	0.1967	0.235	0.2141	0.1193	0.2464	0.8395
SAAT(Ours)	0.3852	**0.3661**	**0.4052**	**0.3847**	**0.4769**	**0.6356**	0.5333	**0.205**	**0.2389**	**0.275**	**0.2557**	**0.2016**	**0.3356**	**0.7192**

Table B.3 Performance Under TextFooler

Methods	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$
	Wassem							**OLID**						
Tw-Star	0.4582	0.5514	0.6224	0.5848	0.5562	0.6864	0.485	0.4817	0.4676	0.4046	0.4338	0.4779	0.569	0.7098
CNN_LSTM	0.4267	0.6013	0.7485	0.6669	0.6203	0.773	0.3012	**0.5743**	0.5804	0.4793	0.5250	0.5697	0.6524	0.4949
CNN_GRU	0.4374	**0.6179**	0.7183	0.6643	**0.6346**	0.758	0.3844	0.5316	**0.5232**	**0.5145**	**0.5188**	**0.5312**	0.6016	0.6293
ToxDectRoBERTa	0.5323	0.5209	0.796	0.6297	0.4969	0.7943	0.3317	0.3806	0.1809	0.2168	0.1972	0.3131	0.6307	0.4428
BERT-HateXplain	0.4561	0.4694	0.6763	0.5542	0.4283	0.7026	0.4062	0.2188	0.0621	0.1204	0.0819	0.1924	0.3876	0.6822
HateBERT	0.4364	0.4206	0.3362	0.3737	0.4303	0.5993	0.4352	0.3746	0.1942	0.1408	0.1632	0.3148	0.6377	0.4777
fBERT	**0.6406**	0.5943	**0.7569**	0.6658	0.6128	**0.8484**	**0.2365**	0.4203	0.3039	0.124	0.1761	0.3643	**0.6806**	**0.4247**
SAAT(Ours)	0.634	0.6106	0.7402	**0.6692**	0.6298	0.8398	0.2454	0.3425	0.2074	0.1125	0.1459	0.3052	0.4973	0.5169
Methods	**HSOL**							**de Gibert**						
Tw-Star	0.5659	0.3065	0.3364	0.3208	0.4034	0.7258	0.4331	0.2506	0.1408	0.372	0.2043	0.2466	0.5763	0.606
CNN_LSTM	0.5507	0.3561	0.3636	0.3598	0.4577	0.6931	0.446	0.2786	0.1108	0.43	0.1762	0.2285	0.6824	0.6484
CNN_GRU	0.5811	0.3376	0.3583	0.3476	0.4645	0.7465	0.4231	0.2654	0.1895	0.428	0.2627	0.2867	0.6548	0.6516
ToxDectRoBERTa	0.6452	0.5375	0.5524	0.5448	0.5376	0.8152	0.2577	0.6211	0.3382	0.4136	0.3721	0.5286	0.6809	0.3654
BERT-HateXplain	0.6335	**0.5516**	0.5262	0.5386	0.5358	0.7627	0.3041	0.6128	0.3405	0.4857	0.4003	0.5616	0.6614	0.378
HateBERT	**0.6624**	0.5172	0.5644	**0.5398**	**0.5988**	**0.8321**	**0.2457**	0.6665	0.3887	**0.4412**	0.4133	0.5841	0.7374	0.3414
fBERT	0.6094	0.5025	**0.5742**	0.4885	0.4927	0.817	0.2738	0.6382	0.3438	0.3818	0.3618	0.5205	0.7497	0.3765
SAAT(Ours)	0.6214	0.5172	0.5255	0.5213	0.5853	0.8217	0.2771	**0.6979**	**0.4194**	0.4322	**0.4257**	**0.6041**	**0.7552**	**0.3237**
Methods	**GHC**							**HateXplain**						
Tw-Star	0.3527	0.1617	0.4958	0.2439	0.5275	0.6428	0.6776	0.2818	0.2872	0.1634	0.2083	0.2988	0.5004	0.724
CNN_LSTM	0.4793	0.2741	0.384	0.3199	0.4743	0.6028	0.603	0.2678	0.2542	0.1792	0.2102	0.2711	0.4635	0.6892
CNN_GRU	0.4811	0.2676	0.4583	0.3379	0.5245	0.6555	0.5823	0.2723	0.2645	0.1766	0.2118	0.2666	0.4946	0.7064
ToxDectRoBERTa	0.4625	0.3133	0.2752	0.2930	0.3576	0.5749	0.5515	0.3982	0.3642	0.2241	0.2775	0.3943	**0.6447**	0.5077
BERT-HateXplain	0.4373	0.2654	0.2395	0.2518	0.3139	0.5403	0.6294	0.3066	0.2461	0.188	0.2132	0.2962	0.497	0.6331
HateBERT	0.4975	**0.3225**	0.2508	0.2822	0.3494	0.5522	0.5478	0.3923	0.369	0.248	0.2966	0.3957	0.6355	0.5187
fBERT	0.4684	0.298	0.2565	0.2757	0.3211	0.4956	0.6533	0.3824	0.3314	0.232	0.2729	0.3678	0.6031	0.5474
SAAT(Ours)	**0.6415**	0.2667	**0.4842**	**0.344**	**0.5244**	**0.7284**	**0.4857**	**0.4105**	**0.3856**	**0.3522**	**0.3716**	**0.4032**	0.6069	**0.4906**
Methods	**Latent Hatred**							**Toxigen**						
Tw-Star	**0.5517**	0.4555	0.066	0.1153	0.3722	0.5601	**0.096**	0.1726	0.1764	0.239	0.203	0.1718	**0.4595**	0.8335
CNN_LSTM	0.5147	0.4595	0.3714	0.4108	0.5101	0.6189	0.601	0.2154	0.2231	0.268	0.2435	0.2134	0.5145	0.781
CNN_GRU	0.5173	0.4716	0.3458	0.3990	0.5071	0.6028	0.6237	0.1809	0.1681	0.254	0.2023	0.2138	0.459	0.8092
ToxDectRoBERTa	0.2808	0.1309	0.2293	0.1667	0.2664	0.4421	0.6585	0.3188	0.3512	0.484	0.407	0.2661	0.4387	0.5149
BERT-HateXplain	0.5464	0.4637	0.4811	0.4722	0.5443	**0.7056**	0.4484	0.1923	0.2548	0.323	0.2837	0.1785	0.3431	0.7454
HateBERT	0.5142	0.3682	0.4378	0.4000	0.4958	0.6099	0.674	0.2515	0.2549	0.261	0.2574	0.2499	0.418	0.6843
fBERT	0.4646	0.288	0.4444	0.3495	0.4469	0.5867	0.642	0.2048	0.2279	0.248	0.2375	0.2025	0.3513	0.7418
SAAT(Ours)	0.5503	**0.4814**	**0.5211**	**0.5005**	**0.5532**	0.6713	0.4064	**0.3154**	**0.3698**	**0.516**	**0.4308**	**0.2908**	0.4502	**0.4769**

Table B.4 Performance Under HotFlip

Methods	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$
	Wassem							**OLID**						
Tw-Star	0.6011	0.5746	0.7704	0.6582	0.5881	0.7738	0.1638	0.4325	0.392	0.245	0.3015	0.4118	0.6135	0.3242
CNN_LSTM	**0.7025**	**0.664**	0.6842	0.6739	0.6983	0.8271	**0.1219**	0.4375	0.2642	0.307	0.2840	0.3497	0.633	0.2616
CNN_GRU	0.6175	0.6063	0.6733	0.638	0.6164	0.7435	0.1541	0.4075	0.1509	0.3204	0.2052	0.315	0.5829	0.3454
ToxDectRoBERTa	0.595	0.5638	0.6684	0.6117	0.5691	0.8438	0.2539	0.4575	0.3111	0.3207	0.3158	0.3616	0.7695	0.3394
BERT-HateXplain	0.5722	0.5483	**0.7951**	0.649	0.5471	0.8092	0.2525	0.4225	0.3086	0.3125	0.3105	0.3664	0.6672	0.35
HateBERT	0.6675	0.6358	0.7054	0.6688	0.667	0.8423	0.163	0.4675	0.3818	0.3195	0.3479	0.3869	0.7633	**0.3048**
fBERT	0.6706	0.6371	0.7192	0.6757	0.6652	0.8604	0.153	**0.4964**	**0.4526**	**0.3289**	**0.3810**	**0.4165**	0.7793	0.3245
SAAT(Ours)	0.675	0.6458	**0.775**	**0.7045**	**0.6717**	**0.8795**	0.1843	0.46	0.3571	0.3149	0.3347	0.3796	**0.7816**	0.3805
Methods	**HSOL**							**de Gibert**						
Tw-Star	0.4975	0.2868	0.2552	0.2701	0.3322	0.7233	0.5701	0.425	0.2414	0.327	0.2778	0.3421	0.6863	0.4138
CNN_LSTM	0.46	0.2143	0.2203	0.2173	0.3375	0.7023	0.5134	0.455	0.1786	0.2125	0.1941	0.3314	0.7291	0.2973
CNN_GRU	0.465	0.2667	0.3304	0.2952	0.3471	0.7436	0.3955	0.43	0.1625	0.2241	0.1884	0.3079	0.6711	0.3255
ToxDectRoBERTa	0.4825	0.3333	0.3035	0.3177	0.3529	0.8146	0.3689	0.4825	0.4557	0.4218	0.4381	0.4304	0.7702	0.3814
BERT-HateXplain	0.4875	0.3333	0.3025	0.3172	0.348	0.8315	0.3442	0.4175	0.3187	0.4145	0.3603	0.3708	0.6605	0.3993
HateBERT	0.4925	**0.4407**	0.3455	0.3871	0.3724	0.822	0.2677	0.4975	0.4909	0.4535	0.4715	0.4215	0.7921	0.3185
fBERT	**0.6543**	0.4283	**0.4655**	**0.4461**	**0.446**	**0.8412**	**0.1747**	0.4575	0.3333	0.3385	0.3359	0.3701	0.7703	0.3624
SAAT(Ours)	0.4709	0.3235	0.355	0.3385	0.3597	0.7905	0.3037	**0.645**	**0.6157**	**0.4816**	**0.5405**	**0.4658**	**0.8508**	**0.2791**
Methods	**GHC**							**HateXplain**						
Tw-Star	0.415	0.1852	0.1053	0.1340	0.3251	0.6844	0.3364	0.4737	0.4612	0.3457	0.3934	0.4616	0.6878	0.3935
CNN_LSTM	0.4512	0.1154	0.115	0.1152	0.3216	0.7217	0.3469	0.4925	0.4839	0.2252	0.3074	0.4534	0.6844	0.4102
CNN_GRU	0.4725	0.1333	0.1114	0.1214	0.329	0.4792	0.4196	0.4925	0.4818	0.2246	0.3064	0.4534	0.7009	0.392
ToxDectRoBERTa	0.646	0.4429	0.318	0.3702	0.4476	0.7008	0.1957	0.5125	0.5309	0.2153	0.3064	0.4652	0.7255	0.3594
BERT-HateXplain	0.4375	0.3932	0.253	0.3079	0.4122	0.6837	0.3048	0.5148	0.5203	0.3846	0.4423	0.5067	0.7009	0.4046
HateBERT	0.5013	0.4986	0.175	0.2591	0.441	0.8169	0.3117	0.5254	0.5424	0.3202	0.4027	0.5042	0.7294	0.3824
fBERT	0.6022	0.305	0.1799	0.2263	0.4793	0.8074	0.3333	0.5278	0.5443	0.3445	0.4219	0.5115	**0.7536**	0.3677
SAAT(Ours)	**0.7375**	**0.5364**	**0.3975**	**0.4566**	**0.5675**	**0.8487**	**0.1582**	**0.5496**	**0.5494**	**0.3956**	**0.46**	**0.5274**	0.7443	**0.3316**
Methods	**Latent Hatred**							**Toxigen**						
Tw-Star	0.4925	0.3816	0.049	0.0868	0.387	0.5977	**0.0238**	0.3945	0.4071	0.463	0.4332	0.3924	0.6007	0.482
CNN_LSTM	0.4975	0.4968	0.3852	0.4339	0.4911	0.6495	0.2867	0.3675	0.3707	0.378	0.3743	0.3674	0.564	0.4966
CNN_GRU	0.535	0.5522	0.3708	0.4431	0.522	0.6327	0.2336	0.3556	0.3478	0.32	0.3333	0.359	0.5336	0.493
ToxDectRoBERTa	0.4596	0.4633	0.505	0.4832	0.4589	0.6787	0.4758	0.3652	0.4182	**0.694**	0.5219	0.29	0.5406	0.5244
BERT-HateXplain	0.3052	0.2044	0.1307	0.1595	0.283	0.4571	0.4602	0.3579	0.3986	0.549	0.4619	0.336	0.581	0.5152
HateBERT	0.5911	0.5978	0.554	0.5729	0.5893	0.8026	0.3444	0.4775	0.4829	0.536	0.5081	0.4755	**0.6981**	0.4494
fBERT	0.4325	0.3373	0.1427	0.1979	0.3794	0.6194	0.4462	0.361	0.3828	0.454	0.4154	0.3554	0.5728	0.5453
SAAT(Ours)	**0.6492**	**0.6121**	**0.5267**	**0.5664**	**0.6641**	**0.8196**	0.2218	**0.5775**	**0.5132**	0.595	**0.5511**	**0.5432**	0.6422	**0.4146**

Table B.5 Performance Under BAE

Methods	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr\downarrow$
	Wassem							**OLID**						
Tw-Star	0.6975	0.6612	0.8071	0.7269	0.6936	0.8021	0.0478	0.585	0.6133	0.5446	0.5769	0.5784	0.6288	0.093
CNN_LSTM	0.6825	0.6652	0.735	0.6983	0.6816	0.7763	0.09	0.6557	0.6333	0.6175	0.6253	0.6605	0.6852	0.0964
CNN_GRU	0.7317	0.7072	0.7585	0.732	0.7292	0.7712	0.0458	0.6967	0.6765	0.5923	0.6316	0.6862	0.6764	0.1146
ToxDectRoBERTa	0.7752	0.7218	0.875	0.7911	0.7717	0.8803	0.0373	**0.7275**	0.7495	0.6365	0.6884	0.7255	**0.8206**	**0.054**
BERT-HateXplain	0.6899	0.6532	0.7881	0.7143	0.6855	0.8163	0.0676	0.7272	0.7164	0.6424	0.6774	**0.7194**	0.7948	0.126
HateBERT	0.7821	**0.7642**	0.8114	0.7871	**0.7798**	0.8831	**0.0341**	0.765	0.7359	0.6355	0.6820	0.7166	0.8054	0.0813
fBERT	0.7401	0.7081	0.8233	0.7609	0.7394	0.8739	0.0402	0.7003	**0.7496**	0.6574	0.7005	0.7168	0.8116	0.055
SAAT(Ours)	**0.7552**	0.7236	**0.8889**	**0.7978**	0.772	0.8652	0.052	0.6864	0.7293	**0.6841**	**0.7060**	0.7099	0.7951	0.0922
Methods	**HSOL**							**de Gibert**						
Tw-Star	0.8825	0.5778	0.3106	0.4040	0.3926	0.7204	0.1188	0.6118	0.3292	0.3511	0.3393	0.4817	0.758	0.0929
CNN_LSTM	0.8327	0.4841	0.305	0.3742	0.3883	0.7697	0.1417	0.575	0.4261	0.1929	0.2656	0.3601	0.7833	0.1221
CNN_GRU	**0.8846**	0.4971	0.3202	0.3895	0.3904	0.7531	0.1451	0.5504	0.4852	0.1845	0.2679	0.3808	0.701	0.0984
ToxDectRoBERTa	0.8425	0.6385	0.4205	0.5071	0.5965	0.8211	0.0881	**0.705**	**0.5106**	0.3247	0.397	0.4967	0.7818	**0.0775**
BERT-HateXplain	0.8221	0.5698	0.4265	0.4878	0.5672	0.8075	0.0686	0.6153	0.4875	0.315	0.3827	0.4769	0.7215	0.1119
HateBERT	0.866	0.6706	0.4833	0.5617	0.6184	**0.8439**	**0.0549**	0.6825	0.4988	0.3346	0.4005	0.5366	0.7484	0.0814
fBERT	0.8819	**0.7571**	0.5268	**0.6213**	**0.6866**	0.8239	0.0155	0.675	0.4823	**0.3541**	**0.4084**	**0.5505**	**0.7868**	0.0825
SAAT(Ours)	0.8275	0.7322	**0.5275**	0.6132	0.6746	0.8183	0.0599	0.655	0.4974	0.3245	0.3928	0.5296	0.7542	0.0952
Methods	**GHC**							**HateXplain**						
Tw-Star	0.6825	0.3672	0.2708	0.3117	0.3373	0.7374	0.1306	0.6575	0.5872	0.6384	0.6117	0.6868	0.8242	0.1271
CNN_LSTM	0.7515	0.5071	0.1785	0.2641	0.2805	0.7235	0.1234	0.7772	0.6388	0.565	0.5996	0.7131	0.8344	0.1122
CNN_GRU	0.7535	0.5182	0.1409	0.2216	0.3202	0.7593	0.1083	0.695	0.6145	0.5045	0.5541	0.6836	0.8543	0.1366
ToxDectRoBERTa	**0.7925**	0.5271	0.5113	**0.5191**	**0.6002**	0.7919	0.1038	0.7725	0.6237	0.6111	0.6173	0.7463	0.8715	0.1283
BERT-HateXplain	0.7665	0.5287	0.395	0.4522	0.5387	0.7748	**0.089**	0.7339	0.6533	0.5575	0.6016	0.7224	0.8684	0.1437
HateBERT	0.7875	**0.5388**	0.4354	0.4816	0.4962	**0.8061**	0.0954	**0.7753**	0.6654	0.6449	0.655	**0.7732**	**0.8915**	0.1146
fBERT	0.7645	0.5242	0.3933	0.4494	0.4817	0.804	0.0583	0.755	0.6364	0.634	0.6352	0.7514	0.8685	0.1293
SAAT(Ours)	0.7325	0.4908	**0.5366**	0.5096	0.5127	0.7864	0.1228	0.7525	**0.6988**	**0.6756**	**0.687**	0.7651	0.8818	**0.0988**
Methods	**Latent Hatred**							**Toxigen**						
Tw-Star	0.5151	0.3931	0.04	0.0726	0.3737	0.5687	**0.0048**	0.6875	0.6712	0.735	0.7017	0.6868	0.7006	0.0646
CNN_LSTM	0.5925	0.4242	0.365	0.3924	0.5458	0.6326	0.1722	0.6225	0.6522	0.525	0.5817	0.6189	0.6803	0.1263
CNN_GRU	0.5725	0.4593	0.3445	0.3937	0.5382	0.6268	0.1349	0.6575	0.6632	0.636	0.6493	0.6574	0.6923	0.0962
ToxDectRoBERTa	0.7175	0.5884	0.6115	0.5997	0.6459	0.6403	0.1287	0.5975	0.5759	0.742	0.6484	0.5892	0.6701	0.1315
BERT-HateXplain	0.6835	0.5515	**0.6375**	0.5914	0.6228	0.6233	0.1046	0.6275	0.6197	0.658	0.6383	0.6271	0.7123	0.1661
HateBERT	0.7255	0.6014	0.5925	0.5969	**0.6543**	**0.6796**	0.1105	**0.75**	**0.7224**	0.713	**0.7177**	**0.7244**	**0.7464**	0.0798
fBERT	0.7125	0.6061	0.5773	0.5913	0.5622	0.6564	0.1275	0.7025	0.6866	**0.745**	0.7146	0.7202	0.7252	**0.0633**
SAAT(Ours)	**0.735**	**0.6118**	0.6308	**0.6212**	0.6444	0.6732	0.0985	0.703	0.6747	0.701	0.6876	0.7077	0.7151	0.0917

Table B.6 Bias and Explainability Performance Evaluation on HateXplain

Methods	Regular						Bias			Explainability		
	acc	*prec*	*rec*	*f1*	*maf1*	*auc*	*gmb-sub*	*gmb-bpsn*	*gmb-bpsp*	*iou-f1*	*auprc*	*compre*
Tw-Star	0.7904	0.6403	0.7326	0.6834	0.7633	0.8433	0.6842	0.6622	0.7142	-	-	-
CNN_LSTM	0.8443	0.789	0.6768	0.7286	0.8097	0.878	0.7645	0.7063	0.738	-	-	-
CNN_GRU	0.8378	0.7844	0.6547	0.7137	0.8003	0.8774	0.7623	0.6997	0.7336	-	-	-
ToxDectRoBERTa	0.8521	**0.8145**	0.6747	0.7381	0.8175	0.9093	**0.7822**	0.7132	**0.7495**	0.126	0.598	0.683
BERT-HateXplain	0.8596	0.7803	0.7589	0.7695	0.8343	0.913	0.7781	0.7066	0.7396	0.13	0.603	0.688
HateBERT	0.8658	0.7954	0.7611	0.7778	0.8408	0.917	0.7641	**0.7144**	0.742	0.116	0.606	0.687
fBERT	0.8671	0.8113	0.7421	0.7752	0.8404	0.9114	0.7726	0.7046	0.7337	0.123	0.602	0.694
SAAT (Ours)	**0.8684**	0.7705	**0.7983**	**0.7842**	**0.8499**	**0.918**	0.7658	0.7084	0.7341	**0.144**	**0.614**	**0.702**

Note: *gmb-sub* represents the Area Under the Receiver Operating Characteristic (AUROC) value for subgroup metrics bias. *gmb-bpsn* denotes the *gmb-auc* value for the background positive subgroup negative, while *gmb-bpsp* indicates the *gmb-auc* value for the background positive subgroup positive. *iou-f1* refers to the Intersection-Over-Union (IOU) F1 score. *aupre* represents the Area Under the Precision-Recall Curve, and *compre* signifies the comprehensiveness metric. The attention-based mechanism is utilized to accurately identify significant words in the input sentences; therefore, we do not present the explainability values for the three other models.

overlap is quantified with *lr* (Lexical Richness), *ttr* (Type Token Ratio), and *hko* (Hateful Keyword Overlap), reflecting vocabulary diversity and toxic markers (Chowdhury et al. 2019). Adversarial phrasing is assessed through *nf* (Negation Frequency), *cf* (Contrast Frequency), and *ss* (Strong Sentiment), with *ss* computed using VADER sentiment scores (Hutto and Gilbert 2014). Metrics are extracted via SpaCy's "*en_core_web_sm*" pipeline, and *hko* relies on a curated lexicon covering commonly used categories of hateful expressions. These indicators collectively capture structural, lexical, and pragmatic characteristics that may influence hateful or adversarial content.

We report the performance of each dataset on nine corpus-level characteristics in Table B.8, with anomalous values highlighted in bold. Toxigen exhibits a longer average sentence length than the other datasets, while its remaining syntactic complexity measures are comparable. In contrast, it shows pronounced deviations in lexical overlap, with substantially lower lexical richness, type–token

Table B.7 Regular and Robust Performance Across Initialization Conditions

Panel A: HateXplain

Regular Performance	acc	$prec$	rec	$f1$	$maf1$	auc
ToxDectRoBERTa	0.8506±0.68	0.8086±1.16	0.6526±1.46	0.7195±1.07	0.8069±0.93	0.9112±0.88
BERT-HateXplain	0.8479±0.72	0.7789±1.04	0.7623±0.77	0.7606±0.87	0.8288±0.82	0.9086±0.79
HateBERT	0.8466±0.76	0.7883±0.87	0.7465±1.12	0.7697±0.93	0.8356±0.86	0.9119±0.82
fBERT	0.8559±0.89	**0.8106**±0.75	0.7433±0.72	0.7772±0.8	**0.841**±0.91	0.9106±0.66
SAAT(Ours)	**0.8676**±0.83	0.7722±0.47	**0.7855**±0.98	**0.7785**±0.52	0.8398±0.97	**0.9142**±0.95
Robust Performance	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*
ToxDectRoBERTa	0.5203±0.43	0.6116±0.53	0.1285±0.47	0.2180±0.54	0.4288±0.46	0.7444±0.46
BERT-HateXplain	0.5101±0.67	0.5342±0.46	0.0771±0.11	0.1258±0.2	0.3422±0.49	0.6816±0.57
HateBERT	0.5529±0.59	0.6493±0.51	0.2334±0.18	0.3365±0.43	0.5039±0.44	0.7416±0.52
fBERT	0.6134±0.57	0.6897±0.52	0.3636±0.25	0.4755±0.25	0.5688±0.64	0.8122±0.65
SAAT(Ours)	**0.8085**±0.7	**0.7007**±0.52	**0.7167**±0.67	**0.71**±0.59	**0.7878**±0.66	**0.8908**±0.64

Panel B: Latent Hatred

Regular Performance	acc	$prec$	rec	$f1$	$maf1$	auc
ToxDectRoBERTa	0.7526 ± 0.75	0.6193 ± 1.12	0.6648 ± 0.69	0.6389 ± 0.72	0.7144 ± 0.63	**0.8222** ± 0.95
BERT-HateXplain	0.7224 ± 0.61	0.5665 ± 0.46	0.6672 ± 0.49	0.6069 ± 0.55	0.7066 ± 0.43	0.8074 ± 0.35
HateBERT	0.7428 ± 0.53	0.6076 ± 0.56	0.6389 ± 0.74	0.6122 ± 0.61	0.7159 ± 0.59	0.7966 ± 0.41
fBERT	0.7556 ± 0.61	0.6122 ± 0.48	0.6744 ± 0.83	0.6126 ± 0.44	0.7299 ± 0.58	0.8144 ± 0.68
SAAT(Ours)	**0.7596** ± 0.62	**0.634** ± 0.59	**0.6813** ± 0.72	**0.6567** ± 0.55	**0.7385** ± 0.37	0.8128 ± 0.54
Robust Performance	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*
ToxDectRoBERTa	0.5489 ±0.66	0.3199 ±0.38	0.2463 ±0.18	0.2755 ±0.23	0.4843 ±0.57	0.6136 ±0.67
BERT-HateXplain	0.6075 ±0.78	0.4112 ±0.4	0.3865 ±0.39	0.4026 ±0.38	0.5545 ±0.59	0.6223 ±0.72
HateBERT	0.4941 ±0.78	0.1942 ±0.14	0.1398 ±0.17	0.1666 ±0.29	0.4038 ±0.49	0.5814 ±0.43
fBERT	0.5029 ±0.65	0.2862 ±0.21	0.2886 ±0.25	0.2865 ±0.18	0.4556 ±0.63	0.6011 ±0.59
SAAT(Ours)	**0.6499** ±0.38	**0.5714** ±0.31	**0.6146** ±0.44	**0.5922** ±0.3	**0.6669** ±0.46	**0.7747** ±0.47

Panel C: HSOL

Regular Performance	acc	$prec$	rec	$f1$	$maf1$	auc
ToxDectRoBERTa	0.9164 ± 3.82	0.7554 ± 4.15	0.7246 ± 5.31	0.7288 ± 4.51	0.7466 ± 3.83	0.8588 ± 3.46
BERT-HateXplain	0.9215 ± 3.51	0.7277 ± 3.93	0.7199 ± 5.13	0.6915 ± 4.11	0.7601 ± 3.94	0.8597 ± 3.77
HateBERT	0.9209 ± 3.56	0.7645 ± 3.82	0.7122 ± 4.82	0.742 ± 4.14	0.7772 ± 4.1	0.8623 ± 3.52
fBERT	0.9221 ± 3.62	0.7428 ± 3.8	0.7429 ± 5.45	0.7456 ± 4.28	0.7879 ± 4.26	0.8568 ± 3.72
SAAT(Ours)	**0.9322** ± 3.67	**0.8036** ± 4.14	**0.7825** ± 5.53	**0.7746** ± 4.36	**0.8133** ± 4.45	**0.8767** ± 3.87
Robust Performance	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*
ToxDectRoBERTa	0.5949 ± 2.46	0.5206 ± 2.56	0.4986 ± 1.98	0.5120 ± 2.48	0.6277 ± 2.88	0.7452 ± 2.07
BERT-HateXplain	0.5859 ± 2.14	0.4774 ± 1.88	0.5298 ± 2.06	0.5006 ± 2.12	0.5551 ± 2.04	0.7266 ± 1.87
HateBERT	0.5693 ± 1.96	0.4726 ± 2.03	0.4822 ± 2.25	0.4888 ± 2.45	0.5406 ± 2.17	0.6972 ± 1.81
fBERT	0.5726 ± 1.94	0.4825 ± 2.11	0.4572 ± 2.24	0.4711 ± 2.29	0.5079 ± 1.98	0.7466 ± 1.79
SAAT(Ours)	**0.8088** ± 2.42	**0.7644** ± 2.43	**0.7405** ± 2.59	**0.7835** ± 2.67	**0.7516** ± 2.48	**0.8325** ± 2.39

Panel D: OLID

Regular Performance	acc	$prec$	rec	$f1$	$maf1$	auc
ToxDectRoBERTa	**0.8589**±2.46	0.7944±3.36	0.7831±3.45	0.7763±4.41	**0.8154**±3.22	0.8564±2.27
BERT-HateXplain	0.8357±1.96	0.8026±3.49	0.7755±4.21	0.7774±3.67	0.8017±3.35	0.8246±2.85
HateBERT	0.8454±1.93	0.7965±3.44	0.7922±4.44	0.7856±4.26	0.7966±3.75	0.8399±3.03
fBERT	0.8526±2.06	**0.8046**±3.54	0.7959±3.89	0.8042±4.07	0.8123±4.15	**0.8522**±2.94
SAAT(Ours)	0.8244±1.89	0.8021±3.66	**0.8136**±3.47	**0.8147**±3.57	0.7955±3.67	0.8379±2.94
Robust Performance	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*
ToxDectRoBERTa	0.6072 ± 1.55	0.4278 ± 2.45	0.3745 ± 2.06	0.3985 ± 2.48	0.4266 ± 2.29	0.7056 ± 2.05
BERT-HateXplain	0.6397 ± 2.21	0.4406 ± 2.02	0.3979 ± 2.48	0.4288 ± 2.66	0.4172 ± 1.87	0.6313 ± 2.32
HateBERT	0.6466 ± 2.45	0.4179 ± 2.59	0.3668 ± 2.54	0.3828 ± 2.41	0.4049 ± 2.35	0.6897 ± 2.44
fBERT	0.6502 ± 2.6	0.4494 ± 2.75	0.3976 ± 2.67	0.3933 ± 2.78	0.4077 ± 2.53	0.6972 ± 2.08
SAAT(Ours)	**0.8174** ± 2.77	**0.7555** ± 3.13	**0.765** ± 2.9	**0.7476** ± 3.41	**0.7116** ± 3.12	**0.7288** ± 2.46

Note: Standard deviations have been scaled by 100 for easier presentation.

Table B.8 Comparative Analysis of Corpus-Level Linguistic Features Across Hate Speech Datasets

Methods	**Syntactic Complexity**			**Lexical Overlap**			**Adversarial Phrasing**		
	td	asl	awl	lr	hko	ttr	nf	cf	ss
HSOL	3.119	8.77	3.953	0.135	0.286	0.133	0.072	0.045	0.424
OLID	3.279	8.83	3.996	0.116	0.214	0.147	0.096	0.046	0.255
Latent Hatred	3.34	8.93	4.005	0.084	0.235	0.086	0.103	0.038	0.267
Toxigen	3.822	**11.23**	3.825	**0.034**	**0.113**	**0.032**	**0.151**	0.043	0.254

ratio, and hateful keyword overlap, indicating a strong reliance on repetitive and common vocabulary. It also displays an unusually high negation frequency, suggesting that hateful intent is often expressed implicitly (Rice et al. 2020).

Table B.9 Robust Performance of Adversarial Training Methods Under White-Box Attacks at Word-level

Methods	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr \downarrow$	*training time*
Panel A: HSOL								
Regular	0.5826	0.514	0.4898	0.5016	0.5156	0.6135	0.7267	15min
FGSM-based	0.5336	0.5251	0.0516	0.0939	0.2579	0.5846	0.5166	28min
PGD-based	0.6245	0.5819	0.0850	0.1483	0.3479	0.5915	0.354	158min
TRADES-based	0.6366	0.6178	0.0745	0.133	0.3561	0.6188	0.3513	161min
AVmixup-based	0.6457	0.6453	0.1227	0.2062	0.389	0.6257	0.3505	158min
SAAT(Ours)	**0.8745**	**0.8044**	**0.7748**	**0.7893**	**0.7905**	**0.8434**	**0.1593**	158min
Panel B: Latent Hatred								
Regular	0.5227	0.3104	0.272	0.2899	0.4532	0.5813	0.3466	17min
FGSM-based	0.4997	0.3683	0.097	0.1536	0.3122	0.5145	0.2842	32min
PGD-based	0.576	0.4055	0.1332	0.2005	0.3379	0.5587	0.2673	162min
TRADES-based	0.5921	0.419	0.1344	0.2035	0.341	0.5588	0.2768	166min
AVmixup-based	0.5946	0.4237	0.1659	0.2384	0.3439	0.5623	0.2664	162min
SAAT(Ours)	**0.6506**	**0.5682**	**0.6147**	**0.5905**	**0.6711**	**0.7885**	**0.2121**	163min
Panel C: Toxigen								
Regular	0.3898	0.3611	0.296	0.3253	0.3681	0.4308	0.4539	182min
FGSM-based	0.4978	0.4213	0.356	0.3855	0.3906	0.5253	0.3846	415min
PGD-based	0.5568	0.4677	0.364	0.426	0.4811	0.6349	0.3481	1963min
TRADES-based	0.5702	0.4826	0.373	0.4233	0.4835	0.6288	0.3492	2011min
AVmixup-based	**0.5921**	0.4867	0.38	0.4246	**0.4856**	**0.6411**	0.3489	1972min
SAAT(Ours)	0.5766	**0.4872**	**0.383**	**0.4289**	0.4838	0.6405	**0.3467**	1966min

Table B.10 Regular and Robust Performance of Aggregation Methods under Word-level Attacks

Methods	**Regular Performance**						**Robust Performance**						
	acc	*prec*	*rec*	*f1*	*maf1*	*auc*	acc^*	$prec^*$	rec^*	$f1^*$	$maf1^*$	auc^*	$asr \downarrow$
Panel A: HateXplain													
SAAT−0	0.8226	0.6963	0.7241	0.7099	0.7729	0.8867	0.6097	0.6766	0.3941	0.4981	0.5956	0.8145	0.3872
Attention	0.8158	0.699	0.7287	0.7135	0.7829	0.8659	0.6144	0.6678	0.4035	0.5031	0.6064	0.8274	0.3744
Center Loss	0.832	0.7268	0.7123	0.7195	0.7895	0.9002	0.6457	0.6845	0.4472	0.541	0.6444	0.8456	0.3499
SAAT	0.8689	0.7872	0.7936	0.7904	0.8112	0.9178	0.8008	0.6824	0.7108	0.6963	0.7712	0.8824	0.1952
Panel B: Latent Hatred													
SAAT−0	0.7401	0.5556	0.6247	0.5881	0.6744	0.7759	0.496	0.4878	0.1889	0.2723	0.4344	0.6478	0.3778
Attention	0.7477	0.5642	0.6207	0.5911	0.6663	0.7823	0.5355	0.4896	0.2025	0.2865	0.4466	0.689	0.3656
Center Loss	0.7588	0.5964	0.5855	0.5909	0.6987	0.8056	0.5668	0.5214	0.2674	0.3535	0.4514	0.7014	0.3342
SAAT	0.7616	0.6279	0.6888	0.6569	0.7601	0.8157	0.6478	0.5663	0.5964	0.581	0.6669	0.7789	0.2199
Panel C: OLID													
SAAT−0	0.7444	0.6225	0.6578	0.6397	0.6775	0.7678	0.6122	0.6277	0.4556	0.528	0.5762	0.6393	0.2998
Attention	0.7458	0.6355	0.6598	0.6474	0.7001	0.7883	0.6364	0.6354	0.4889	0.5526	0.606	0.6782	0.2986
Center Loss	0.7789	0.6823	0.6214	0.6504	0.7241	0.8049	0.6566	0.6836	0.4769	0.5618	0.6277	0.7054	0.2646
SAAT	0.8089	0.7966	0.8012	0.7989	0.7862	0.8221	0.7989	0.7457	0.7226	0.734	0.699	0.7225	0.2244
Panel D: Toxigen													
SAAT−0	0.7596	0.7885	0.7299	0.758	0.7404	0.8303	0.5794	0.4814	0.3619	0.4132	0.4845	0.6376	0.3571
Attention	0.7542	0.7876	0.7221	0.7534	0.7255	0.8289	0.5833	0.4848	0.3688	0.4189	0.4904	0.6415	0.3524
Center Loss	0.7456	0.7458	0.7186	0.7319	0.7259	0.8216	0.5676	0.4794	0.3651	0.4145	0.4878	0.6185	0.3546
SAAT	0.7676	0.7634	0.7119	0.7368	0.7288	0.8354	0.5827	0.4769	0.3649	0.4135	0.4768	0.6355	0.3556

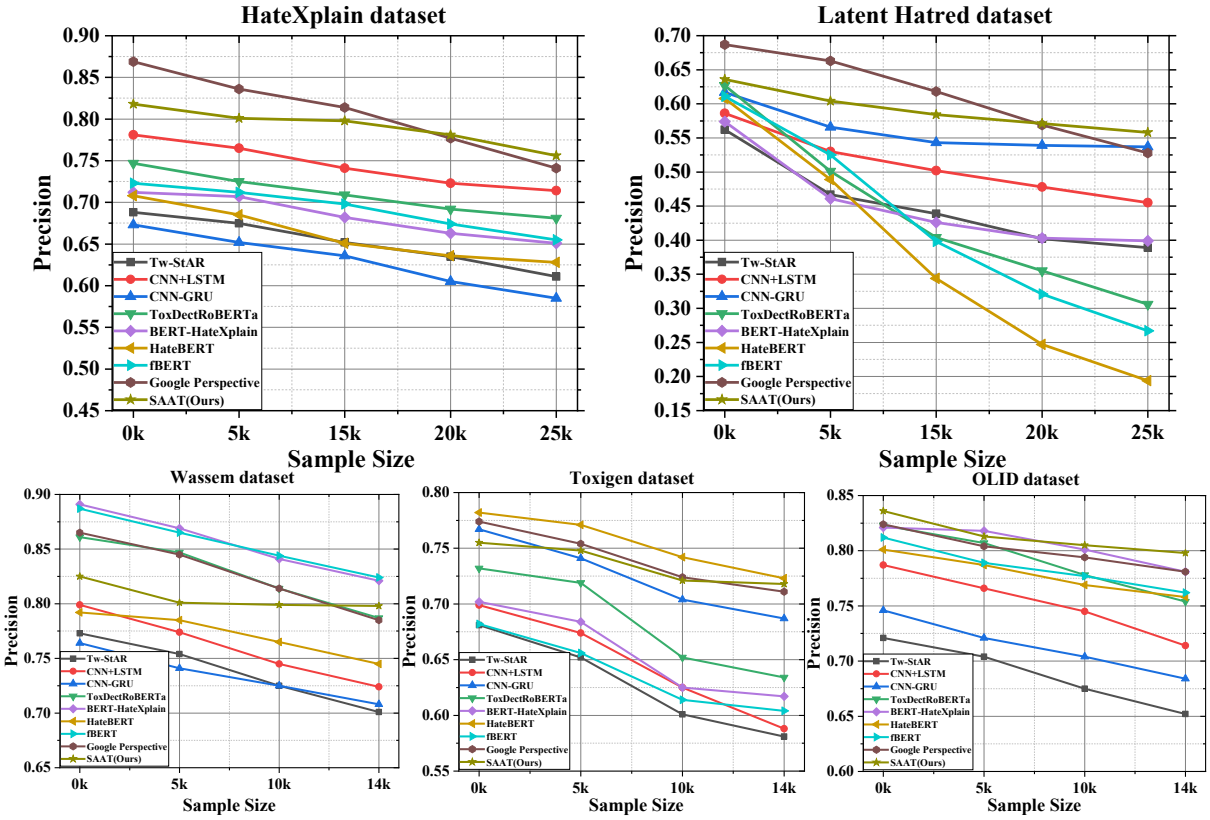

Figure B.1 **Robust Performance under Black-box Attacks with Varied Sample Size**

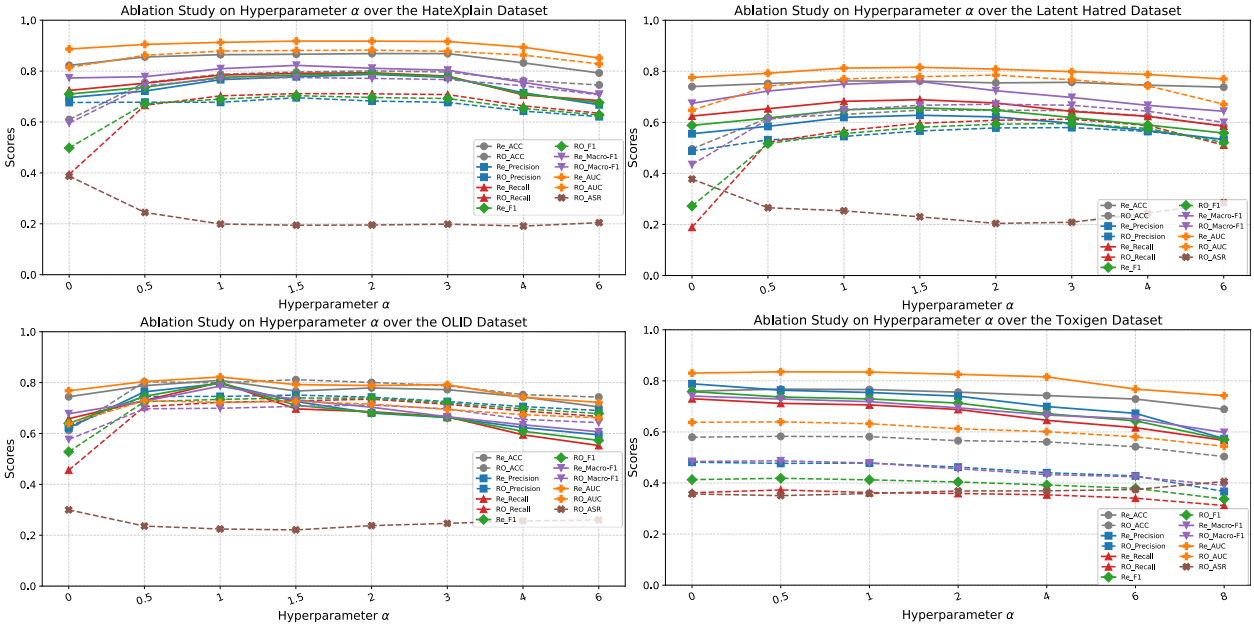

Figure B.2 **Ablation Study under Varying Hyperparameters**

References

Bartlett PL, Mendelson S (2002) Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3(Nov):463–482.

Chowdhury SA, Stepanov EA, Danieli M, Riccardi G (2019) Automatic classification of speech overlaps: feature representation and algorithms. *Computer Speech & Language* 55:145–167.

Dobriban E, Hassani H, Hong D, Robey A (2023) Provable tradeoffs in adversarially robust classification. *IEEE Transactions on Information Theory* 69(12):7793 – 7822.

Hutto C, Gilbert E (2014) VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 216–225 (AAAI, Palo Alto, CA).

Pauli P, Koch A, Berberich J, Kohler P, Allgöwer F (2021) Training robust neural networks using lipschitz bounds. *IEEE Control Systems Letters* 6:121–126.

Rice L, Wong E, Kolter Z (2020) Overfitting in adversarially robust deep learning. *Proceedings of the 37th International Conference on Machine Learning*, 8093–8104 (PMLR).

Wang Y, Buschmeier H (2024) Revisiting the phenomenon of syntactic complexity convergence on german dialogue data. *Proceedings of the 20th Conference on Natural Language Processing*, 75–80 (ACL, Cedarville, OH).

Xu H, Mannor S (2012) Robustness and generalization. *Machine learning* 86(3):391–423.

Yin D, Kannan R, Bartlett P (2019) Rademacher complexity for adversarially robust generalization. *Proceedings of the 36th International conference on machine learning*, 7085–7094 (PMLR).