Why models fail? An Evaluation Framework to Characterize Model-Dataset Interactions

Anonymous ACL submission

Abstract

Datasets escape accountability while models face relentless evaluation. We introduce a framework that inverts this paradigm, subjecting 21 hate speech datasets to systematic scrutiny across robustness, fairness, and explainability – revealing which training corpora produce trustworthy models and why most fail. Our framework treats datasets as the primary experimental variable: each corpus undergoes cross-domain stress testing (generalization to 20 holdout distributions), demographic audit (classification parity across protected groups), and explanation reliability assessment (post-hoc explainability faithfulness in distribution and under distribution shift). We report cross-corpus generalization failures despite shared semantics and quantify pervasive demographic biases with the majority of datasets scoring near random chance. We demonstrate that post-hoc explanation faithfulness varies substantially with training distributions, independent of attribution method or architecture, and produces contradictory responses under shift.

1 Introduction

Data quality determines AI system performance more than algorithmic sophistication. Facial detection systems broke through performance barriers not through computational advances but through access to massive training corpora that enabled robust feature learning (Torralba and Efros, 2011). Despite increasing attention to data's role in model performance, a fundamental question persists: *How do differences between datasets targeting the same task manifest in model behavior across multiple dimensions of trustworthiness?*

Robustness to distribution shift, fairness across demographic groups, and faithfulness of post-hoc explanations are increasingly recognized as critical desiderata for trustworthy machine learning (OECD, 2024). Yet these properties remain siloed in research. No existing work jointly quantifies how training data composition affects all three dimensions simultaneously: *Do dataset properties that improve generalization also promote fairness? Does high accuracy guarantee faithful explanations under distribution shift? Does model architecture or dataset choice have greater impact on cross-domain generalization?* These questions are critical because distribution shifts in deployment can invalidate fairness guarantees established on in-distribution data while simultaneously compromising explanation reliability that practitioners depend on for regulatory compliance and debugging. To the best of our knowledge, this represents the first investigation into how training data shape explanations faithfulness when evaluated under distribution shift and in conditions of model degradation.

We address these questions through a unified evaluation framework applied to 21 publicly available hate speech datasets from MetaHate (Piot et al., 2024), spanning diverse platforms, annotation schemes, and targeted domains. Hate speech detection provides an ideal testbed – systems exhibit poor cross-corpus generalization despite shared semantic spaces, systematic demographic disparities, and opaque decision boundaries (Arango et al., 2019; Davidson et al., 2019). Our framework measures robustness through cross-domain performance metrics, fairness through AUC-based classification parity across demographic subgroups, and explainability through Sufficiency and Comprehensiveness of post-hoc explanations. We selectively apply TRAK data attribution (Park et al., 2023) to trace failures back to influential training examples.

Contributions. Our main contributions in this work are the following: (**i**) We introduce a framework jointly examining robustness, fairness, and explainability as interconnected properties shaped by training data §3. (**ii**) We subject 21 hate speech datasets and four model architectures to systematically evaluate models across three trustworthi-

ness dimensions. (**iii**) We demonstrate complex interdependencies: fairness and accuracy decouple, explainability sometimes inversely correlates with accuracy under shift, and architectural upgrades can raise AUC but rarely improve fairness.

2 Background

Data Impact. Machine learning research has increasingly shifted towards data as a central determinant of model performance. This includes theoretical studies on how data distributions shape learning and generalization (Adebayo et al., 2018; Arpit et al., 2017; Badjatiya et al., 2017; Jiang et al., 2019; Yang et al., 2022, 2024), as well as their role in model fairness and bias (Dwork et al., 2012; Feldman et al., 2015; Hardt et al., 2016; Romei and Ruggieri, 2014; Zliobaite, 2015). In post-hoc explainability, however, the role of data has been largely overlooked, with Ribeiro et al. (2021) providing the only notable investigation. The growing recognition of data's impact has spurred data-centric ML methodologies (DMLR, 2024), leading to advances in dataset construction (Almohaimeed et al., 2023; Mosquera Gómez et al., 2023; Pingle et al., 2023; Shinde et al., 2024), applications in new domains (Arnaiz-Rodriguez and Oliver, 2024; Deng and Ma, 2024; Kohli et al., 2024; Vysogorets and Kempe, 2024), and ethical debates around data governance in AI (Janssen et al., 2020). We present the first integrated analysis of generalization, fairness, and explainability across NLP datasets, bridging previously siloed research directions.

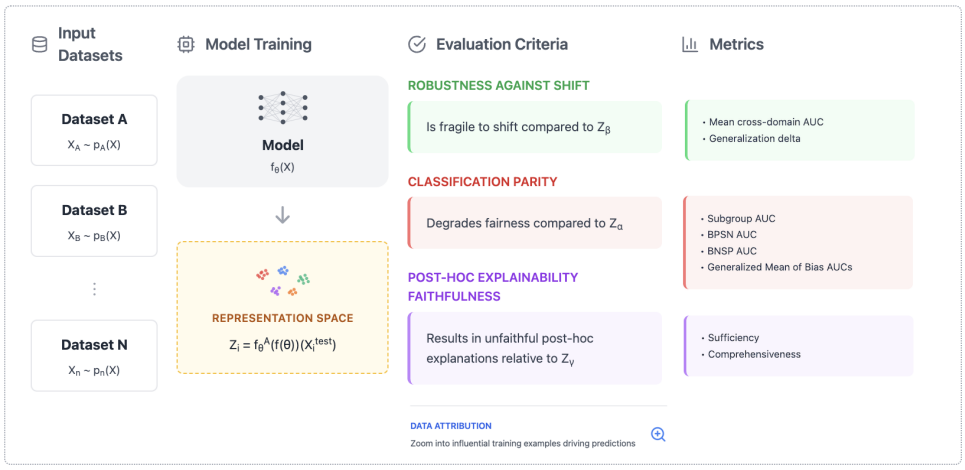
Hate Speech Detection. Hate speech detection research has prioritized isolated dataset performance over generalization, fairness, and explainability, with academic evaluations diverging from real-world distributions and producing poor generalization (Tonneau et al., 2024). Cross-dataset evaluation studies have consistently demonstrated that models trained on individual datasets fail to transfer across different data sources (Bourgeade et al., 2023; Antypas and Camacho-Collados, 2023; Khurana et al., 2024; Fortuna et al., 2021), exhibiting performance degradation attributable to the learning of dataset-specific artifacts and annotation biases. Compounding these generalization failures, models systematically misclassify identity mentions as hateful content, resulting in disparate false positive rates across protected demographic groups (Garg et al., 2023). Yoder et al. (2022) investigate patterns in hate speech datasets structuring variation by targeted identity, finding that classifiers fail to generalize across different demographic groups. Post-hoc explainability methods have been increasingly applied to hate speech detection models to understand decision-making processes and identify potential biases (Mehta and Passi, 2022). However, no work has investigated how training data composition affects the quality of post-hoc explanations when evaluated in and out-of-distribution, either in hate speech detection or in other NLP tasks.

3 Evaluation Framework

Prior work in evaluation has largely remained fragmented: robustness, fairness, and explainability are usually studied in isolation, with no unifying methodology to understand how training data shapes these outcomes. In this section, we introduce a framework that jointly examines these dimensions, treating them as complementary lenses on how dataset properties drive model behavior. Our framework consists of three components (Figure 1): (i) robustness under shift in §3.1 (mean cross-domain AUC, generalization delta), (ii) classification parity across demographic subgroups in §3.2 (Subgroup AUC, BPSN AUC, BNSP AUC, Generalized Mean of Bias AUCs), and (iii) faithfulness of generated post-hoc explanations in §3.2.2 (Sufficiency and Comprehensiveness). To trace specific failure modes back to training examples, we selectively apply TRAK (Park et al., 2023), a data attribution method that identifies which training samples most influenced a given prediction. While the specific choice of models and metrics can vary, the key contribution is methodological: these components together provide a holistic view of when failures originate from training data, the model, or their interaction.

Datasets. We use 21 hate speech datasets from MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection (Piot et al., 2024). This selection provides representative coverage of the hate speech detection landscape, spanning diverse platforms such as Twitter (Basile et al., 2019; Zampieri et al., 2019), Reddit (Kurrek et al., 2020), Gab (Kennedy et al., 2022), and Wikipedia (Wulczyn et al., 2016); varied annotation approaches; and multiple targeted domains, including gender-based (Basile et al., 2019), political (Grimminger and Klinger, 2021), racial (de Gibert et al., 2018), and cyber-bullying content (Bin Abdur Rakib and Soon, 2018). Piot et al. (2024) standardized all datasets

Figure 1: Overview of the evaluation framework. We train models on individual datasets and generate their respective representation spaces, which are then evaluated across robustness against shift, classification parity, and post-hoc explainability faithfulness. TRAK (Park et al., 2023) is used to identify influential training examples driving model predictions.



into a binary hate vs. non-hate classification. Table 2 in Appendix A presents a description of each dataset used in the study.

Models. Due to the high performance of BERT-based models on hate speech classification (Samghabadi et al., 2020; Mozafari et al., 2019), we trained and evaluated DistilBERT (DB; Sanh, 2019), RoBERTa (RB; Liu et al., 2019), and DeBERTaV3 (DeB; He et al., 2021) from Hugging Face. All three models were fine-tuned for 3 epochs using the AdamW optimizer (Loshchilov et al., 2017), with a learning rate of 2e-5. Note that state-of-the-art models contain billions of parameters; computational constraints prevented experimentation with larger models. As a baseline, we evaluated a logistic regression model with TF-IDF over the entire vocabulary. All models were trained using an 80/20 train-test split. To ensure apples-to-apples comparisons, we focused on models compatible with uniform post-hoc explainability methods. While model-agnostic by design, the framework benefits from diverse models for validation.

3.1 Robustness against Distribution Shift

Machine learning models operate under the closed-world assumptions that the training and inference regimes align. This premise rarely holds in deployment environments, where annotation processes are inherently constrained by incomplete domain expertise, systematic sampling biases, and finite coverage of the target distribution's support (Paullada et al., 2021). Curating datasets often involves multiple degrees of freedom (e.g. source selection, linguistic constraints, perspective samplings, and annotation demographics). Each of them can introduce model degradation: source selection can lead to domain mismatch, linguistic constraints may create artificial patterns that do not generalize, perspective sampling can embed unwanted correlations, and annotation demographics may encode biases in the ground truth. Hence, despite aiming to capture real-world phenomena, datasets become constrained snapshots of the represented field.

For each training distribution, we compute two complementary metrics: (a) the mean cross-domain performance, measured as the average model AUC across all out-domain test sets, and (b) the generalization delta, calculated as the difference between in-distribution test performance and mean cross-domain performance. In doing so, we quantify for each source training distribution, both the absolute cross-domain generalization capacity and the relative performance degradation under shift.

3.2 Classification Parity

A model's decision boundary is shaped by its training data, where negative examples implicitly define "the rest of the world" (Torralba and Efros, 2011). Non-representative sampling produces overconfident classifiers with poor discriminative power and can result in unfair treatment across demographic groups. We investigate how different training distributions affect model performance across demographic groups. For each source training distribution, we evaluate the resulting trained model using the comprehensive AUC-based metric suite developed by Borkan et al. 2019. The evaluation framework quantifies classification parity through: Subgroup AUC, Background Positive Subgroup Negative (BPSN) AUC, Background Negative Subgroup Positive (BNSP) AUC, Generalized Mean of Bias AUCs (GMB). A detailed description of these

metrics can be found in Appendix B. The models are evaluated on the grounds of how much they are able to reduce the unintended bias towards a target community. We evaluate on the Jigsaw Unintended Bias dataset (Xiao et al.), which explicitly labels demographic groups mentioned in each comment.

3.2.1 Data Attribution

Beyond measuring outcomes, we also consider how failures can be traced back to specific training data. Recent advances in data attribution, such as TRAK (Park et al., 2023), provide scalable methods to identify which training examples most influenced a model's predictions. While our robustness and fairness evaluations show what failures occur, we complement them by linking model behavior to its training data to explain why they arise. For details on this method we defer to the original paper.

3.2.2 Post-hoc Explainability Faithfulness

Post-hoc explanations are widely used to interpret model predictions, but their reliability varies with model architecture, training configuration, and explanation method. Recent work shows that the same explanation method can produce different results across architectures, and a single model can report inconsistent attributions across different explainability methods (Alvarez-Melis and Jaakkola, 2018; Yeh et al., 2019; Kindermans et al., 2019). Despite the crucial role of training data in shaping model representations, its influence on the faithfulness of explanations remains underexplored.

We focus on explanation faithfulness, whether explanations reflect model decision-making, and investigate how training distributions affect it. To this end, we assess faithfulness through input ablations – removing top-ranked tokens and observing the resulting changes in model predictions. We adopt the Sufficiency and Comprehensiveness metrics from the ERASER benchmark (DeYoung et al., 2019), which are widely used in explainability research (Mathew et al., 2020; Carton et al., 2020; Chan et al., 2022; Zhou and Shah, 2022; Wiegreffe and Marasović, 2021). These metrics measure how well the identified important features alone can sustain the prediction (Sufficiency), and how much the prediction changes when those features are removed (Comprehensiveness). Definitions and implementation details for both metrics are provided in Appendix C. To assess faithfulness across explanation granularities, we compute the metrics for the top $k\%$ tokens ($k = 10, 20, \ldots, 100$) following ERASER, then we average to obtain AOPC (DeYoung et al., 2019).

4 Results

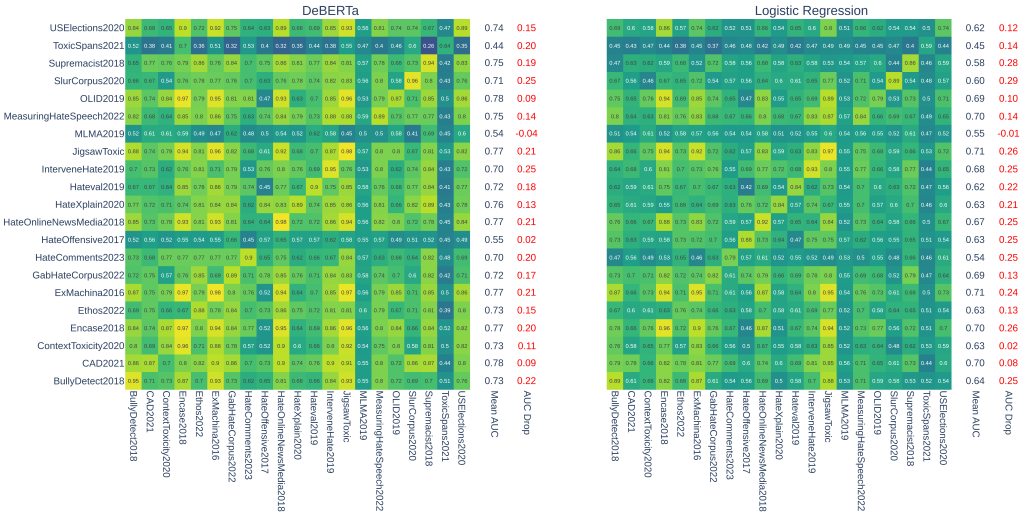
4.1 Robustness against Distribution Shift

We evaluate how well a model trained on one dataset generalizes on a representative set of other datasets. Figure 2 present the cross-dataset generalization performance for DeB and LR.

`CAD2021` and `OLID2019` show the strongest cross-dataset generalization. For `CAD2021`, performance drops 0.07-0.09 across models (DeB: 0.78, RB: 0.78, DB: 0.75, LR: 0.70). `OLID2019` shows similar robustness with drops of 0.09-0.12 (DeB: 0.78, RB: 0.78, DB: 0.76, LR: 0.69). A second tier of datasets, `Encase2018`, `ExMachina2016`, `JigsawToxic`, and `HateOnlineNewsMedia2018`, also demonstrate solid generalization as the AUC scores average in the range of 0.75-0.77 but the generalization drops are higher than 0.2. Despite `Encase2018` being 29 times larger (91,950 vs. 3,214 examples), both this dataset and `HateOnlineNewsMedia2018` achieve identical 0.70 mean AUC in transfer performance with DeB. In contrast, `ToxicSpans2021` performs and transfers poorly, especially for transformer models (e.g., DeB 0.44, drop 0.20; DB 0.39, drop 0.27). `SlurCorpus2020` and `Supremacist2018` similarly exhibit large transfer drops (e.g., DeB drops of 0.26 and 0.20, respectively), and `HateOffensive2017` shows inconsistent behavior: DB drops sharply (0.24), while RB does not. These patterns reflect their design: specialized domains or annotation schemes targeting explicit slurs capture narrower linguistic phenomena that generalize poorly to broader offensive language in other corpora. The findings are stable across models, indicating that dataset semantics, rather than architecture, primarily determine cross-dataset generalizability. `OLID2019`, `HateEval2019`, `HateSpeech2018`, and `JigsawToxic` function as strong hubs, both exporting and importing well, while datasets like `ToxicSpans2021` and `SlurCorpus2020` fail.

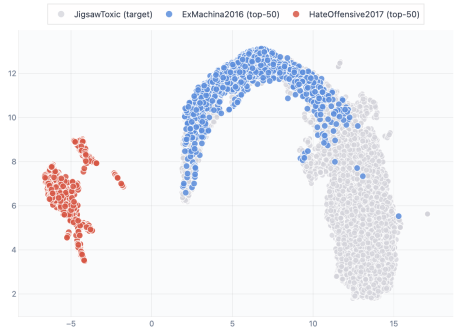
To probe why some datasets generalize better than others, we compute TRAK-based data attribution. TRAK produces, for each target test data point, a ranked list of source training points according to their contribution to the prediction. Thus, every target data point has its own "attribution profile" over the source dataset. We collect the top-K most influential points across all target test

Figure 2: Cross-dataset generalization between DeB (left) and LR (right), reporting AUC classification scores when training on one dataset (rows) and testing on another (columns). Diagonal cells represent in-domain performance, while off-diagonal cells indicate cross-domain generalization. Results for DB and RB are reported in Appendix D.



points, identifying the subset of the source data that the model most heavily relied on when transferring to the target. We compare these attribution profiles for two source datasets with contrasting generalization to the same target: DB trained on `ExMachina2016` vs. `HateOffensive2017` and evaluated on `JigsawToxic` (AUC: 0.96 vs. 0.72). From the UMAP visualization (Figure 3), `ExMachina2016` influential examples cluster within the `JigsawToxic`, while those from `HateOffensive2017` form a separate cluster, consistent with their contrasting AUC scores.

Figure 3: UMAP of TRAK-selected top-50 influential training examples from `ExMachina2016` (blue) and `HateOffensive2017` (red) on `JigsawToxic` test data (gray).



Observations. (1) Models trained on datasets with diverse offensive language phenomena transfer well to specialized domains, but not viceversa. (2) The source/sink hierarchy remains identical from LR to DeB, showing that linguistic coverage drives generalization. (3) Few hub corpora dominate performance.

4.2 Classification Parity

We assess model bias across demographic groups using the AUC-based metrics from Borkan et al. (2019); Table 1 reports the GMB scores.

The highest GMB scores are reached on `OLID2019` (DeB, 0.80; DB, 0.78; RB, 0.78), `HateOnlineNewsMedia2018` (RB, 0.75; DeB, 0.75), `Encase2018` (DeB, 0.75), and `ExMachina2016` (LR, 0.66). In some cases, the range of fairness values is wide (e.g. the `OLID2019` range is 0.17) while in other cases the range is very small (e.g. `ExMachina2016` range is 0.02). In some cases, fairness gains are substantial: for example, `OLID2019` improves from 0.63 with LR to 0.80 with DeB, and `USElections2020` rises from 0.55 (LR) to 0.63 (RB). In others, such as `ExMachina2016`, differences between models are minimal (0.65–0.67). The fairness impact of the data is most apparent with LR, which lacks pretraining and thus exposes dataset-induced bias more directly. In the majority of datasets, GMB values cluster between 0.50 and 0.65, regardless of their temporal origin, collection methodology, or annotation protocol. Even high with high AUC values (>0.85), fairness remains close to 0.50.

We zoom in RB trained on `GabHateCorpus2022` (GMB = 0.58; AUC = 0.76). A comment misclassified as non-toxic (prob = 0.57) reads: *"Mike – will you personally take responsibility for any terrorist attacks by Muslim refugees? If not, you need to stand down."* The presupposition that Muslim refugees *will* commit terrorist attacks encodes a

harmful stereotype, but the model misses it – revealing a blind spot for implicit hate beyond surface markers (slurs, profanity). The most TRAK-influential training points include *"Muslims. That's the greatest fraud ever."* (label = 0) and *"Die Jew die"* (label = 0) – mislabeled hate speech. Exposure to such errors might teach the model that hostile rhetoric toward minorities is acceptable. A neutral sentence, *"My neighbor is lesbian,"* is wrongly flagged as toxic (prob = 0.87). All top-5 influential training examples involve sexuality and gender in toxic contexts, despite no explicit references to LGBTQ+ identities.

> **Observations.** (1) High predictive performance can coexist with fairness collapse. (2) Upgrading from linear models to large transformers raises AUC but rarely shifts GMB beyond the 0.50-0.65 band, suggesting diminishing returns without dataset refinement. (3) Training on mislabeled examples and imbalanced identity contexts teaches models to overlook implicit hate while flagging neutral identity mentions.

4.3 Post-hoc Explainability Faithfulness

In-domain faithfulness of post-hoc explanations. Figure 4 presents a comparative analysis of SHAP and LIME explanations through Sufficiency and Comprehensiveness metrics. LIME provides more faithful explanations than SHAP across models, reporting stronger Comprehensiveness (0.31 vs. 0.28) and slightly better Sufficiency on average. We observe that explanation faithfulness varies substantially across training distributions, even when holding model architecture and explanation method constant. When using LIME on RoBERTa, `HateOnlineNewsMedia2018`, `SlurCorpus2020`, and `HatEval2019` report the highest comp_AOPC with scores of 0.80, 0.64, and 0.61, respectively. In contrast, `MLMA2019`, `ToxicSpans2021`, and `ContextToxicity2020` exhibit much lower comp_AOPC scores. `HateComments2023` also reports a high comp_AOPC (0.52), but its suff_AOPC remains elevated at 0.06 – suggesting that, while identified features are impactful when removed, they are not sufficient to preserve the model's prediction when retained.

Variations in faithfulness persist even in cases where models demonstrate comparable AUC. RB on `HateOnlineNewsMedia2018` and `JigsawToxic` both achieve an AUC of 0.98, yet differ by over 0.70 in comp_AOPC; similarly, `InterveneHate2019` and `SlurCorpus2020` have AUC scores of 0.95 and 0.96, but differ by 0.19 in comp_AOPC, while maintaining similar suff_AOPC values. LR achieves lower explanation faithfulness than transformer models, with comp_AOPC scores of 0.20–0.26 vs. 0.26–0.36. The observed pattern likely reflects transformers' ability to handle the out-of-distribution perturbations created by AOPC metrics.

Table 1: Comparison of the four models on the Jigsaw Unintended Bias dataset, reporting AUC and GMB (in bold the highest scores and the closest ones per classifier). The additional fairness metrics (Subgroup AUC, BPSN AUC, and BNSP) AUC are reported in Appendix E.

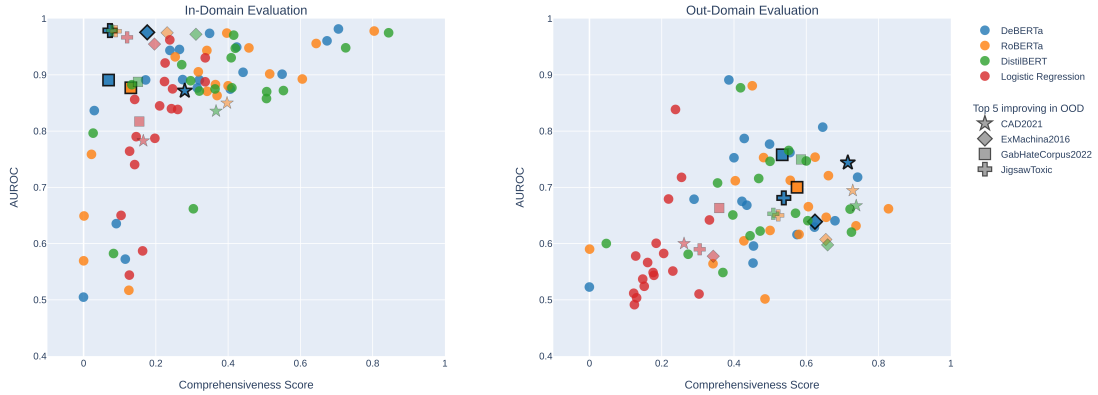
Dataset	LR		DB		DeB		RB	
	AUC	GMB	AUC	GMB	AUC	GMB	AUC	GMB
BullyDetect2018	0.68	0.54	0.72	0.55	0.79	0.56	0.78	0.56
CAD2021	0.70	0.61	0.78	0.68	0.82	0.71	0.81	0.70
ContextToxicity2020	0.69	0.52	0.86	0.66	0.88	0.66	0.81	0.60
Encase2018	0.81	0.64	0.89	0.74	0.91	**0.75**	0.91	0.73
Ethos2022	0.58	0.55	0.71	0.62	0.74	0.65	0.79	0.67
ExMachina2016	0.81	**0.66**	0.89	0.68	0.91	0.68	0.90	0.65
GabHateCorpus2022	0.68	0.57	0.73	0.58	0.77	0.62	0.76	0.58
HateComments2023	0.55	0.52	0.74	0.58	0.71	0.60	0.76	0.59
HateOffensive2017	0.63	0.54	0.65	*0.50*	0.55	0.53	0.49	*0.50*
HateOnlineNewsMedia	0.69	0.60	0.87	0.73	0.87	**0.75**	0.88	**0.75**
HateXplain2020	0.58	0.53	0.72	0.54	0.74	0.56	0.73	0.56
Hateval2019	0.57	0.55	0.65	0.53	0.68	0.54	0.70	0.54
InterveneHate2019	0.65	0.52	0.66	0.51	0.68	0.52	0.66	0.51
JigsawToxic	0.79	0.56	0.89	0.54	0.89	0.54	0.90	0.55
MeasuringHateSpeech	0.67	0.54	0.72	0.55	0.75	0.57	0.79	0.56
MLMA2019	0.53	0.52	0.66	0.58	0.56	*0.50*	0.73	0.63
OLID2019	0.78	0.63	0.92	**0.78**	0.93	**0.80**	0.92	**0.78**
SlurCorpus2020	0.56	0.55	0.59	0.57	0.62	0.58	0.62	0.58
Supremacist2018	0.57	0.52	0.72	0.59	0.73	0.60	0.72	0.63
ToxicSpans2021	0.45	0.48	0.47	0.43	0.51	0.46	0.41	0.46
USElections2020	0.63	0.55	0.74	0.59	0.76	0.61	0.76	0.63

Out-domain faithfulness of post-hoc explanations. We evaluate all models on a common out-of-distribution test set (`HateXplain`). This setup provides a controlled comparison where all models face identical test conditions, allowing us to isolate how different training environments affect explanations faithfulness. Figures 5 reports the in-domain and out-domain LIME Comprehensiveness scores against predictive performance. Some datasets report consistently faithful explanations regardless of shift – `HateOnlineNewsMedia2018` maintains the highest Comprehensiveness both in-domain (0.70–0.84) and out-domain (0.72–0.83).

We hypothesize that in out-of-domain settings, Comprehensiveness and Sufficiency scores should decline alongside predictive performance, as they rely on predictive likelihood, which drops in well-calibrated models (Desai and Durrett, 2020). This evaluation tests whether faithfulness metrics reflect performance degradation or vary independently, controlling for method and archi-

Figure 4: AOPC faithfulness: LIME and SHAP Sufficiency (lower is better) and Comprehensiveness (higher is better).

Figure 5: Relationship between LIME omprehensiveness and model performance (AUROC) for in-domain (left) and out-domain (right) evaluation. Points are colored by model architecture. Special marker shapes identify datasets in the top-5 model-dataset combinations by out-domain improvement. Faded shapes indicate the same datasets paired with other models outside the top-5. LIME Sufficiency and SHAP Comprehensiveness and Sufficiency scores are reported in Appendix F.



tecture. We observe, however, notable improvements in both metrics for specific combinations: Comprehensiveness increases by 0.43-0.46 for DeB on `JigsawToxic`, `GabHateCorpus2022`, `ExMachina2016`, and `CAD2021`, plus RB on `GabHateCorpus2022`; Sufficiency improves by 0.06-0.09 for LR on `JigsawToxic`, DeB on `HateOffensive2017` and `Ethos2022`, and DB on `HateOffensive2017` and `SlurCorpus2020`. The relationship between performance and faithfulness varies substantially: DeB on `HateOffensive2017` maintains nearly identical AUC across domains (0.57 in-domain vs. 0.57 out-domain) while showing a $+0.08$ Sufficiency improvement, whereas DeB on `JigsawToxic` exhibits a $+0.46$ Comprehensiveness gain alongside a severe 0.30 AUC drop from near-perfect in-domain performance (0.98).

Observations. (1) Faithfulness demonstrates complex dependencies on learned representations. (2) It sometimes appears as weakest when the model is strongest, and most reliable when performance fails.

5 Discussion

We analyzed how learned representations in hate speech detection models are shaped by 21 different training datasets. The three dimensions evaluated are not arbitrary but interconnected: robustness determines the breadth of generalization, fairness reflects which subgroups are disproportionately harmed under that generalization, and explainability reveals how models operationalize those decisions. Failures cascade – a lack of robustness amplifies fairness harms, while both shape the faithfulness of explanations.

Case studies. This framework aims to make this cascade visible. `JigsawToxic`, for example, delivers near-perfect AUC in-domain (≈ 0.98) and functions as one of the hubs in cross-dataset generalization. Yet its fairness collapses (GMB ≈ 0.54–0.55). The faithfulness of post-hoc explanations falls below that of the same model architecture trained on several other datasets. Cross-dataset evaluation reveals a striking inversion: when `JigsawToxic`-trained models are evaluated on `HateXplain2020`,

AUROC drops substantially (≈ -0.30) while Comprehensiveness increases ($+0.46$). Faithfulness appears as weakest when the model is strongest, and most reliable when performance fails. This is precisely why these dimensions must be studied together: a model can be robust but unfair, fair but uninterpretable, or accurate but misleadingly explainable, depending on the dataset that trained it. The models that use `JigsawToxic` in training are highly accurate in-domain and robust to shift, but exhibit poor fairness and contradictory explanation reliability. BERT-based models that use `HateOnlineNewsMedia2018` in training achieve high in-domain accuracy (AUC $\approx$ 0.98), strong fairness (GMB = 0.75), and demonstrate the most consistently faithful explanations both in-domain (Comprehensiveness 0.70-0.84) and under distribution shift (0.72-0.83 on `HateXplain2020`). In contrast, BERT-based models that use `ContextToxicity2020` in training achieve lower in-domain accuracy (AUC 0.74–0.88), weaker fairness (GMB 0.60–0.66), and demonstrate the lowest explanation faithfulness among all evaluated datasets.

General evaluations. Machine learning models assume distributional alignment between training and test data – an assumption our cross-domain experiments systematically invalidate. We demonstrate substantial distributional heterogeneity causing significant performance degradation, even among datasets sharing domain objectives and annotation frameworks (`HateOffensive2017` vs. `HateXplain2020`). Performance drops can also reflect dataset design rather than model failure: `SlurCorpus2020` and `ToxicSpans2021` deliberately emphasize narrow linguistic phenomena.

From a fairness perspective, this heterogeneity is equally pronounced. Across 21 datasets, 12 model-dataset pairs show near-random GMB (≈ 0.5), with 20 more barely above chance (0.52–0.55). This pattern manifests in two distinct outcomes: models either maintain predictive accuracy while violating fairness criteria (e.g., DeB on `OLID2019`, AUC $= 0.92$, GMB $= 0.79$; DB on `ContextToxicity2020`, AUC $= 0.85$, GMB $= 0.66$), or fail at both metrics (e.g., LR on `Hateval2019`, AUC $= 0.51$, GMB $= 0.54$).

Representation gaps exposed by robustness and fairness also surface in explanation faithfulness. Post-hoc explainability methods, when evaluated on models trained and tested on the same distribution, exhibit volatility independent of feature attribution methods and model architectures. This instability manifests even across models with comparable predictive performance (RoB on `HateOnlineNewsMedia2018` vs. `JigsawToxic`, AUROC $\approx$ 0.98 but comp_AOPC that differs by $>$ 0.70). In out-domain evaluation, where multiple models trained on different datasets are tested against a common distribution, we observe that while predictive performance degrades predictably, explanation faithfulness metrics show inconsistent and often contradictory responses (DeB on `JigsawToxic` drops -0.30 AUROC on `HateXplain2020` yet comp_AOPC rises $+0.46$).

Dataset creation. Two of the most effective training datasets across all evaluation dimensions are `OLID2019` and `HateOnlineNewsMedia`. Despite not being the largest, both are carefully curated. For instance, `HateOnlineNewsMedia` achieves high out-of-domain AUC similar to `ENCASE2018`, yet the latter is 29 times larger. Both were created using domain-specific keyword strategies and piloting phases to ensure collection quality.

Dataset selection. Some datasets induce brittle representations that degrade under shift, amplify fairness violations, and destabilize explanations – even when in-domain accuracy is high. This is especially clear in `MLMA2019`, where diffuse task scope (e.g., multiple abuse types across just 5,593 examples) prevent the model from learning stable decision boundaries. As our findings show, dataset size alone is not predictive of quality – what matters is whether the data adequately represents the complexity of the phenomenon and aligns with the intended evaluation goals.

6 Conclusion

Rather than advocating for larger and enhanced datasets, an approach that reinforces the field's fixation on scale, we aimed to foster a deeper reflection on the impact of dataset selection. While achieving high AUC on individual hate speech benchmarks might suggest progress, our analysis of learned representations reveals: pervasive distributional divergence evidenced by cross-domain performance degradation, the inability to simultaneously ensure robustness and demographic fairness, and complex dependencies with explainability faithfulness.

7 Limitations

While numerous metrics exist for evaluating model behavior, we deliberately restricted our focus to

a core set that are both widely validated in literature and directly relevant to our research objectives. Importantly, we ensured that all selected metrics and methods were compatible across the range of model configurations considered. The faithfulness metrics (Sufficiency and Comprehensiveness) require producing counterfactual inputs that are inherently out-of-distribution to models. Our concerns about this methodological constraint echo those raised in prior work (Hase et al., 2021). To address this limitation, we also consider soft Sufficiency and Comprehensiveness (Zhao and Aletras, 2023), which attenuate token embeddings proportionally to their importance scores rather than removing them entirely. This reduces the risk of evaluating explanations on out-of-distribution inputs (Zhao and Aletras, 2023). Results are reported in Appendix F. We do not compute these metrics for the logistic regression models, since they use TF–IDF rather than contextual embeddings, making proportional masking ill-defined. As outside the scope of this paper, in future work we will investigate the mechanisms behind high levels of post-hoc explainability faithfulness observed in conditions of model degradation on out-of-distribution sentences. Preliminary experiments, reported in Appendix H, seem to suggest that when models encounter out-of-distribution inputs, they resort to simpler heuristics – using fewer features with more concentrated importance. This leads to increased explanation faithfulness despite degraded model performance, as explanation methods more accurately capture these simplified decision patterns rather than more complex reasoning exhibited in-distribution. As future work, we plan to conduct controlled experiments that will help establish causal mechanisms underlying the correlations identified in our current study.

7.1 Ethical Considerations

This study examines hate speech dataset variations through three model desiderata, recognizing that performance differences often reflect legitimate contextual distinctions rather than methodological flaws. Examples appear without identifying metadata, and research received institutional ethics approval.

8 Acknowledgments

We used AI language models for proofreading portions of the paper.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In Advances in Neural Information Processing Systems, pages 9505–9515.

Saad Almohaimeed, Saleh Almohaimeed, Ashfaq Ali Shafin, Bogdan Carbunar, and Ladislau Bölöni. 2023. THOS: A benchmark dataset for targeted hate and offensive speech. In Workshop on Data-centric machine learning.

David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. In ICML Workshop on Human Interpretability in Machine Learning.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. arXiv preprint arXiv:2307.01680.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In Proceedings of the 42nd international acm sigir conference on research and development in information retrieval, pages 45–54.

Adrian Arnaiz-Rodriguez and Nuria Oliver. 2024. Towards algorithmic fairness by means of instance-level data re-weighting based on Shapley values. In Workshop on Data-centric machine learning.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In International conference on machine learning, pages 233–242. PMLR.

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.

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 Proceedings of the 13th International Workshop on Semantic Evaluation, pages 54–63. ACL.

Tazeek Bin Abdur Rakib and Lay-Ki Soon. 2018. Using the reddit corpus for cyberbully detection. pages 180–189. Springer.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In Companion proceedings of the 2019 world wide web conference, pages 491–500.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3495–3508.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. arXiv preprint arXiv:2010.04736.

Chun Sik Chan, Huanqi Kong, and Guanqing Liang. 2022. A comparative study of faithfulness metrics for model interpretability methods. arXiv preprint arXiv:2204.05514.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. Proceedings of the ICWSM 2017, 11(1):512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In Proceedings of the ALW2 2018. ACL.

Junwei Deng and Jiaqi Ma. 2024. Computational copyright: Towards a royalty model for AI music generation platforms. In Workshop on Data-centric machine learning.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. arXiv preprint arXiv:2003.07892.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. arXiv preprint arXiv:1911.03429.

DMLR. 2024. Call for papers DMLR 2024. https://dmlr.ai/cfp-icml24/. Accessed: 2025-02-15.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pages 214–226. ACM.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259–268. ACM.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Information Processing & Management, 58(3):102524.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, others, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. Proceedings of the ICWSM 2018, 12(1).

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. ACM Computing Surveys, 55(13s):1–32.

Lukas Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. In Proceedings of the WASSA 2021, pages 171–180. ACL.

Sarthak Gupta, Pranav Priyadarshi, and Manish Gupta. 2023. Hateful comment detection and hate target type prediction for video comments. In Proceedings of the CIKM 2023, CIKM '23, pages 3923–3927. ACM.

Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, pages 3315–3323.

Peter Hase, Harry Xie, and Mohit Bansal. 2021. The out-of-distribution problem in explainability and search methods for feature importance explanations. Advances in neural information processing systems, 34:3650–3666.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543.

Marijn Janssen, Paul Brous, Elsa Estevez, Luis S Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy artificial intelligence. Government information quarterly, 37(3):101493.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2019. Fantastic generalization measures and where to find them. arXiv preprint arXiv:1912.02178.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, others, and Morteza Dehghani. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. Language Resources and Evaluation, 56(1):79–108.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Caterina von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application.

Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2024. Defverify: Do hate speech models reflect their dataset's definition? arXiv preprint arXiv:2410.15911.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (un) reliability of saliency methods. In Explainable AI: Interpreting, explaining and visualizing deep learning, pages 267–280. Springer.

Ravin Kohli, Matthias Feurer, Katharina Eggensperger, Bernd Bischl, and Frank Hutter. 2024. Towards quantifying the effect of datasets for benchmarking: A look at tabular machine learning. In Workshop on Data-centric machine learning.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In Proceedings of the WOAH 2020, pages 138–149, Online. ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101, 5.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI 2020.

Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (xai). Algorithms, 15(8):291.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. Complex & Intelligent Systems, 8(6):4663–4678.

Rafael Mosquera Gómez, Julian Eusse, Juan Ciro, Daniel Galvez, Ryan Hileman, Kurt Bollacker, and David Kanter. 2023. Speech Wikimedia: A 77 language multilingual speech dataset. In Workshop on Data-centric machine learning.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In International conference on complex networks and their applications, pages 928–940. Springer.

OECD. 2024. Recommendation of the council on artificial intelligence. Accessed: 16 May 2025.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In Proceedings of the EMNLP-IJCNLP 2019, pages 4675–4684. ACL.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: Attributing model behavior at scale. arXiv preprint arXiv:2303.14186.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11).

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter?

John Pavlopoulos, Jeffrey Sorensen, Léa Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In Proceedings of the SemEval 2021, pages 59–69. ACL.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3CubeMahaSent-MD: A multi-domain Marathi sentiment analysis dataset and transformer models. In Workshop on Data-centric machine learning.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection. In Proceedings of the International AAAI Conference on Web and Social Media, volume 18, pages 2025–2039.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In Proceedings of the EMNLP-IJCNLP 2019, pages 4755–4764. ACL.

José Ribeiro, Raíssa Silva, Lucas Cardoso, and Ronnie Alves. 2021. Does dataset complexity matters for model explainers? In 2021 IEEE International Conference on Big Data (Big Data), pages 5257–5265. IEEE.

Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 29(05):582–638.

Pranav Sachdeva, Ricardo Barreto, Geoff Bacon, Alexander Sahn, Caterina von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In Proceedings of the LREC 2022, pages 83–94. ELRA.

Joni Salminen, Hind Almerekhi, Milos Milenkovic, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. Proceedings of the ICWSM 2018, 12(1).

Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using bert: A multi-task approach. In Proceedings

of the second workshop on trolling, aggression and cyberbullying, pages 126–131.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Rajat Shinde, Sujit Roy, Christopher E Phillips, Aman Gupta, Aditi Sheshadri, Manil Maskey, and Rahul Ramachandran. 2024. WINDSET: Weather insights and novel data for systematic evaluation and testing. In Workshop on Data-centric machine learning.

Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A Hale, Samuel P Fraiberger, Victor Orozco-Olvera, and Paul Röttger. 2024. Hateday: Insights from a global hate speech dataset representative of a day on twitter. arXiv preprint arXiv:2411.15462.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In CVPR 2011, pages 1521–1528. IEEE.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing cad: the contextual abuse dataset. In Proceedings of the NAACL 2021, pages 2289–2303. ACL.

Artem Vysogorets and Julia Kempe. 2024. Towards robust data pruning. In Workshop on Data-centric machine learning.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. arXiv preprint arXiv:2102.12060.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex machina: Personal attacks seen at scale.

Yao Xiao, Yaoyao Chang, Cheng Peng, Siyu Li, and Zhiyu Yuan. Jigsaw unintended bias in toxicity classification.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. International Journal of Computer Vision, pages 1–28.

Rubing Yang, Jialin Mao, and Pratik Chaudhari. 2022. Does the data induce capacity control in deep learning? In International Conference on Machine Learning, pages 25166–25197. PMLR.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. Advances in neural information processing systems, 32.

Michael Miller Yoder, Lynnette Hui Xian Ng, David West Brown, and Kathleen M Carley. 2022. How hate speech varies by target identity: A computational analysis. arXiv preprint arXiv:2210.10839.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In Proceedings of the NAACL 2019, pages 1415–1420. ACL.

Zhixue Zhao and Nikolaos Aletras. 2023. Incorporating attribution importance for improving faithfulness metrics. arXiv preprint arXiv:2305.10496.

Yilun Zhou and Julie Shah. 2022. The solvability of interpretability evaluation metrics. arXiv preprint arXiv:2205.08696.

Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148.

9 NLP Checklist

A1. Did you describe the limitations of your work? Yes. Please refer to Section 7.

A2. Did you discuss any potential risks of your work? Yes. We have discussed some ethical considerations in Section 7.1.

B. Did you use or create scientific artifacts? Yes, we used existing scientific artifacts (datasets, pre-trained models, evaluation metrics).

B1. Did you cite the creators of artifacts you used? Yes. Please refer to Section 3.

B2. Did you discuss the license or terms for use and/or distribution of any artifacts? Yes. Please refer to Section 3.

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)? Our use of the MetaHate dataset follows its intended research purpose, accessed through a signed terms of use agreement. Any derivatives from our work maintain the original research-only restrictions and cannot be used outside research contexts.

B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it? The datasets used contain offensive language. The sources are publicly available, however, to avoid any distressing feeling to our readers we avoided presenting and cite content in the full body of the paper that can affect the readers.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? The datasets are explained in detail by the authors of the MetaHate paper. The descriptions in this paper include only what is necessary to this work.

B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created? Yes, please refer to Sections 3 and 4.

C. Did you run computational experiments? Yes.

C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used? Yes. Please refer to Section 3.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values? Yes. We did not perform hyperparameter tuning.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run? Yes. We report performance results in Section 4 and Appendix D.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used? Yes. Please refer to Section 3.

D. Did you use human annotators (e.g., crowdworkers) or research with human participants? No.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? N/A.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)? N/A.

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? N/A.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board? N/A.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data? N/A.

E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing? Yes. We used AI language models for proofreading portions of the manuscript to check for grammatical errors and clarity.

E1. Did you include information about your use of AI assistants? Yes. Please refer to Section 8.

A Datasets

Dataset	Size	Description	Original Annotation	Source	References
Binary Classification					
Hateval 2019	12,747	Hate speech against women and immigrants	Hate, Non-hate	Twitter	Basile et al., 2019
OLID 2019	14,052	Hierarchical offensive language	Hate, Non-hate	Twitter	Zampieri et al., 2019
US 2020 Elections	2,999	Political hate speech	Hate, Non-hate	Twitter	Grimminger and Klinger, 2021
BullyDetect 2018	6,562	Cyberbullying	Cyberbullying, No cyberbullying	Reddit	Bin Abdur Rakib and Soon, 2018
Intervene Hate 2019	45,170	Counter-speech and hate speech	Hate, Non-hate	Reddit, Gab	Qian et al., 2019
Hate in Online News	3,214	News comments	Hate, Non-hate	Facebook	Salminen et al., 2018
Supremacist 2018	10,534	White supremacist content	Hate, Non-hate	Stormfront	de Gibert et al., 2018
Gab Hate Corpus	27,434	Hate speech	Assault on Human Dignity / No	Gab	Kennedy et al., 2022
HateComments 2023	2,070	Hate speech	Hate, Non-hate	YouTube	Gupta et al., 2023
Ex Machina 2016	115,705	Toxicity detection	Attack, No Attack	Wikipedia	Wulczyn et al., 2016
Context Toxicity 2020	19,842	Context-aware toxicity	Toxic, No Toxic	Wikipedia	Pavlopoulos et al., 2020
Multi-class / Multi-label Classification					
Hate Offensive 2017	24,783	Offensive language	Hate Speech, Offensive, Neither	Twitter	Davidson et al., 2017
ENCASE 2018	91,950	Cyberbullying and hate speech	Abusive, Normal, Spam, Hateful	Twitter	Founta et al., 2018
MLMA 2019	5,593	Hate speech	Multiple abuse categories	Twitter	Ousidhoum et al., 2019
HateXplain 2020	20,109	Explainable hate speech	Hate, Offensive, Normal	Twitter, Gab	Mathew et al., 2020
Slur Corpus 2020	39,960	Slur-based hate speech	Multiple slur categories	Reddit	Kurrek et al., 2020
CAD 2021	23,060	Contextual abuse	Multiple abuse types	Reddit	Vidgen et al., 2021
Severity Scale					
Measuring Hate 2020-22	39,565	Linear hate speech scale	Severity scale	Twitter, Reddit, YouTube	Kennedy et al., 2020; Sachdeva et al., 2022
ETHOS 2020	998	Multi-target hate speech	Severity scale	Reddit, YouTube	Mollas et al., 2022
Span-level Annotation					
Toxic Spans 2021	10,621	Token-level toxicity	Span-level annotation	Comments	Pavlopoulos et al., 2021

Table 2: Description of the dataset adopted for the experimental evaluation.

Note: Datasets are grouped by classification type. For a comprehensive description of each dataset, please refer to Piot et al., 2024. While the original Toxic Spans 2021 dataset (Pavlopoulos et al., 2021) identified specific text segments indicating toxicity, in MetaHate (Piot et al., 2024) the authors have standardized its format to match other datasets, providing binary classifications of whether comments contain hate speech or not. For MLMA 2019, they (Piot et al., 2024) have selected only text in English.

B Classification parity metrics

- Subgroup AUC: We restrict the data set to only the examples that mention the specific identity subgroup. A low value in this metric means the model does a poor job of distinguishing between toxic and non-toxic comments that mention the identity.

- BPSN AUC: We restrict the test set to the non-toxic examples that mention the identity and the toxic examples that do not. A low value in this metric means that the model confuses non-toxic examples that mention the identity with toxic examples that do not, likely meaning that the model predicts higher toxicity scores than it should for non-toxic examples mentioning the identity.

- BNSP AUC: We restrict the test set to the toxic examples that mention the identity and the non-toxic examples that do not. A low value here means that the model confuses toxic examples that mention the identity with non-toxic examples that do not, likely meaning that the model predicts lower toxicity scores than it should for toxic examples mentioning the identity.

- GMB AUC: This metric was introduced by the Google Conversation AI Team as part of their Kaggle competition. [1] This metric combines the per-identity Bias AUCs into one overall measure as:

$$M_p(m_s) = \left(\frac{1}{N} \sum_{s=1}^{N} m_s^p \right)^{\frac{1}{p}} \tag{1}$$

  where:

  - M_p = the pth power-mean function
  - m_s = the bias metric m calculated for subgroup s
  - N = number of identity subgroups (10)

  We use $p = -5$ as was done in the competition.

C Post-hoc explainability faithfulness metrics

- Comprehensiveness represents the impact of replacing most important rationales by a baseline. For each input x_i, we construct a contrast example $\tilde{x}_i = x_i \setminus r_i$ by removing the predicted rationales r_i. Let $m(x_i)_j$ denote the prediction probability assigned by model m to class j for the original input. Comprehensiveness is defined as *comprehensiveness* $= m(x_i)_j - m(x_i \setminus r_i)_j$, where $m(x_i \setminus r_i)_j$ is the prediction probability for the contrast example. When important rationales are removed, we expect the model's confidence to decrease, yielding a higher Comprehensiveness score that indicates more faithful interpretations (DeYoung et al., 2019).

- Sufficiency represents the impact of adding most important features to a baseline in the predictive behavior. We measure Sufficiency as *sufficiency* $= m(x_i)_j - m(r_i)_j$, where $m(r_i)_j$ is the prediction probability when only rationales are provided. A lower Sufficiency score implies the rationales contain essential information for the model's prediction, suggesting more faithful interpretations (DeYoung et al., 2019).

[1] Jigsaw Unintended Bias in Toxicity Classification competition: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

D Cross-dataset generalization performance

Figure 6: Cross-dataset generalization performance comparison between RobBERTa (left) and DistilBERT (right) models reporting AUC classification scores when training on one dataset (rows) and testing on another (columns), with diagonal cells representing in-domain performance and off-diagonal cells indicating cross-domain generalization.

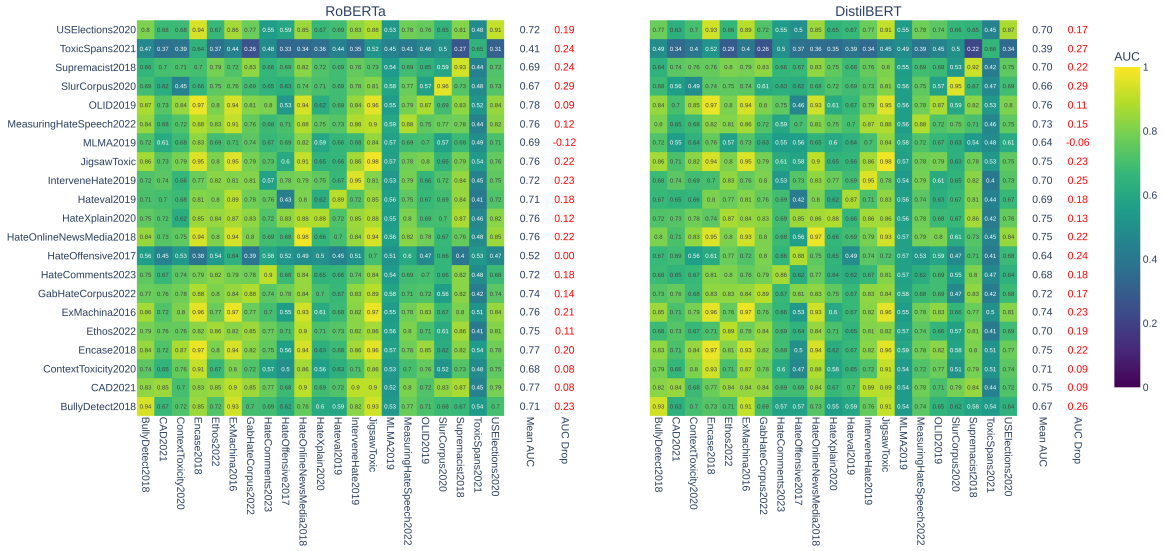


Table 3: Within-dataset AUROC performance comparison.

Dataset	LR	DB	DeB	RB
BullyDetect2018	0.888	0.930	0.945	0.943
CAD2021	0.783	0.835	0.872	0.850
ContextToxicity2020	0.650	0.796	0.837	0.735
Encase2018	0.962	0.971	0.974	0.974
Ethos2022	0.764	0.890	0.877	0.863
ExMachina2016	0.955	0.972	0.975	0.974
GabHateCorpus2022	0.817	0.887	0.891	0.880
HateComments2023	0.790	0.858	0.904	0.902
HateOffensive2017	0.875	0.883	0.645	0.523
HateOnlineNewsMedia2018	0.921	0.975	0.982	0.978
HateXplain2020	0.839	0.877	0.891	0.881
Hateval2019	0.840	0.872	0.901	0.893
InterveneHate2019	0.930	0.947	0.949	0.948
JigsawToxic	0.967	0.979	0.979	0.978
MeasuringHateSpeech2022	0.845	0.875	0.892	0.883
MLMA2019	0.544	0.582	0.499	0.570
OLID2019	0.787	0.871	0.875	0.871
SlurCorpus2020	0.888	0.948	0.960	0.956
Supremacist2018	0.856	0.918	0.943	0.932
ToxicSpans2021	0.587	0.662	0.636	0.649
USElections2020	0.741	0.870	0.891	0.905

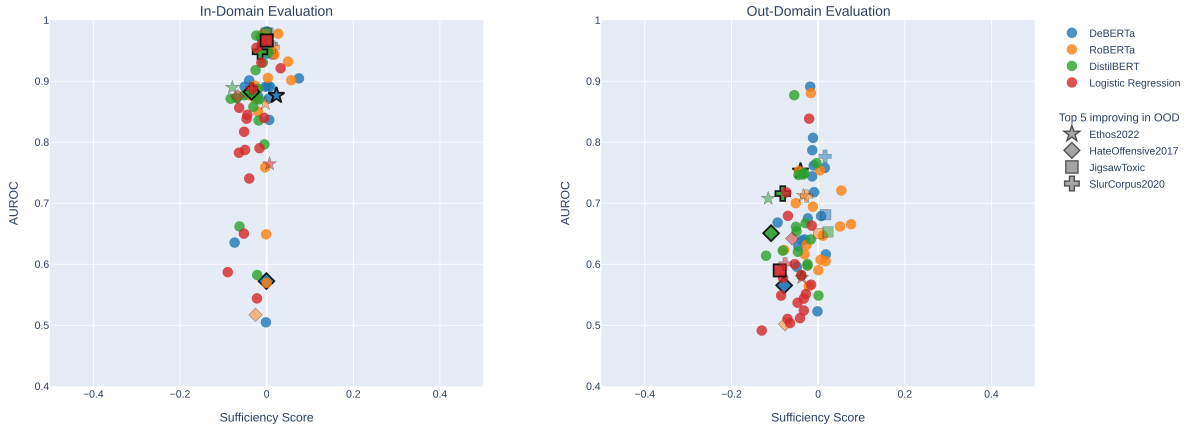
E Subgroup, BPSN, and BNSP AUC metrics

Table 4: Comparison of fairness metrics (Subgroup, BPSN, and BNSP AUC) across the four models. Metrics evaluate performance across subgroups including gender (male, female), race (white, black), religion (christian, jewish, muslim), sexual orientation (homosexual/gay/lesbian), and mental health (psychiatric or mental illness).

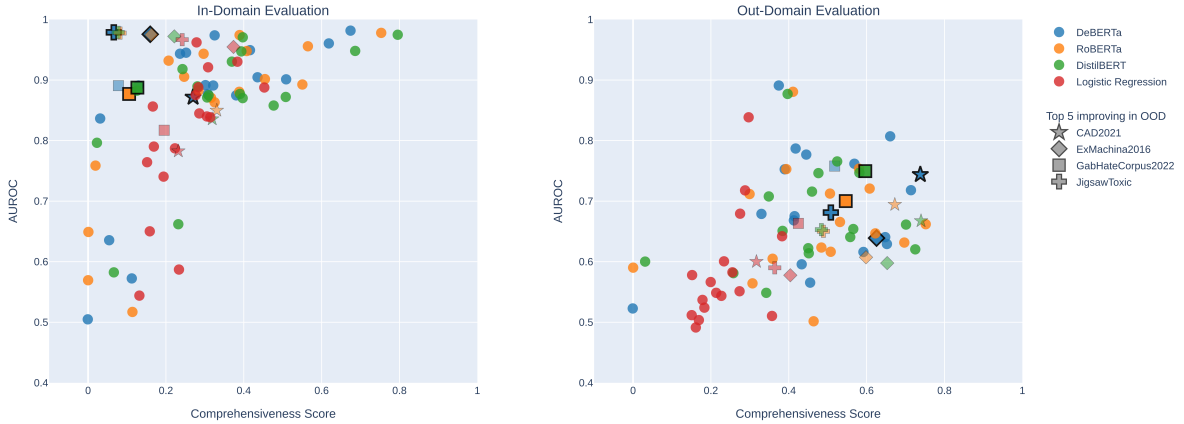
Dataset	**Subgroup AUC**				**BPSN AUC**				**BNSP AUC**			
	LR	RB	DeB	DB	LR	RB	DeB	DB	LR	RB	DeB	DB
BullyDetect2018	0.506	0.517	0.519	0.512	0.562	0.581	0.581	0.569	0.497	0.514	0.516	0.512
CAD2021	0.562	0.637	0.644	0.640	0.391	0.503	0.582	0.552	0.809	0.853	0.791	0.785
ContextToxicity2020	0.522	0.561	0.574	0.587	0.574	0.608	0.687	0.668	0.517	0.568	0.589	0.612
Encase2018	0.644	0.668	0.673	0.684	0.566	0.707	0.608	0.690	0.754	0.732	0.869	0.760
Ethos2022	0.551	0.637	0.632	0.617	0.525	0.593	0.571	0.417	0.560	0.724	0.708	0.816
ExMachina2016	0.632	0.582	0.592	0.595	0.721	0.694	0.738	0.735	0.632	0.585	0.583	0.586
GabHateCorpus2022	0.546	0.558	0.639	0.587	0.386	0.244	0.380	0.309	0.773	0.919	0.852	0.860
Hateval2019	0.522	0.524	0.524	0.504	0.579	0.543	0.561	0.532	0.493	0.520	0.509	0.502
HateComments2023	0.496	0.533	0.549	0.528	0.490	0.519	0.544	0.509	0.537	0.621	0.621	0.622
HateOffensive2017	0.553	0.500	0.507	0.526	0.437	0.500	0.358	0.525	0.682	0.500	0.691	0.528
HateXplain2020	0.534	0.544	0.543	0.531	0.594	0.557	0.543	0.531	0.501	0.544	0.550	0.533
InterveneHate2019	0.513	0.513	0.519	0.502	0.497	0.507	0.521	0.503	0.526	0.516	0.519	0.503
JigsawToxic	0.542	0.511	0.516	0.507	0.598	0.568	0.580	0.552	0.541	0.511	0.515	0.507
MLMA2019	0.499	0.577	0.500	0.540	0.471	0.618	0.500	0.652	0.544	0.598	0.500	0.488
MeasuringHateSpeech2022	0.518	0.530	0.535	0.529	0.566	0.582	0.576	0.560	0.510	0.524	0.530	0.529
HateOnlineNewsMedia2018	0.570	0.654	0.647	0.645	0.437	0.798	0.789	0.753	0.769	0.662	0.658	0.684
OLID2019	0.563	0.665	0.695	0.668	0.517	0.682	0.806	0.697	0.753	0.823	0.742	0.812
SlurCorpus2020	0.527	0.550	0.562	0.545	0.550	0.597	0.629	0.665	0.520	0.557	0.531	0.455
Supremacist2018	0.481	0.586	0.577	0.573	0.509	0.529	0.565	0.520	0.523	0.706	0.602	0.641
ToxicSpans2021	0.439	0.437	0.468	0.410	0.496	0.479	0.595	0.545	0.407	0.428	0.348	0.303
USElections2020	0.529	0.560	0.545	0.542	0.591	0.669	0.646	0.629	0.511	0.563	0.545	0.537

F Out-domain faithfulness of post-hoc explanations

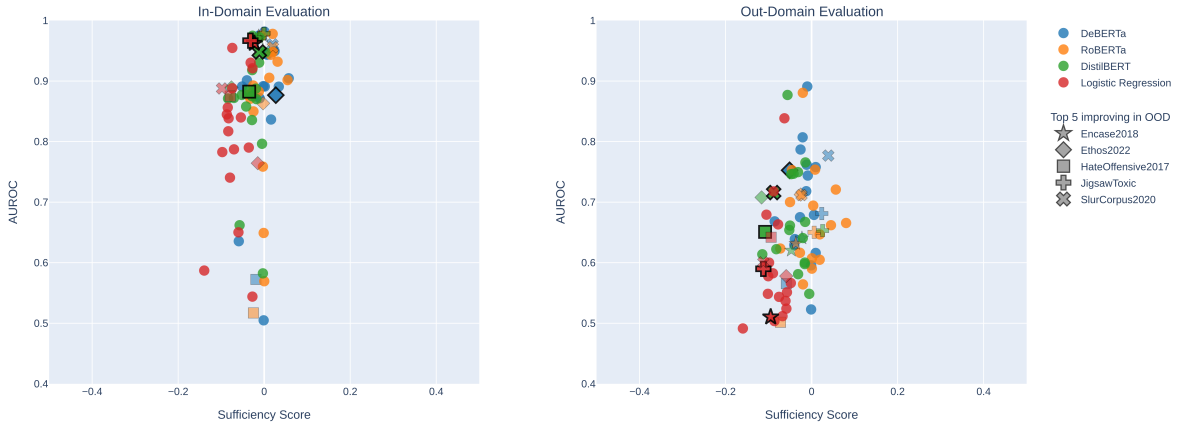
Figure 7: Relationship between SHAP AOPC scores and model performance (AUROC) for in-domain (left) and out-of-domain (right) evaluation. Points are colored by model architecture. Special marker shapes identify datasets in the top-5 model-dataset combinations by out-of-domain improvement. Highlighted markers (full opacity, thick border) represent actual top-5 combinations; faded shapes indicate the same datasets paired with other models outside the top-5.



(a) LIME Sufficiency.



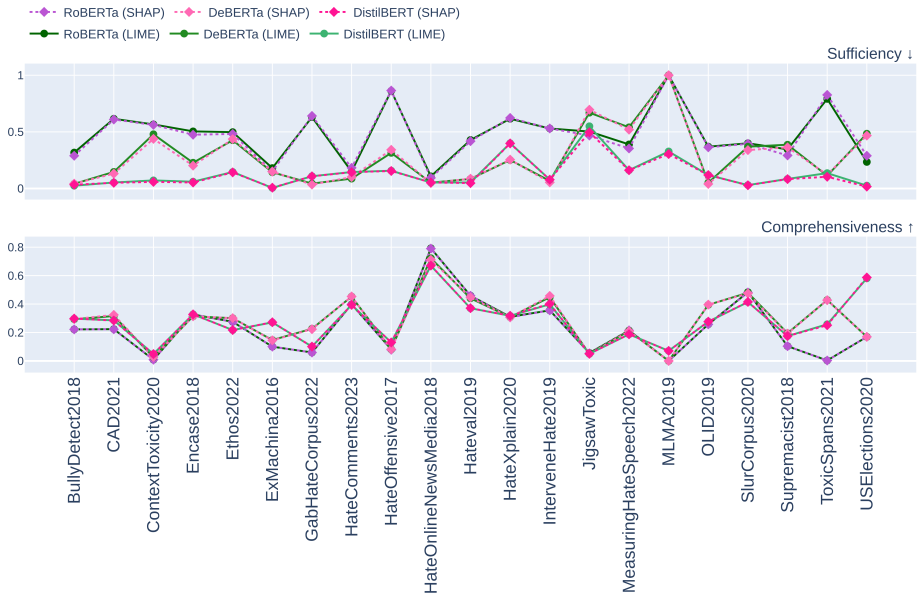
(b) SHAP Comprehensiveness.



(c) SHAP Sufficiency.

G Soft Sufficiency and Comprehensiveness

Figure 8: Soft faithfulness of LIME (solid line) and SHAP (dotted line) explanations for in-domain evaluation: Sufficiency (top, lower is better) and Comprehensiveness (bottom, higher is better). We find mininal difference between the two post-hoc explainability methods.



H Post-hoc explainers and out-of-distribution-induced heuristic simplification

We conduct a preliminary experiment to investigate how distribution shift affects post-hoc explainability faithfulness. We used a logistic regression model trained on `ExMachina2016`, and compared SHAP explanations between in-distribution (`ExMachina2016`) and out-of-distribution (`HateXplain2020`) inputs ($n = 50$ each). We chose `ExMachina2016` specifically because it demonstrates higher faithfulness scores (both Comprehensiveness and Sufficiency) when evaluating out-of-domain inputs compared to in-domain inputs. Results revealed that explanations on out-of-distribution inputs relied on significantly fewer significant features (464 vs. 1,210), exhibited higher importance concentration (Gini coefficient 0.994 vs. 0.985), and lower feature entropy (6.83 vs. 8.06). We also observe a dramatic reduction in feature diversity, with out-of-distribution inputs using only 58.28 average unique features per sample compared to 205.92 for in-distribution data. Moreover, there is minimal semantic overlap between distributions – only 20% of top features were shared between in-distribution and out-of-distribution inputs. These findings might indicate that when encountering unfamiliar inputs, models resort to simplified decision heuristics that explanation methods can more accurately capture, creating misleading impressions of explanation reliability.