IMC: A Benchmark for Invariant Learning under Multiple Causes

Supplementary Material

A. Details on Text Data

A.1. HateXplain

A.1.1. Original Generation Process

The creation process of HateXplain benchmark dataset [21] involves several systematic steps, beginning with data collection. Researchers gather textual data from various online platforms, including social media and forums, where users frequently post hateful or offensive content. They select posts containing explicit hate speech, offensive language, or neutral statements to ensure a diverse and representative dataset.

In the next stage, human annotators manually label the collected data. Annotators classify each text into predefined categories such as hate speech, offensive language, or neutral content. Additionally, annotators highlight specific words or phrases within the text that contribute to their labeling decision, providing rationales for their annotations. Multiple annotators independently label each text to ensure reliability and mitigate individual biases.

Finally, researchers aggregate the annotations using consensus methods to resolve discrepancies among annotators. They apply statistical measures such as majority voting or inter-annotator agreement metrics to finalize labels and rationales. The resulting dataset includes both labeled texts and explicit rationales explaining annotation decisions, making it valuable for training and evaluating hate speech detection models that require interpretability and transparency.

A.1.2. Editing Process

In our study, we focus on a binary classification task using two specific classes from the dataset: hate speech and offensive speech. We narrow our scope to three primary target groups: homosexual, African, and women. These target categories are further organized into single-target groups (instances targeting only homosexuals, Africans, or women individually) and multi-target groups (instances simultaneously targeting combinations such as homosexual & African, homosexual & women, or African & women). To ensure data quality and annotation reliability, we only include instances where the class label and target community identification have been confirmed by at least two independent annotators, thus establishing a consensus-based approach to dataset curation.

The dataset exhibits an interesting structural characteristic in its distribution across training, validation, and testing splits. While the training and validation datasets predominantly contain multi-target instances (where the offensive or hateful content targets multiple groups simultaneously), the testing dataset primarily consists of single-target instances. This deliberate configuration allows us to evaluate how well models can generalize from learning patterns in multi-target scenarios to identifying hate speech in more focused, single-target contexts. For the validation dataset specifically, we maintain balanced weights between the hate speech and offensive content classes to ensure fair evaluation metrics. The training dataset's distribution encompasses various proportions of single and multi-target groups, providing the model with exposure to diverse manifestations of problematic content.

A.2. Synthetic NLP

This section describes how our synthetic dataset is generated. Initially, a list of two-letter words is generated using combinations of uppercase and lowercase alphabets, resulting in a total of 676 unique words. These words are then divided into specific categories based on predefined counts for each class and feature category. The categories refer to the invariant feature, and the query in the category `inv0` and `inv1` contains only one feature, while the query in `both` contains both features. These three feature categories are further divided into three classes: `class0`, `class1`, and `class2`. The words not selected for these categories are classified as `regular_vocab`.

Each text sample is defined as a combination of five words, which includes a mix of category-specific and regular words. For instance, the `class0_both` category includes one word from `class0_inv0`, one from `class0_inv1`, and three from `regular_vocab`. This ensures that each text sample contains a balanced representation of features. The text samples are then shuffled to ensure randomness. This process is repeated for each category, resulting in a comprehensive dataset with a balanced distribution of features across different classes and categories.

The generated text data is split into training and validation datasets. The training dataset is constructed by sampling a specific number of text samples from each category, ensuring that all features are represented. The validation dataset is created by ensuring that each feature from the `inv0` and `inv1` lists is present at least once. This is achieved by selecting text samples that contain these features or generating new samples if necessary. Once the datasets are prepared, they are saved in both pickle and JSON formats for easy access and readability. The feature lists are also saved separately.

B. Details on LLM experiment

B.1. Example of Prompt

Tab. 6 shows the example of prompts for LLM experiments in Section 5.3 of the main paper.

Table 6. System and User Prompt utilized in HateXplain and Synthetic NLP.

Role	Prompts
System	"You are a Classification chatbot."
User	First, read this pair. This is a training dataset. 'prompt' is given data, and 'label' is the target answer. prompt: {`Sample 1 Query`}. label: {`Sample 1 Label`} prompt: {`Sample 2 Query`}. label: {`Sample 2 Label`} ... Based on the given training pair, fill in the completion based on the prompt's sentence. When presenting the results, use the 'label: <answer>' format. prompt: {`input`}. label: <fill this label>

B.2. Model Configuration

Tab. 7 provides a summary of the model configuration settings for the four models used in the LLM experiments in Section **??**, where HateXplain and SyntheticNLP follow the same configuration. In this experiment, we use all models with their default settings without any modification to their parameters. For Claude 3.7 Sonnet, the default value of top_p is not officially disclosed, and thus it is indicated as – in the table.

Table 7. Comparison of model configurations

Models	**Snapshots**	**Temperature**	**top_p**
GPT-4o	gpt-4o-2024-08-06	1.0	1.0
GPT-4o-mini	gpt-4o-mini-2024-07-18	1.0	1.0
Claude 3.7 Sonnet	claude-3-7-sonnet-20250219	1.0	–
Google Gemini 2.0 Flash	gemini-2.0-flash	1.0	0.95

C. Additional Results

Table 8. **Model Selection: Average Accuracy** † stands for the algorithm trained over the original training distribution.

	PlayingCard–10			OddsEvens			CelebA–Jew			COCO–Bag			Synthetic NLP			HateXplain		
	WG	BC	WC	WG	BC	WC	WG	BC	WC	WG	BC	WC	WG	BC	WC	WG	BC	WC
ERM†	76.6±3.7	99.8±0.3	56.7±6.7	0.5±0.8	87.5±6.1	45.6±24.7	12.3±1.2	97.0±1.8	10.6±2.1	67.5 ± 1.4	94.9 ± 0.9	31.7 ± 2.2	**65.5±0.7**	**100.0±0.0**	16.0±4.1	33.4±11.8	91.9±2.0	21.6±8.1
ERM	87.7±1.1	**100.0±0.0**	76.0±1.4	0.0±0.0	77.6±0.0	63.8±0.0	28.6±3.2	96.0±1.3	14.5±4.2	56.8 ± 3.0	97.1 ± 0.5	21.8 ± 3.5	**65.5±0.7**	**100.0±0.0**	16.0±4.1	31.4±14.6	92.2±3.2	20.1±9.8
SAM†	85.3±2.9	99.4±1.0	72.2±5.3	0.0±0.0	79.8±2.0	58.5±5.2	13.8±2.1	**97.8±0.4**	8.8±2.3	62.5 ± 4.8	95.7 ± 0.4	26.9 ± 2.0	**65.5±0.7**	**100.0±0.0**	**16.8±3.6**	33.3±13.5	**92.8±3.2**	21.6±7.6
SAM	**88.6±1.3**	**100.0±0.0**	**79.5±3.5**	0.0±0.0	77.6±0.0	63.8±0.0	37.7±4.0	95.9±2.1	15.5±1.6	56.9 ± 5.4	**97.6±0.2**	19.3 ± 1.0	**65.5±0.7**	**100.0±0.0**	**16.8±3.6**	**40.8±14.6**	81.0±11.0	**31.6±13.7**
GDRO	84.8±4.3	99.8±0.4	72.4±4.2	35.3±26.8	91.9±11.4	**66.2±4.5**	**45.7±2.3**	94.8±2.1	**17.9±1.8**	59.4 ± 3.1	96.4 ± 0.5	23.6 ± 3.5	65.4±0.7	99.7±0.3	14.9±4.0	34.1±17.7	84.5±11.8	25.4±12.9
IRM	77.5±1.7	99.5±0.4	63.3±5.7	**78.8±2.6**	**100.0±0.0**	13.7±7.7	31.7±1.0	96.3±1.2	14.2±3.1	**68.2±1.0**	97.4 ± 2.4	**31.7±0.1**	65.3±0.3	**100.0±0.0**	16.0±4.8	31.5±13.4	88.5±2.1	23.4±8.5

Table 9. **Model Selection: Worst Class Accuracy** † stands for the algorithm trained over the original training distribution.

	PlayingCard–10			OddsEvens			CelebA–Jew			COCO–Bag			Synthetic NLP			HateXplain		
	WG	BC	WC	WG	BC	WC	WG	BC	WC	WG	BC	WC	WG	BC	WC	WG	BC	WC
ERM†	82.3±3.6	**100.0±0.0**	70.3±6.2	0.0±0.0	77.6±0.0	63.8±0.0	49.4 ± 3.4	92.0 ± 2.4	21.5 ± 5.1	66.0 ± 8.5	93.3 ± 3.1	39.2 ± 7.2	**65.9±0.1**	**100.0±0.0**	**19.0±0.0**	40.7±21.6	84.2±10.7	23.9±11.9
ERM	83.2±2.7	99.8±0.4	70.0±3.6	32.3±2.8	89.8±0.9	66.9±0.5	**54.0±2.1**	90.3 ± 1.3	**30.4±7.4**	75.3 ± 6.7	97.0 ± 3.0	43.3 ± 7.7	**65.9±0.1**	**100.0±0.0**	**19.0±0.0**	34.3±11.3	89.1±2.7	24.5±5.9
SAM†	9.5±2.3	**100.0±0.0**	**77.8±1.3**	0.0±0.0	77.6±0.0	63.8±0.0	41.6±2.1	**92.2±2.1**	24.4±5.7	67.7 ± 2.2	96.8 ± 1.7	29.6 ± 1.5	65.1±0.3	**100.0±0.0**	14.6±1.2	44.4±15.0	84.6±7.4	32.0±13.6
SAM	**84.6±0.3**	99.6±0.6	67.8±1.9	31.8±3.3	89.3±0.6	**70.2±0.5**	53.5 ± 1.3	91.6 ± 1.1	30.2 ± 3.2	75.1 ± 6.3	95.3 ± 1.9	50.0 ± 2.2	65.1±0.3	**100.0±0.0**	14.6±1.2	36.1±15.0	**90.9±4.2**	24.8±9.9
GDRO	83.8±3.9	99.6±0.4	72.4±3.1	39.8±30.7	92.4±11.9	65.7±2.3	43.3 ± 1.3	90.5 ± 1.7	30.0 ± 4.2	73.5 ± 5.5	96.9 ± 2.4	**50.8 ± 4.4**	64.9±0.8	99.5±0.3	11.5±5.3	**62.6±5.7**	70.4±9.7	39.5±1.5
IRM	81.0±5.1	99.5±0.4	67.4±14.9	54.3±10.5	99.1±1.4	19.7±22.8	37.7 ± 2.4	90.8 ± 2.1	26.6 ± 2.0	74.5 ± 4.1	95.3 ± 3.8	49.9 ± 2.9	65.3±0.2	**100.0±0.0**	16.1±4.7	49.2±7.2	77.1±7.5	36.8±7.6
V-REx	78.0±3.5	99.1±0.2	61.0±3.0	**79.0±3.1**	**100.0±0.0**	11.0±2.8	30.7 ± 3.1	91.8 ± 1.7	23.5 ± 6.9	75.4 ± 1.0	98.6 ± 0.2	40.4 ± 1.5	65.1±0.9	**100.0±0.0**	12.6±5.3	59.9±6.6	76.7±0.5	40.8±5.1
RDM	74.2±7.4	99.6±0.7	47.7±14.2	24.9±31.9	86.0±12.1	45.5±31.6	44.1 ± 5.2	91.3 ± 1.0	27.6 ± 4.5	**78.5 ± 1.4**	**98.9 ± 0.2**	50.2 ± 2.6	64.2±0.2	99.7±0.1	9.6±3.0	52.1±10.3	75.5±12.6	**40.8±11.6**
URM	83.1±3.4	98.4±0.7	72.2±7.1	0.0±0.0	77.6±0.0	63.8±0.0	48.1 ± 1.8	91.9 ± 0.8	24.8 ± 5.1	74.4 ± 4.7	96.7 ± 3.5	50.4 ± 5.6	62.7±1.0	99.7±0.1	5.2±3.1	32.0±8.2	83.1±6.6	26.8±7.2

Table 10. **Subgroup Analysis.** Model Selection: Average Accuracy

	PlayingCard–10				OddsEvens				CelebA–Jew				COCO–Bag			
	Avg	WG	WC	WS	Avg	WG	WC	WS	Avg	WG	WC	WS	Avg	WG	WC	WS
ERM†	94.5 ± 0.2	84.6 ± 2.2	76.0 ± 1.4	71.5 ± 5.7	50.0 ± 0.0	0.0 ± 0.0	50.0 ± 0.0	0.0 ± 0.0	47.3±2.0	9.2±3.2	10.6±3.8	3.8±1.7	55.5 ± 3.2	43.4 ± 3.1	21.1 ± 3.5	10.4 ± 2.0
ERM	90.0 ± 0.2	75.5 ± 1.1	56.7 ± 6.7	56.7 ± 6.7	49.5 ± 3.5	0.4 ± 0.7	32.4 ± 17.4	0.0 ± 0.1	51.8±1.3	20.0±3.2	14.5±3.3	9.3±2.0	62.0 ± 1.8	54.0 ± 1.1	**30.9 ± 1.6**	18.9 ± 1.8
SAM†	**94.8 ± 0.4**	**86.2 ± 1.0**	**79.5 ± 3.5**	68.3 ± 2.9	50.1 ± 0.0	0.0 ± 0.0	50.0 ± 0.0	0.0 ± 0.0	46.6±2.7	8.8±1.3	8.8±3.7	2.3±1.0	55.1 ± 2.5	43.0 ± 4.3	18.3 ± 1.1	9.2 ± 2.8
SAM	93.3 ± 0.1	82.8 ± 3.4	72.2 ± 5.3	65.5 ± 5.1	50.4 ± 0.3	0.0 ± 0.0	43.7 ± 5.7	0.0 ± 0.0	54.9±1.3	25.5±3.1	15.5±5.6	9.6±2.3	58.0 ± 2.0	48.5 ± 3.4	25.4 ± 1.3	12.0 ± 1.4
GDRO	92.4 ± 0.6	81.6 ± 3.8	72.4 ± 4.2	66.7 ± 7.6	67.5 ± 14.3	35.8 ± 27.5	**53.2 ± 6.6**	**20.5 ± 18.0**	**57.3±4.3**	**32.5±1.7**	**17.9±4.1**	**14.0±2.3**	57.0 ± 2.8	46.4 ± 2.7	22.9 ± 3.1	13.9 ± 1.6
IRM	90.8 ± 0.9	75.7 ± 2.8	63.3 ± 5.7	61.3 ± 3.2	**82.7 ± 1.2**	**68.6 ± 4.8**	12.4 ± 6.9	9.8 ± 5.1	52.3±3.9	23.4±2.3	14.2±2.0	10.4±3.3	**63.3 ± 0.8**	**55.1 ± 0.6**	30.7 ± 1.2	**19.2 ± 1.6**

Table 11. **Subgroup Analysis.** Model Selection: Worst Class Accuracy

	PlayingCard–10				OddsEvens				CelebA–Jew				COCO–Bag			
	Avg	WG	WC	WS	Avg	WG	WC	WS	Avg	WG	WC	WS	Avg	WG	WC	WS
ERM†	92.4 ± 0.8	81.9 ± 3.9	70.3 ± 6.2	**68.2 ± 5.8**	50.1 ± 0.0	0.0 ± 0.0	50.0 ± 0.0	0.0 ± 0.0	61.2 ± 3.4	37.5 ± 1.3	21.5 ± 3.2	16.8 ± 3.8	64.0 ± 4.7	53.6 ± 8.0	36.4 ± 6.4	21.0 ± 8.9
ERM	93.1 ± 0.2	79.8 ± 3.1	70.0 ± 3.6	61.3 ± 14.1	65.3 ± 1.8	33.0 ± 4.1	55.5 ± 1.7	20.7 ± 1.4	**63.0±2.1**	**40.5±3.3**	**30.4±3.5**	**25.8±2.1**	70.6 ± 5.9	64.4 ± 10.4	42.8 ± 10.5	31.5 ± 16.3
SAM†	**94.8 ± 0.9**	**85.9 ± 3.0**	**77.8 ± 1.3**	65.0 ± 10.0	50.1 ± 0.1	0.0 ± 0.0	50.0 ± 0.0	0.0 ± 0.0	59.8 ± 4.1	31.0 ± 3.2	24.4 ± 6.0	14.5 ± 3.5	63.0 ± 1.1	53.3 ± 2.3	29.0 ± 1.6	15.8 ± 2.6
SAM	92.9 ± 0.4	79.9 ± 3.2	67.8 ± 1.9	64.6 ± 4.4	64.9 ± 0.5	31.3 ± 4.5	**56.7 ± 2.5**	**20.9 ± 7.0**	-62.9 ± 3.5	39.2 ± 4.3	30.2 ± 3.3	24.1 ± 4.3	71.5 ± 4.9	65.9 ± 8.2	**48.8 ± 5.1**	**35.5 ± 9.2**
GDRO	92.0 ± 0.3	80.7 ± 2.6	72.4 ± 3.1	65.0 ± 0.0	69.0 ± 15.6	40.3 ± 31.4	52.7 ± 5.2	19.8 ± 17.5	61.2 ± 1.1	35.2 ± 2.2	30.0 ± 1.7	25.5 ± 4.4	72.5 ± 4.5	65.1 ± 7.9	47.5 ± 6.3	34.0 ± 9.1
IRM	92.2 ± 1.4	80.5 ± 4.4	67.4 ± 14.9	61.7 ± 16.1	63.3 ± 3.4	41.6 ± 7.2	16.5 ± 19.2	8.4 ± 10.9	59.4 ± 4.5	29.9 ± 6.3	26.6 ± 8.3	17.2 ± 3.2	71.4 ± 1.2	64.4 ± 3.6	47.3 ± 0.5	33.0 ± 2.1
V-REx	89.8 ± 1.6	75.0 ± 6.1	61.0 ± 3.0	54.9 ± 7.9	**80.3 ± 1.6**	**66.9 ± 5.4**	10.0 ± 2.9	8.0 ± 3.5	57.1 ± 3.1	25.3 ± 6.3	23.5 ± 3.1	14.4 ± 4.3	68.9 ± 0.6	62.1 ± 1.0	38.6 ± 0.5	24.8 ± 2.5
RDM	88.7 ± 1.1	73.5 ± 4.5	47.7 ± 14.2	47.7 ± 14.2	58.2 ± 9.5	19.7 ± 21.3	35.2 ± 24.7	4.9 ± 3.9	61.7 ± 2.6	37.2 ± 1.7	27.6 ± 4.5	20.3 ± 4.6	**73.5 ± 0.6**	**68.3 ± 2.3**	47.3 ± 2.2	35.3 ± 4.1
URM	92.4 ± 0.7	80.3 ± 2.6	72.2 ± 7.1	58.9 ± 12.1	50.0 ± 0.0	0.0 ± 0.0	50.0 ± 0.0	0.0 ± 0.0	60.5 ± 1.1	34.4 ± 1.5	24.8 ± 4.0	16.8 ± 6.1	72.2 ± 3.5	63.5 ± 4.7	44.9 ± 6.1	28.7 ± 8.2