Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection

Christopher Clarke[1], Matthew Hall[2], Gaurav Mittal[2], Ye Yu[2], Sandra Sanjeev[2], Jason Mars[1], Mei Chen[2]
[1]University of Michigan, Ann Arbor, MI.
[2]Microsoft, Redmond, WA

Problem Statement

Classic approaches to content moderation typically apply a rule-based heuristic approach to flag content. Recent advances in deep learning have demonstrated the promise of using highly effective deep neural models to overcome these challenges. However, despite the improved performance, these data-driven models lack transparency and explainability, often leading to mistrust from everyday users and a lack of adoption by many platforms. In this paper, we present **Rule By Example (RBE)**: a novel exemplar-based contrastive learning approach for learning from logical rules for the task of textual content moderation. RBE is capable of providing rule-grounded predictions, allowing for more explainable and customizable predictions compared to typical deep learning-based approaches.

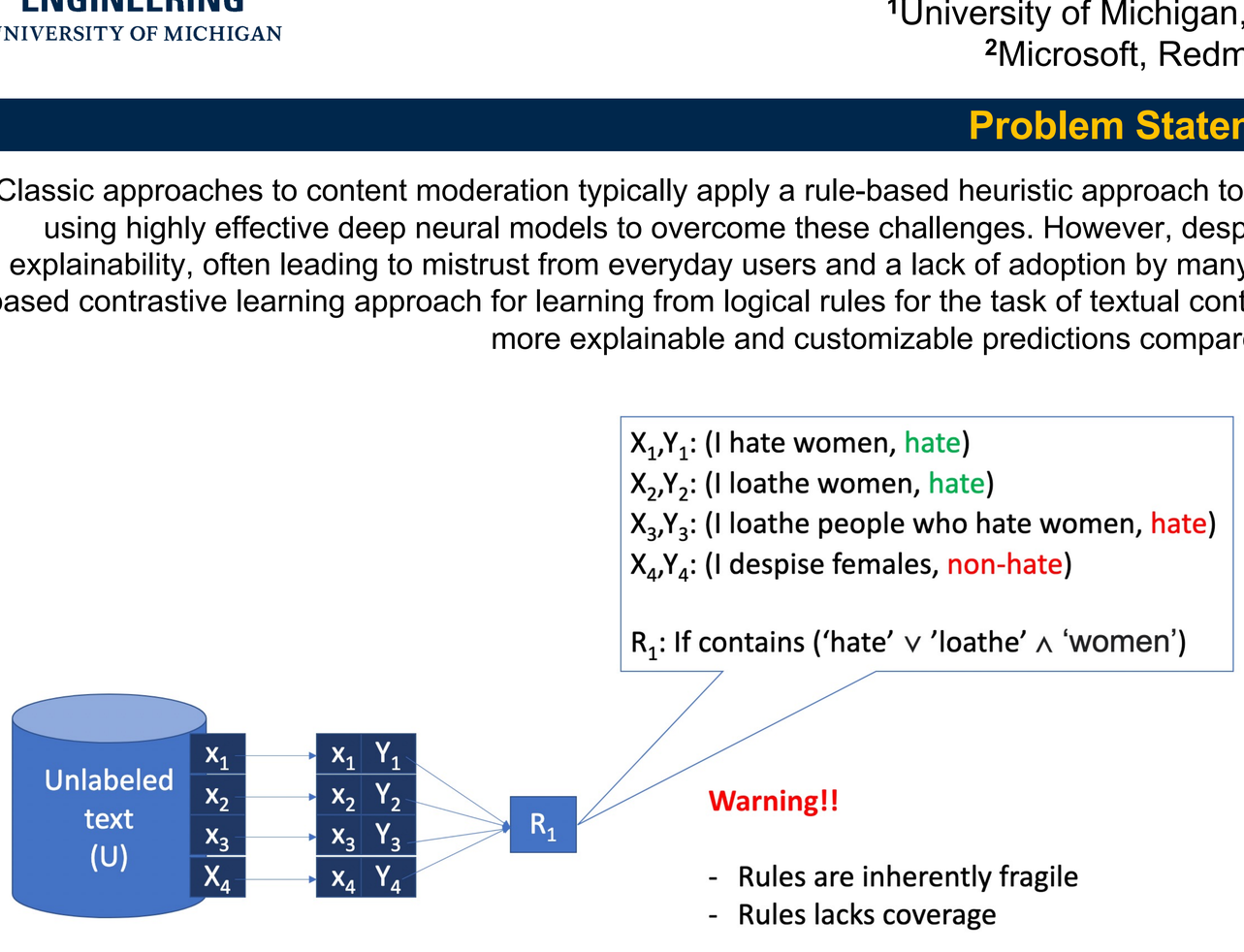


Figure 1: Generalization problem of rules. Logical rules, while easy to explain, are inherently fragile to the nuances of natural language.

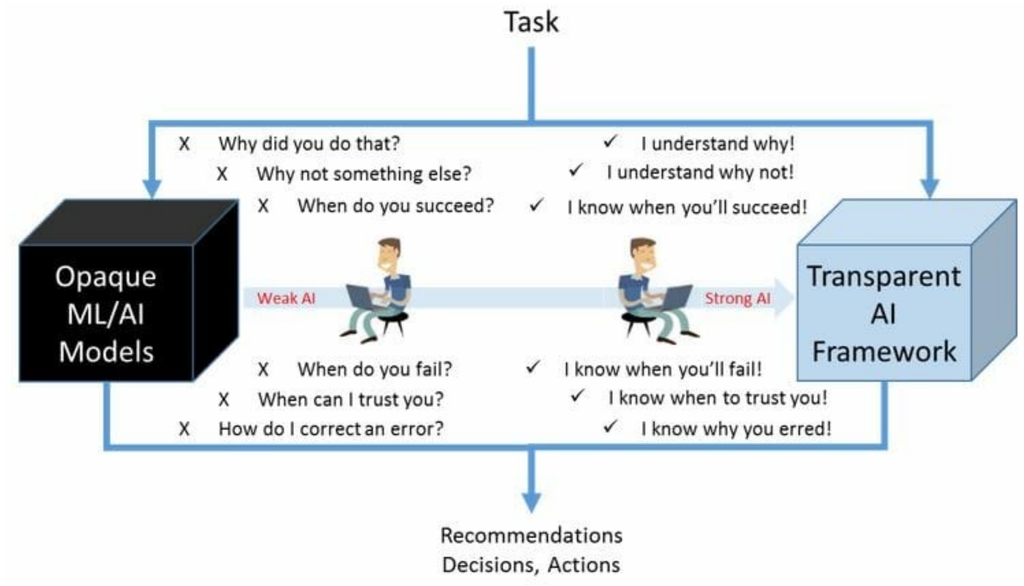
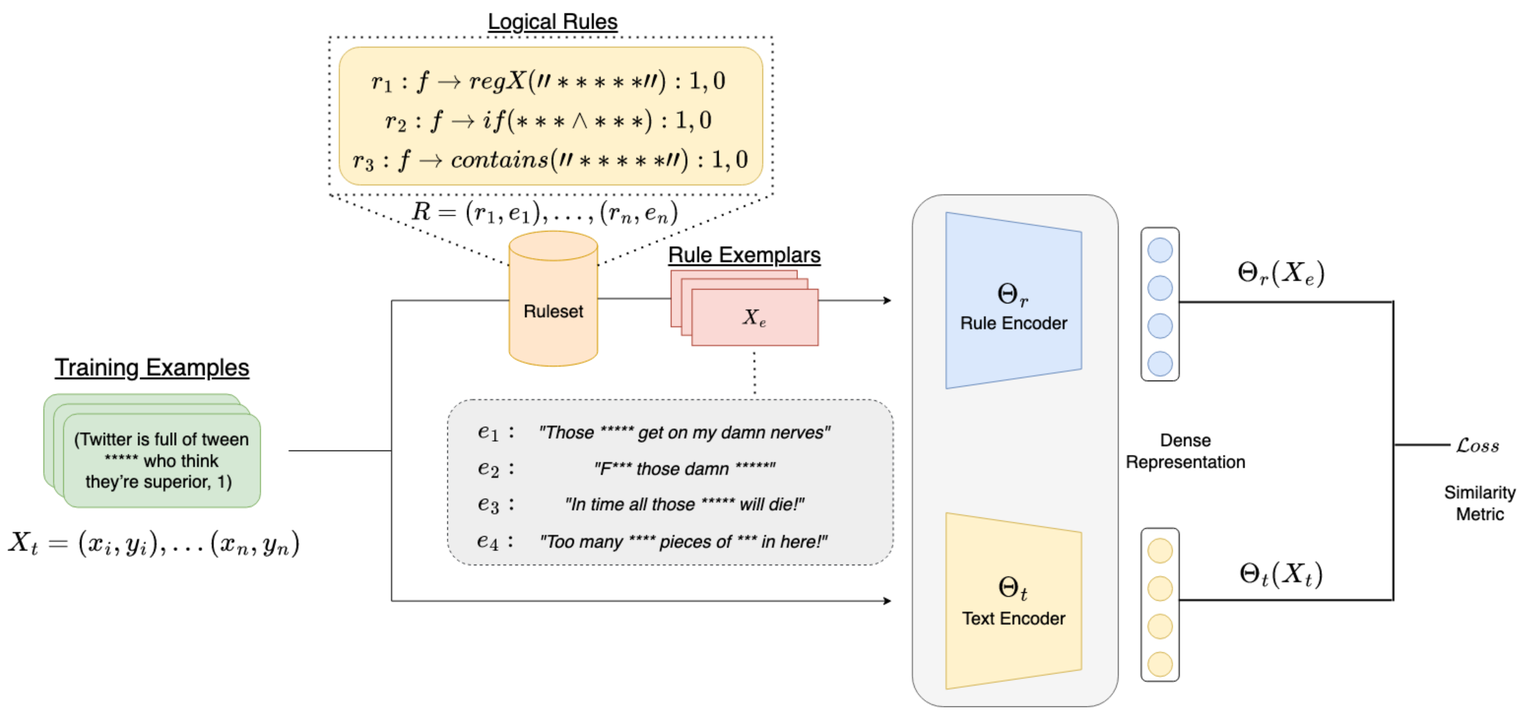


Figure 2: Neural models while performant lack, transparency, customizability, predictability & explainability.

Methods and Approaches



Algorithm 1 Supervised Dual Encoder Training

Require: Rule Encoder Θ_r Text Encoder Θ_t
Input: Training Data $X=(x_1, y_1)...(x_n, y_n)$, Ruleset $R=(r_1, e_1), ..., (r_n, e_n)$
Output: Updated parameters Θ_r, Θ_t

```
Initialize Θ_r and Θ_t
while not converged do
    Get mini-batch X_b
    for each instance x_i in X_b do
        Get exemplars e_i = doRuleset(R, x_i)
        Concatenate exemplars e_i
    end for
    Get Θ_r(E_b) and Θ_t(X_b)
    Compute L = 1/2 (Y_b D^2 + (1 − Y_b) max(margin − D, 0)^2)
    Update parameters of Θ_r and Θ_t
end while
```

Figure 3: Rule By Example Framework (RBE) is comprised of two neural networks, a rule encoder and a text encoder, which jointly learn rich embedding representations for hateful content and the logical rules that govern them. Through Contrastive learning, RBE utilizes a semantic similarity objective that pairs hateful examples with clusters of rule exemplars that govern it.

Results

Content Moderation Using Rules (Fully Supervised)			
	HateXplain	Jigsaw	CAD
Model	**F1**	**F1**	**F1**
HateXplain Rationale Rules	0.752	-	-
Hate+Abuse Rules	0.719	0.226	0.290
CAD Rules	-	-	0.194
MPNet SeqCLS	0.823	0.581	0.463
BERT SeqCLS	0.824	0.563	0.433
Rule By Example (BERT|CAD Rules)	-	-	0.435
Rule By Example (MPNet|CAD Rules)	-	-	**0.476**
Rule By Example (BERT|Hate+Abuse Rules)	0.824	0.602	0.445
Rule By Example (MPNet|Hate+Abuse Rules)	**0.837**	**0.604**	0.476
Rule By Example (BERT|HateXplain Rationale)	0.816	-	-
Rule By Example (MPNet|HateXplain Rationale)	0.832	-	-

Table 1: Experiment Results in Fully Supervised Setting on hate speech classification datasets.

Rule Grounding

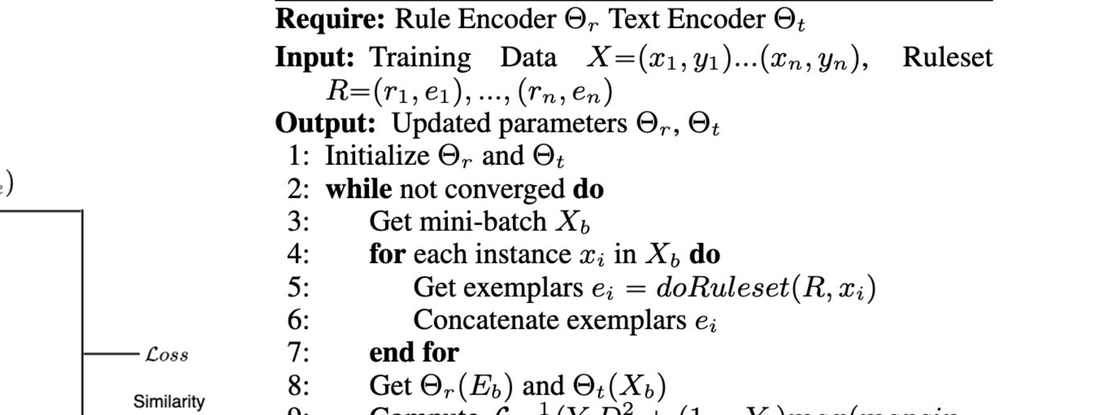


Figure 4: By displaying the rules and exemplars responsible, rule authors and users are better able to understand model predictions and can automatically adjust their ruleset to further improve model performance.

Acknowledgements

We thank our anonymous reviewers for their feedback and suggestions. This work was conducted by the ROAR (Responsible & Open AI Research) team at Microsoft Cloud & AI.

Repo & Dataset

We release with RBE each of our derived rulesets for the HateXplain & Contextual Abuse datasets as well as a small subset of our internal Hate+Abuse list for public use! 😃

SCAN ME!