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]

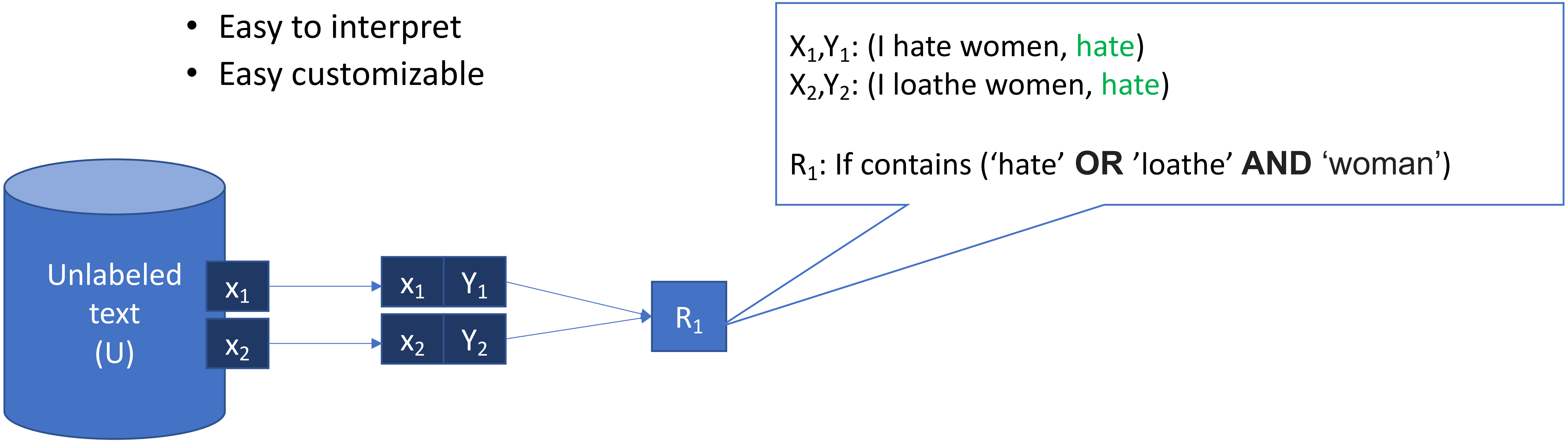
[1]University of Michigan, Ann Arbor, MI.

[2]Microsoft, Redmond, WA

Warning! These slides contain content that may be offensive or upsetting.

Background

- Modern approaches to content moderation typically apply a rules-based heuristic to flagging content.
  - Advantages:
    - Easy to interpret
    - Easy customizable

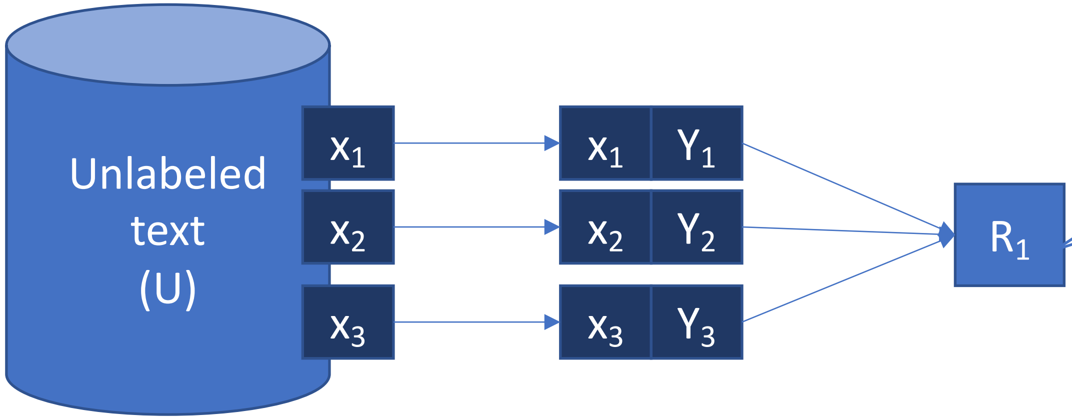


Background

- Modern approaches to content moderation typically apply a rules-based heuristics
  - Advantages:
    - Easy to interpret
    - Easy customizable

X_1,Y_1: (I hate women, hate)
X_2,Y_2: (I loathe women, hate)
X_3,Y_3: (I loathe people who hate women, hate)

R_1: If contains (‘hate’ $\lor$ ’loathe’ $\land$ ‘woman’)

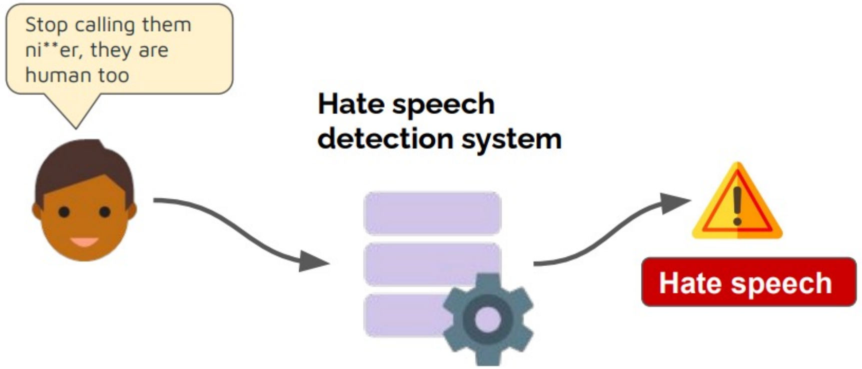


Warning!!

- Rules are inherently fragile
- Rules lacks coverage

Background: Deep Learning Approaches

- Neural models while performant lack:
  - Transparency
  - Customizability/Personalization
  - Predictability/Explainability



- *"60% of users would prefer social media companies provide users with greater choice and control over the content they see"* (CATO, 2021)
- *"Explainability is a crucial aspect in social dimensions"* (Mukherjee et al. 2022)

75% Don't Trust Social Media to Make Fair Content Moderation Decisions, 60% Want More Control over Posts They See | Cato Institute (Dec 2021)

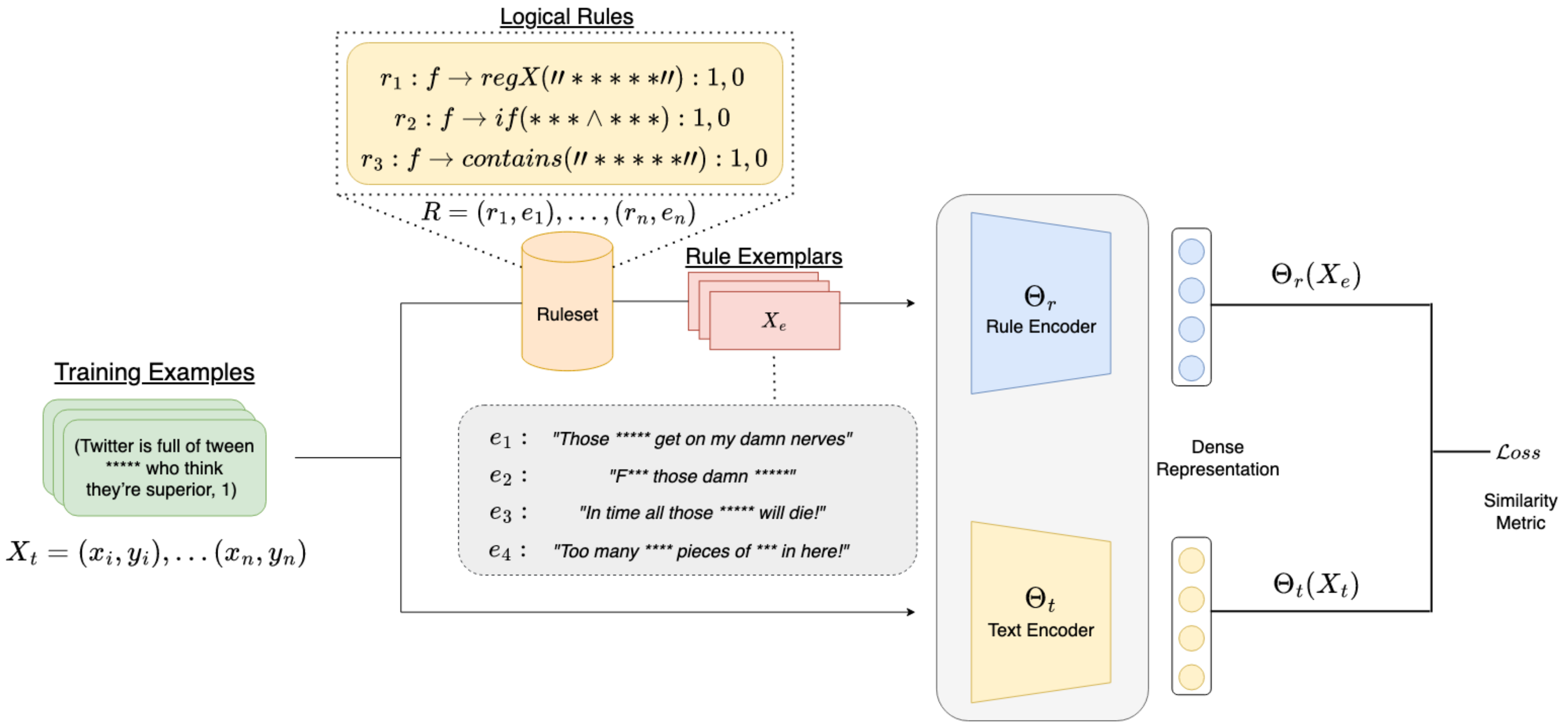
Research Questions/Goals

How can we combine the best of both worlds for HateSpeech Detection?

Goals:

Maintain customizability, transparency & predictability of logical rules	Improve flexibility, coverage and scale of rules by leveraging deep learning.	Provide applicability in scenarios with and without labeled data

Rule By Example Framework



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.

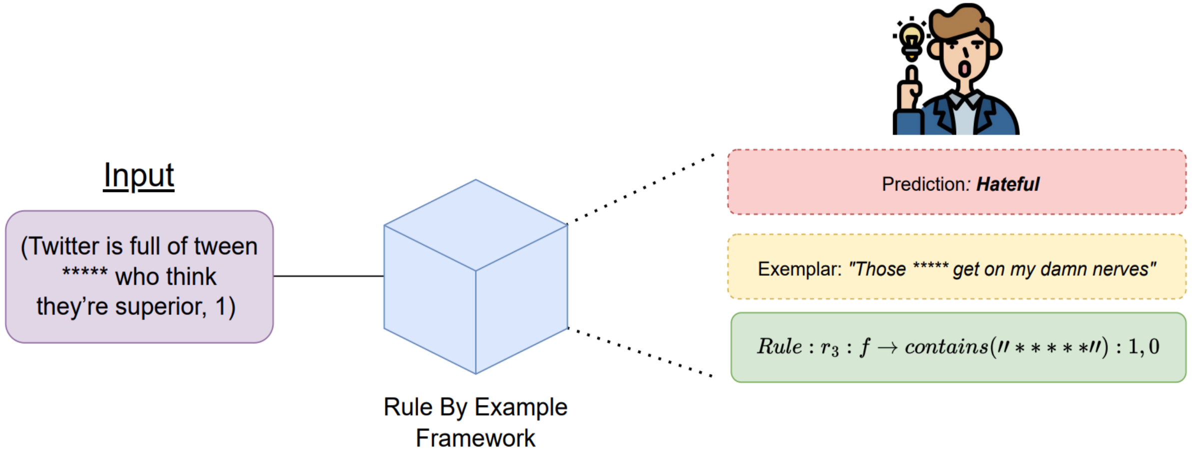
Results

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	-	-

Highlight: RBE outperforms SOTA model on F1-score by up to 4% on three popular hate speech classification datasets.

Advantage: Rule Grounding

Rule By Example Rule Grounding			
Dataset	**Text**	**Fired Rules**	**Exemplar**
HateXplain	fully agree every personal interaction with these queers reinforces what i already knew severe mental illness and obnoxious to boot	if contains("queers") $\rightarrow 1, 0$	yes but queers are too self righteous to let them be included originally the gay and pedophile communities were working together nambla was started by gay men who liked boys now all subsets of pedophiles are members assuming nambla still exists
Jigsaw	Why do they put so many gay people on the damn show since when it was okay to be gay.	if contains("so" $\wedge$"gay") $\rightarrow 1, 0$	stop reverting my edit your so f**** gay get a f**** life your f**** or go get laid or something
CAD	What a little b****	if contains("b****") $\rightarrow 1, 0$	Nope, today is tuna b****



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

Conclusion

- 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!