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

期刊： ACL-2023-1

针对场景、问题：

- content moderation， 检测不良言论
  - rule-based方法 : 过于特异性，或者过于宽泛
  - data-driven 方法：
    - 学习到了更丰富的表达，对未知的数据有着更好的泛化性
    - adoption challenges due to the inability of customers to understand how these complex models reason about their decisions
    - data-driven models lack transparency and explainability, ofen leading to mistrust from eveyday users and a lack of adoption by many platforms.
- the forefront of demands: explainability and transparency
- challenging open question: how we can leverage the robustness and predictive performance of complex deep-learning models whilst allowing the transparency, customizability, and interpretability that rule-based approaches provide?

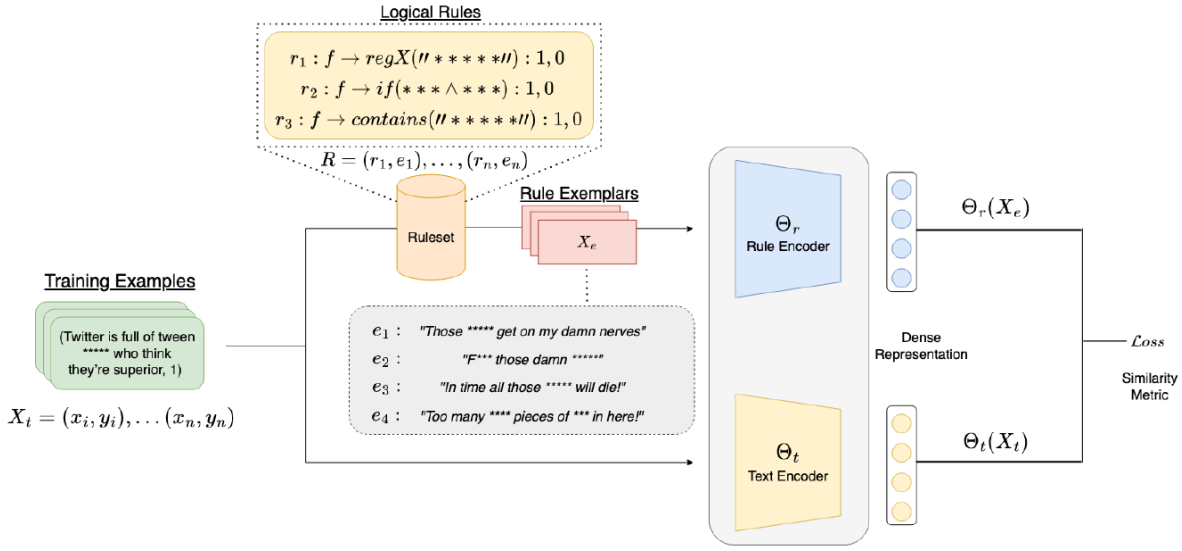
finding:

找到了可以把rule-based和deep learning结合的办法，解决了rule-based的泛化性差和deep-learning的可解释性差的问题

contribution:

- Rule By Example (RBE)—— a novel exemplar-based contrastive learning approach for learning from logical rules for the task of textual content moderation
  - comprised of 2 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
  - advantage
    - RBE is capable of providing rule-grounded predictions, allowing for more explainable and customizable predictions compared to typical deep learning-based approaches.
    - RBE is capable of learning rich rule embedding representation using only a few data examples.
- demonstate how RBE can be easily integrated to boost model F1-score by up to 4% on three popular hate speech classification datasets.
- a detailed analysis and insights into the customizability and interpretability features of RBE to address the problem of merging hateful content and model transparency.

RBE方法



- pre-indexing of exemplars ---> faster inference at runtime after training

是否开源：

是

https://github.com/ChrisIsKing/Rule-By-Example

这种方法相较于其他方法的优越性：

- REB is able to outperform state-of-the-art deep learning classifiers as well as the use of rules in both supervised and unsupervised settings while providing explainable model predictions via rule-grounding

数据集：

3 popular hate speech classification datasets

- HateXplain
  - a large-scale benchmark dataset for explainable hate speech detection.
  - 20k samples across 3 labels “hateful”,”offensive”, “normal”
  - ruleset construction
- Jigsaw
  - a large-scale dataset of Wikipedia comments labeled by human raters for toxic behavior
  - 230k samples (“toxic”, “severe toxic”, “obscene”, “insult”, “identity hate”)
- Contextual Abuse Dataset(CAD)
  - annotated dataset of 25k Reddit entries

Construct Ruleset，是用自己构建的规则集上训练，然后在上面三个流行的数据集上测试结果

- Hate+Abuse List
  - n-grams representing harmful languages such as slurs or hate verbs
  - consists of 2957 distinct identity hate rules
- HateXplain Rationale Ruleset //不太明白

    - extract 1, 2, and 3-gram substrings from the annotator rationales and cluster them by annotator-identified target demographic groups.
    - take the top N n-grams per each demographic group and automatically create rules for each of them
    - 670 distinct rules
- Contextual Abuse Rationale Ruleset
    - similar to HateXplain Rationale Ruleset
    - 2712 distinct rulse for CAD
- Exemplar Selection
    - pairing each rules with accompanying exemplars
        - run our Ruleset on the dataset trainset and extract instances for which a rule correctly fires.
        - for each tule that correctly fires, we then randomly select N instances to act as the exemplars
        - 保证两条规则不能映射到同一个范例——避免潜在的过于笼统的规则（除非另有说明，否则均利用一条范例一个规则）

实验

- 无监督下设计了三个评价指标，在实验中都测了
    - Mean
    - Concat
    - Distance clustering
- precision, recall, F1 score(main), acc
- 结论：
    - 两个主要因素
        - Ruleset Size and Alignment
        - Embedding Initialization

个人理解&不理解：

1. 365页左半部分prior开头段第9行soft implication loss是什么
    1. 不重要，一般hard是指只有0,1，而soft指还可以有其他
2. 367页（2）公式，搜索了对比学习常用的随时函数公式，和这里的不太一样，不是很理解公式的含义
    1. 就是普通的函数loss公式

2. Let Y_t be the correct label of the texts X_t, D be the cosine distance of (x_e, x_t) and m be the margin, our contrastive learning loss function is defined as follows:

$$\mathcal{L} = \frac{1}{2}(Y_t D^2 + (1 - Y_t)max(m - D, 0)^2) \qquad (2)$$

    3. Yt是label，值为0或者1，如果是1的话取D^2,D^2越小越好（loss函数），如果是0的话取max(m-D,0)，D应该大于m，那么取0。
3. 368页第一段数据集的构建中，Jigsaw只将identity hate定义为hateful，是否公平客观？我认为对于其他有害的言论也应该检测出来，同理对CAD数据集页不理解
    1. hate指的是讨厌，暴力言论，网络暴力之类的东西，并不是不良言论
4. n-gram部分还需要再看一下
    1. 表示几元的item（eg: hate x plain 1元
        1. hatex xplain 2元
            1. hatexplain 3元

HateXplain Rationale Ruleset Using the labeled annotator rationales included in the HateXplain dataset, we programmatically generate a Ruleset for HateXplain. To do so, we extract 1, 2, and 3-gram substrings from the annotator rationales and cluster them by annotator-identified target demographic groups. We then take the top N n-grams per each demographic group and automatically create

1. 369页 4.1中的F1-score 1.3/1.4%是什么意思
    1. 这都没仔细看吗？是对于Bert的增长啊，4.1是RBE的，2.3是bert的

On Jigsaw and Contextual Abuse datasets using the Hate+Abuse List and derived CAD Ruleset, RBE outperforms SOTA by an increased margin of 4.1/2.3%, and 4.3/1.3% respectively. Contrary to HateXplain, these two datasets are more heav-

下面这里 引起的问题没看懂

不重要

8 Ethics

Hate speech detection is a complex task. Reducing the task to authoring a set of simple logical rules can potentially lead to rule authors encoding hard biases in those rules. This can cause problems of erasure, for example, if an in-group word or an identity term is used as a rule to identify content as hate speech.

The Rule by Example method can potentially

不会找finding可怎么办啊

嗯，看为什么提出这个方法