Explanation-based Learning for NLP Models

Dong-Ho Lee

*Slide contains harmful contents

Problem of "**label**" only dataset

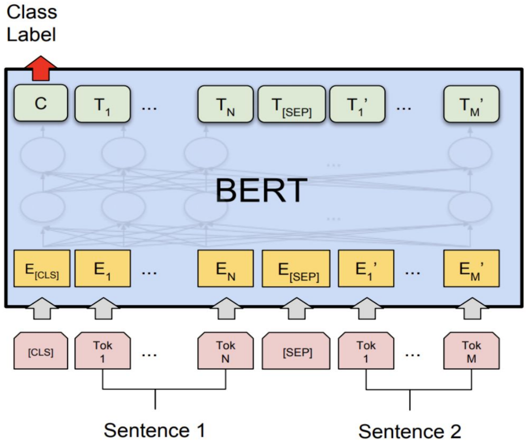
black = HATE ?

whether a scientist or janitor, **black** people are all obedient brutes … **HATE**

blacks have been programmed to love watermelon … **HATE**

black people from the middle ages have always been watermelon-eating … **HATE**

Labeled Dataset
(Machine-Generated Data)



Model may sensitive to **spurious correlations**

ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection., Hartvigsen et al., ACL 2022

*Slide contains harmful contents

Labeling Explanation

whether a scientist or janitor, **black** people are all obedient brutes … **HATE**

WHY is this **HATE**?

whether a scientist or janitor, black people are all obedient brutes …

Salient Spans

Zaidan et al., 2007
Dua et al., 2020

It implies that black people are submissive.

Natural Language

Camburu et al., 2018
Rajani et al., 2019

Using Annotator Rationales to Improve Machine Learning for Text Categorization., Zaidan et al., NAACL 2007
Benefits of Intermediate Annotations in Reading Comprehension., Dua et al., ACL 2020
e-SNLI: Natural Language Inference with Natural Language Explanation., Camburu et al., NeurIPS 2018
Explain Yourself! Leveraging Language Models for Commonsense Reasoning., Rajani et al., ACL 2019

*Slide contains harmful contents

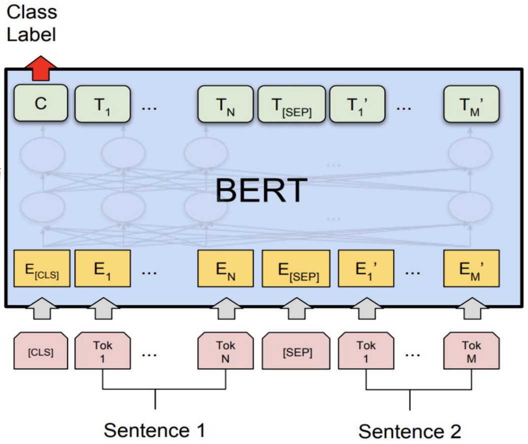
Can we leverage explanation?

whether a scientist or janitor, **black** people are all obedient brutes … **HATE**

WHY is this **HATE**?

whether a scientist or janitor, black people are all obedient brutes …

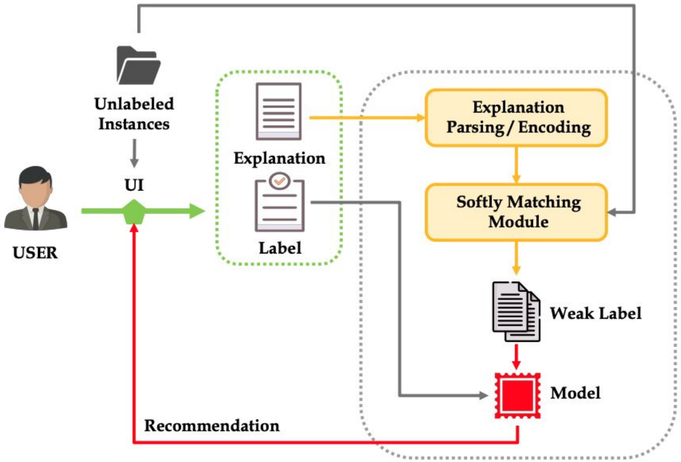
It implies that black people are submissive.



1) Leveraging explanation can accelerate model training?
2) Can we align human explanation with model explanation?

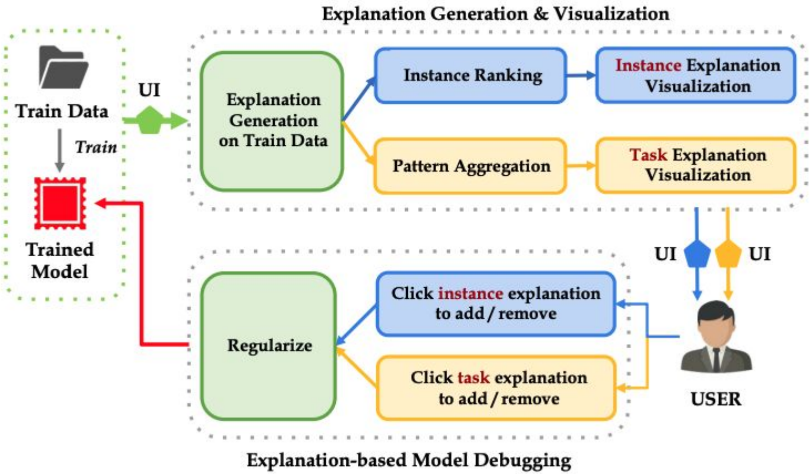
This Talk

RQ1) Label-efficient training with human explanation



LEAN-LIFE (Lee et al., ACL 2020 Demo)
TriggerNER (Lee et al., ACL 2020)
NExT (Wang et al., ICLR 2020)

RQ2) Explanation-based Model Debugging



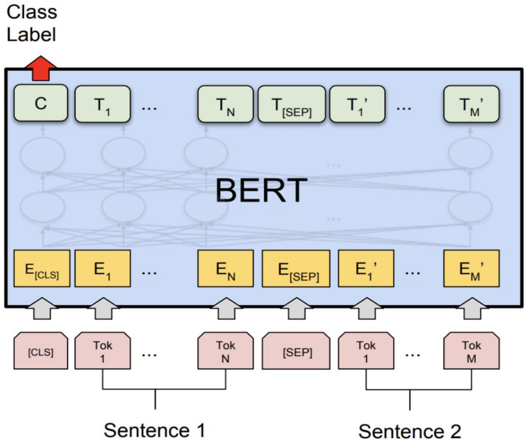
XMD (Lee et al., EMNLP 2022 Demo Submission)
ER-Test (Joshi et al., TrustNLP@NAACL 2022)

Label-Efficient Training with Human Explanation

Simple Recipe for Modern NLP

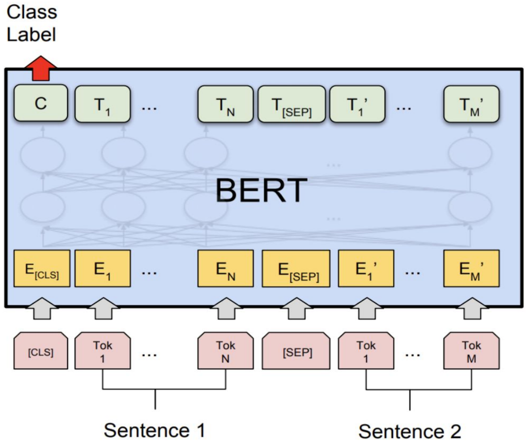
Computing Power

Labeled Dataset



Model

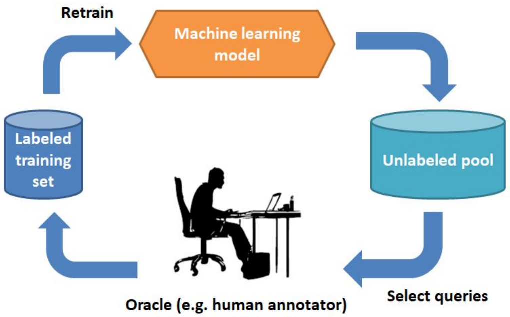
Expensive Cost of Labeled Data



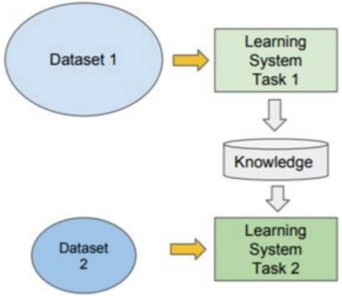
Computing Power | **Labeled Dataset** | **Model**

Model and **Computing power** are transferable across applications,
but **labeled data** is not. Humans need to annotate for each application.

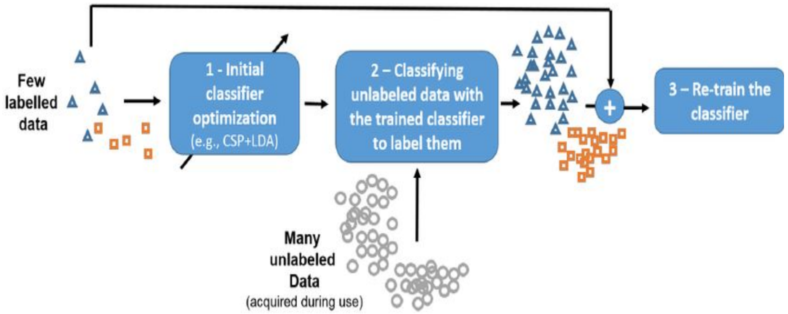
Previous Efforts Toward Label-Efficient Learning



Active Learning



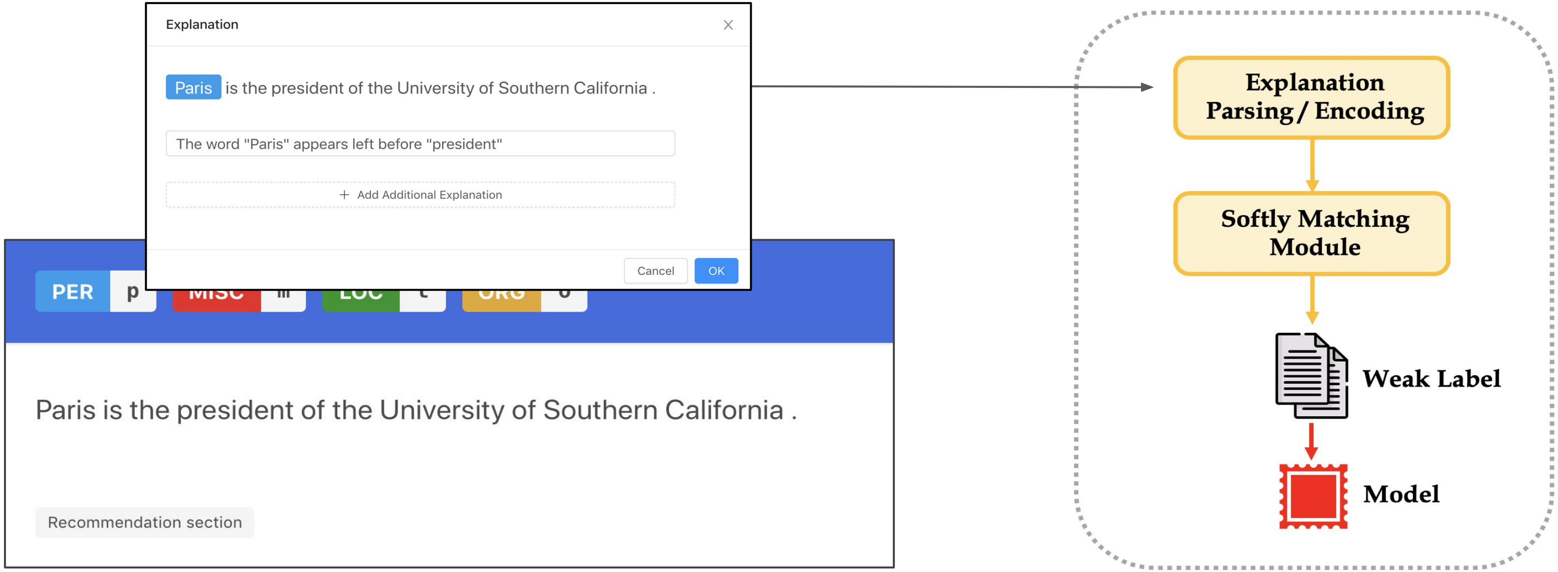
Transfer Learning



Semi-Supervised Learning

Capture and Leverage High-level Supervision

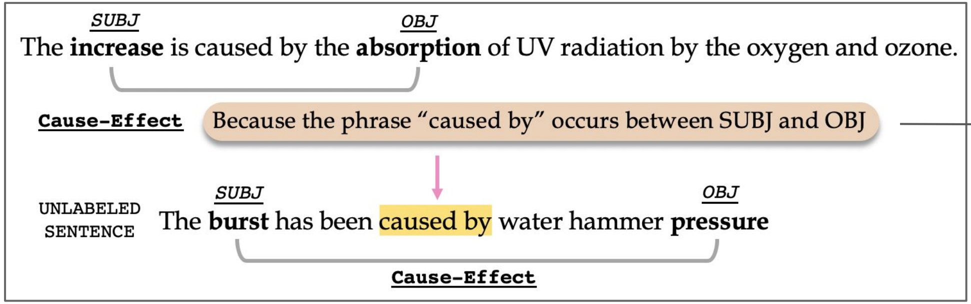
Easy to Use Annotation Framework | **Faster learning w/ Explanations**



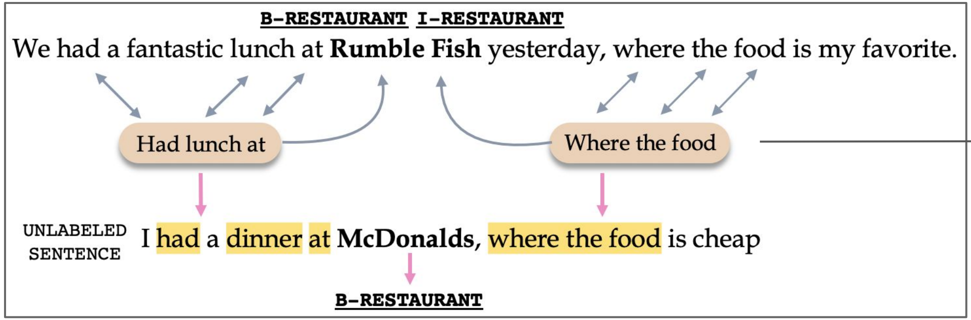
LEAN-LIFE (Lee et al., ACL 2020 Demo)

TriggerNER (Lee et al., ACL 2020) -> NER
NExT (Wang et al., ICLR 2020) -> RE

Form of High-level Supervisions (Explanation)

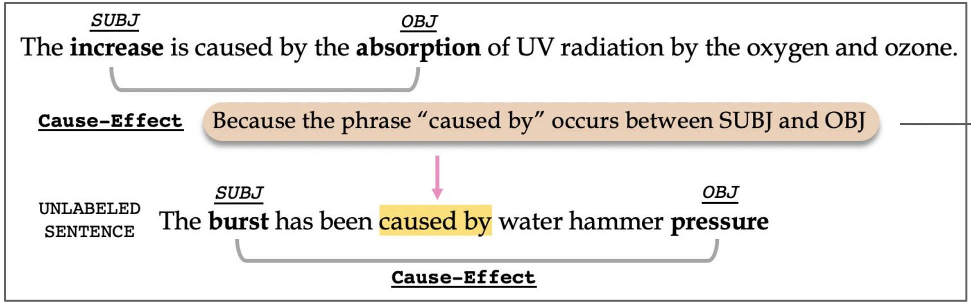


Natural Language



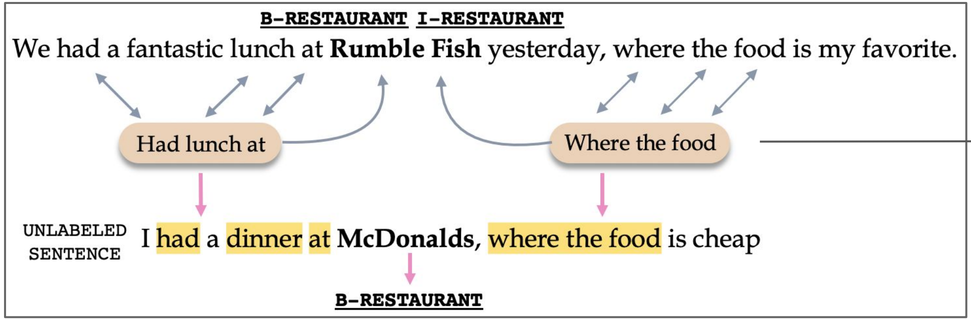
Trigger (Rationale, Attention)

Form of High-level Supervisions (Explanation)



Natural Language

Neural Execution tree
for NL explanation
(Wang et al., ICLR 20)



Trigger
(Rationale, Attention)

TriggerNER
(Lee et al., ACL 20)

Neural Execution tree for Soft Matching

Explanation Parsing / Encoding

Logical Form

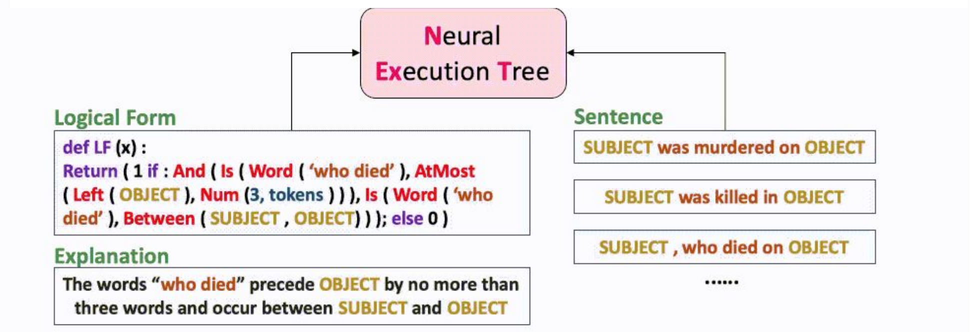
```
def LF (x) :
Return ( 1 if : And ( Is ( Word ( 'who died' ), AtMost
( Left ( OBJECT ), Num (3, tokens ) ) ), Is ( Word ( 'who
died' ), Between ( SUBJECT , OBJECT) ) ); else 0 )
```

Explanation

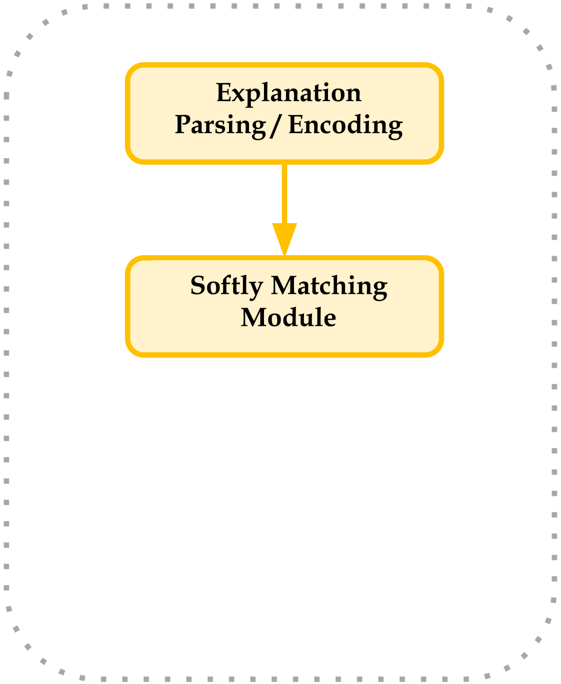
The words "who died" precede OBJECT by no more than three words and occur between SUBJECT and OBJECT

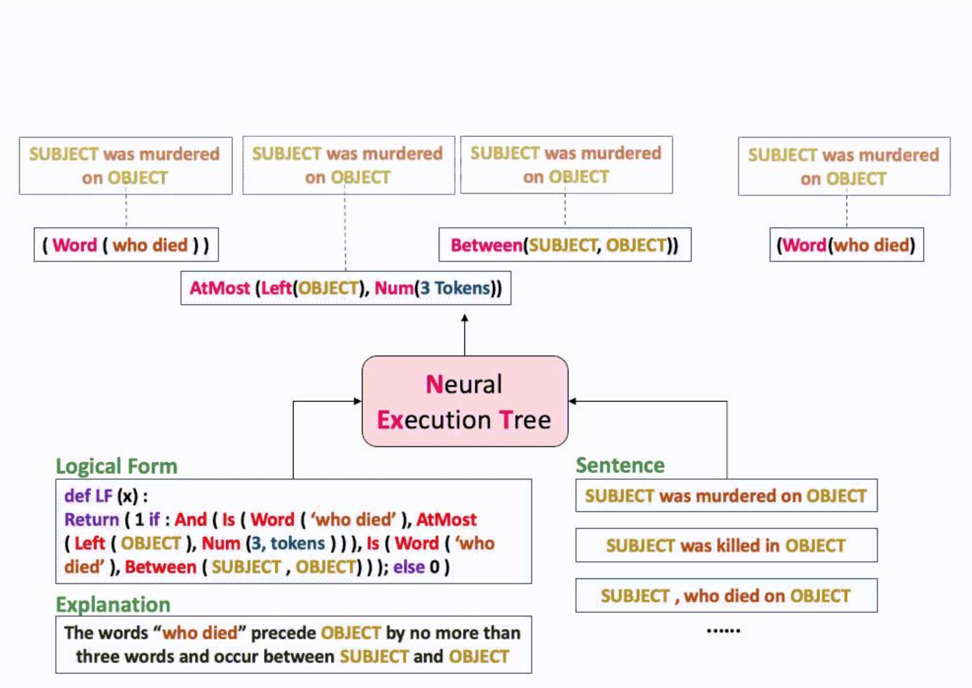
Neural Execution tree for Soft Matching

Explanation
Parsing / Encoding

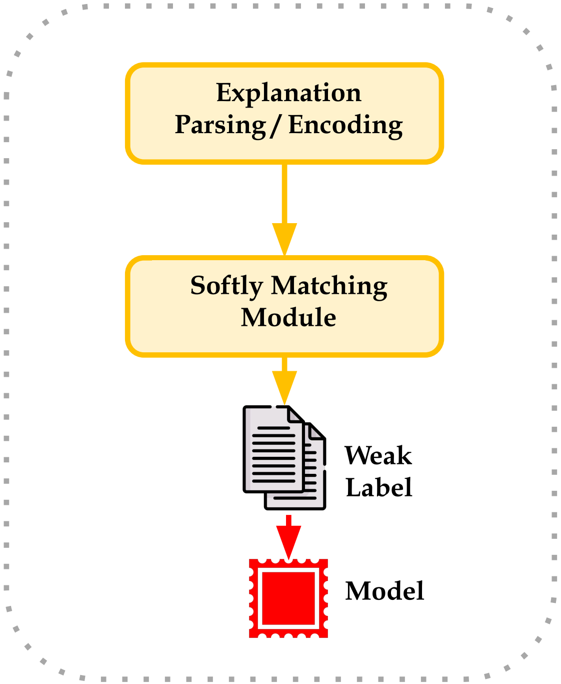


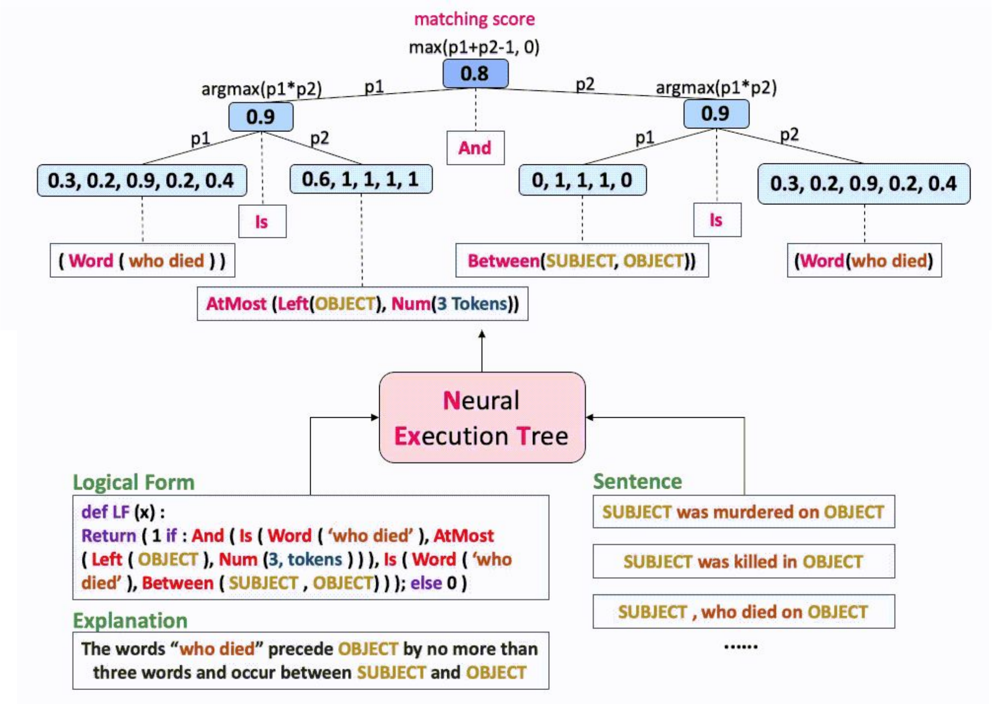
Neural Execution tree for Soft Matching



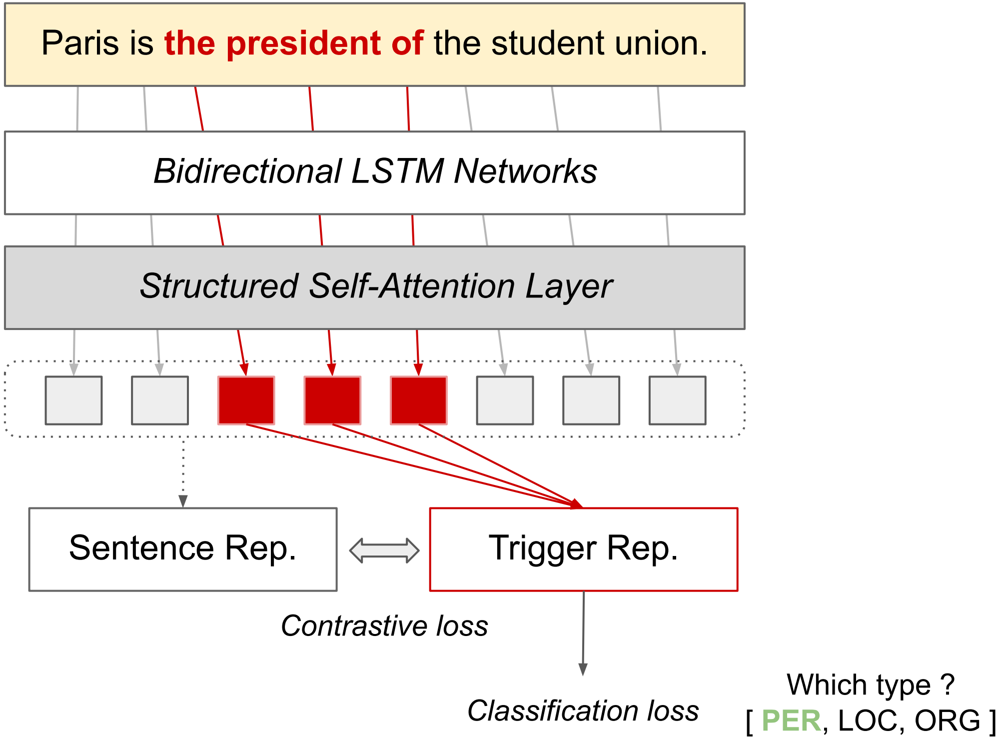


Neural Execution tree for Soft Matching

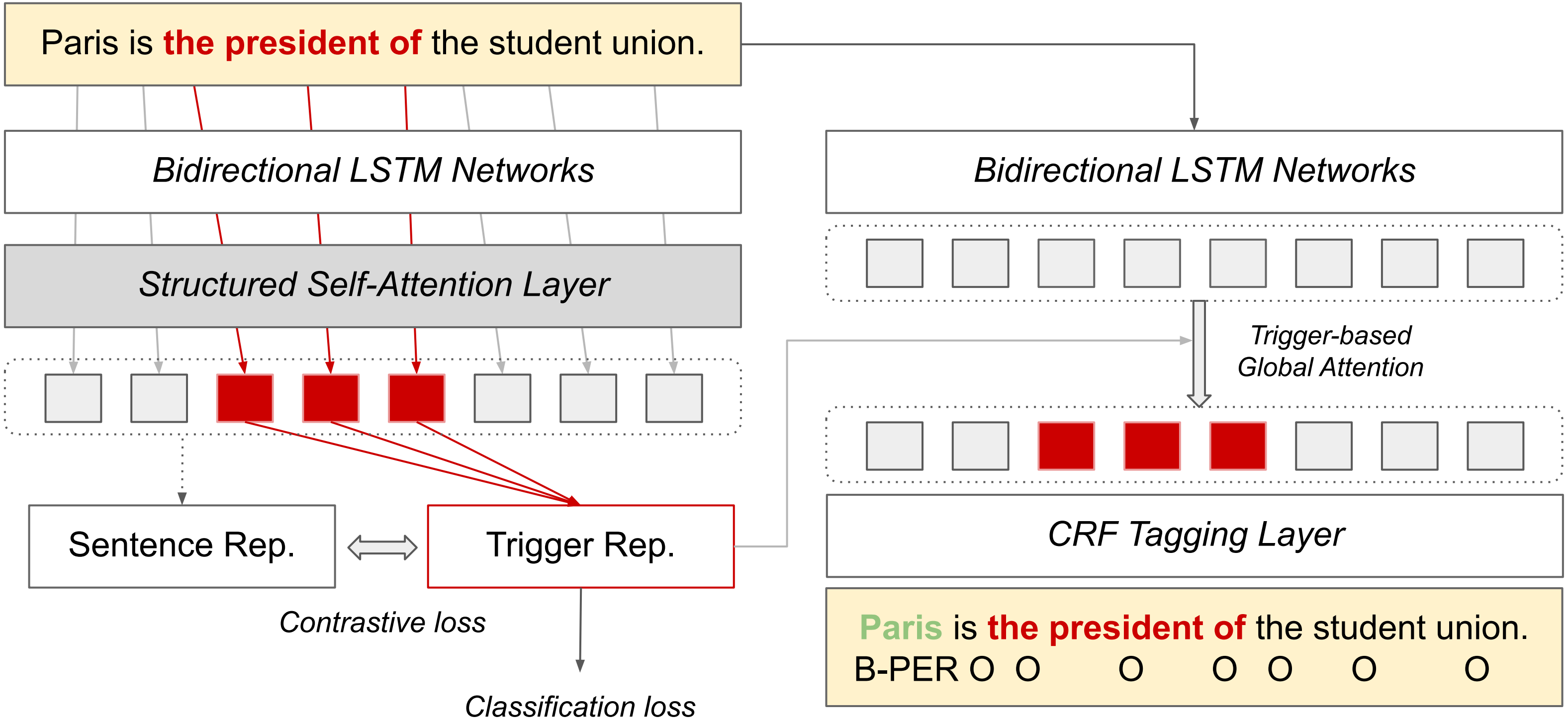




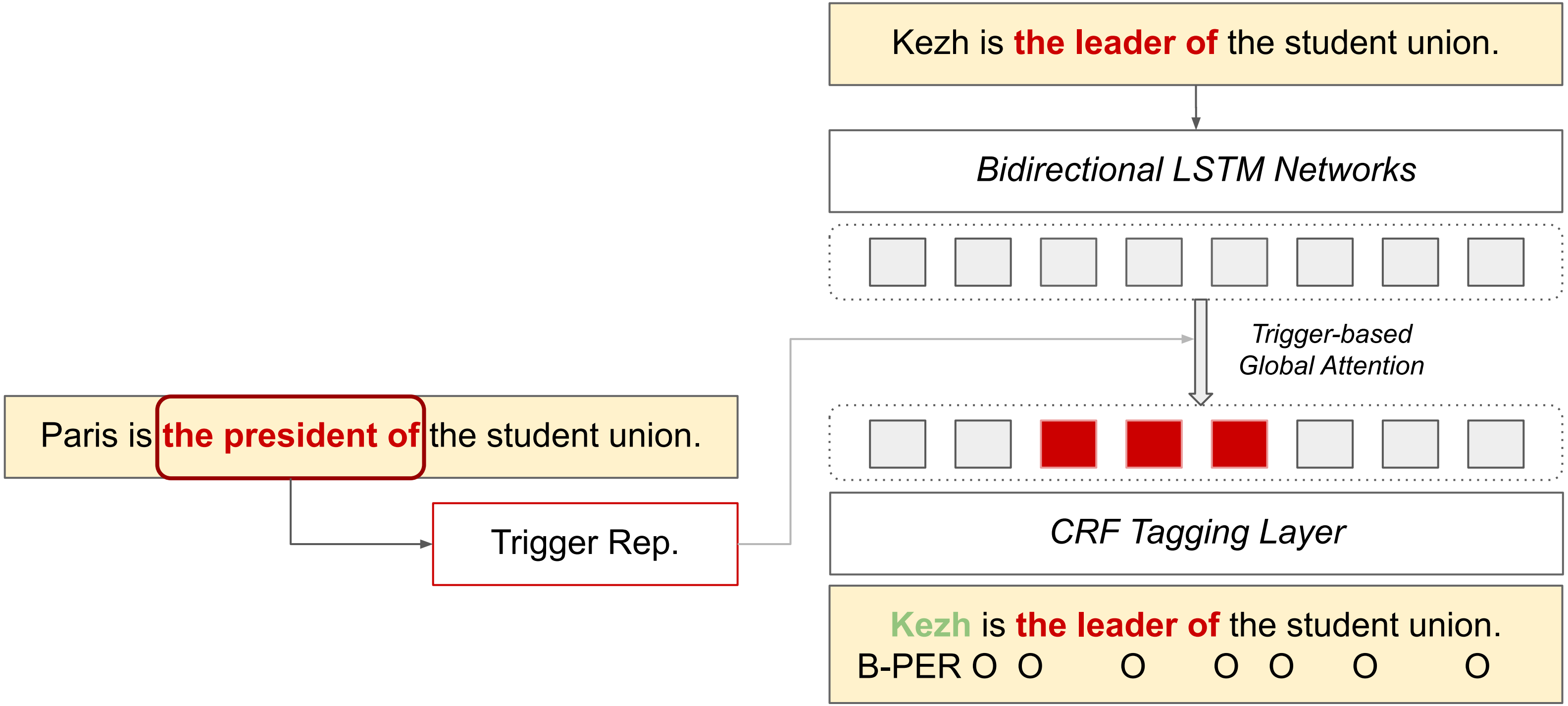
TriggerNER (Train)



TriggerNER (Train)

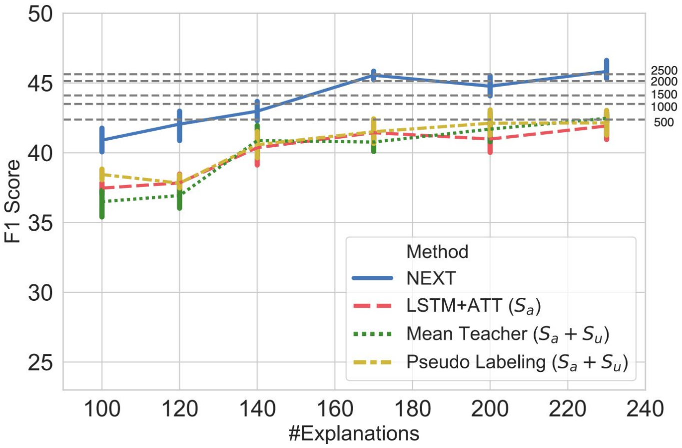


TriggerNER (Inference)



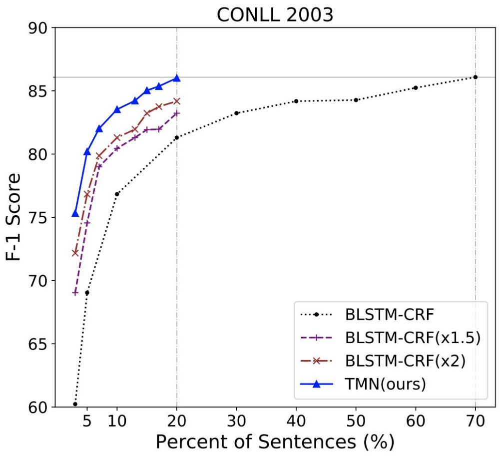
Label Efficiency

NeXT., ICLR 20



Annotation time cost :
Label + Explanation ~= 2X label

TriggerNER., ACL 20

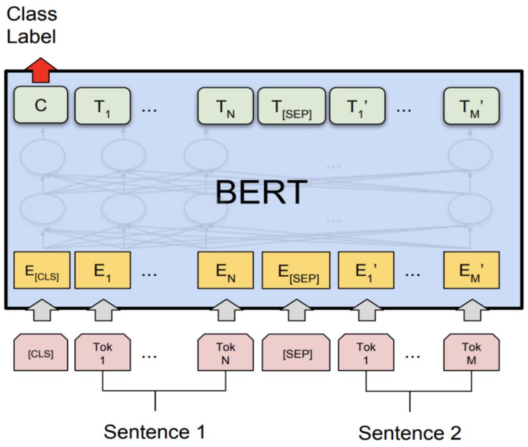


Annotation time cost :
Label + Trigger ~= 1.5X Label

Explanation-based Model Debugging

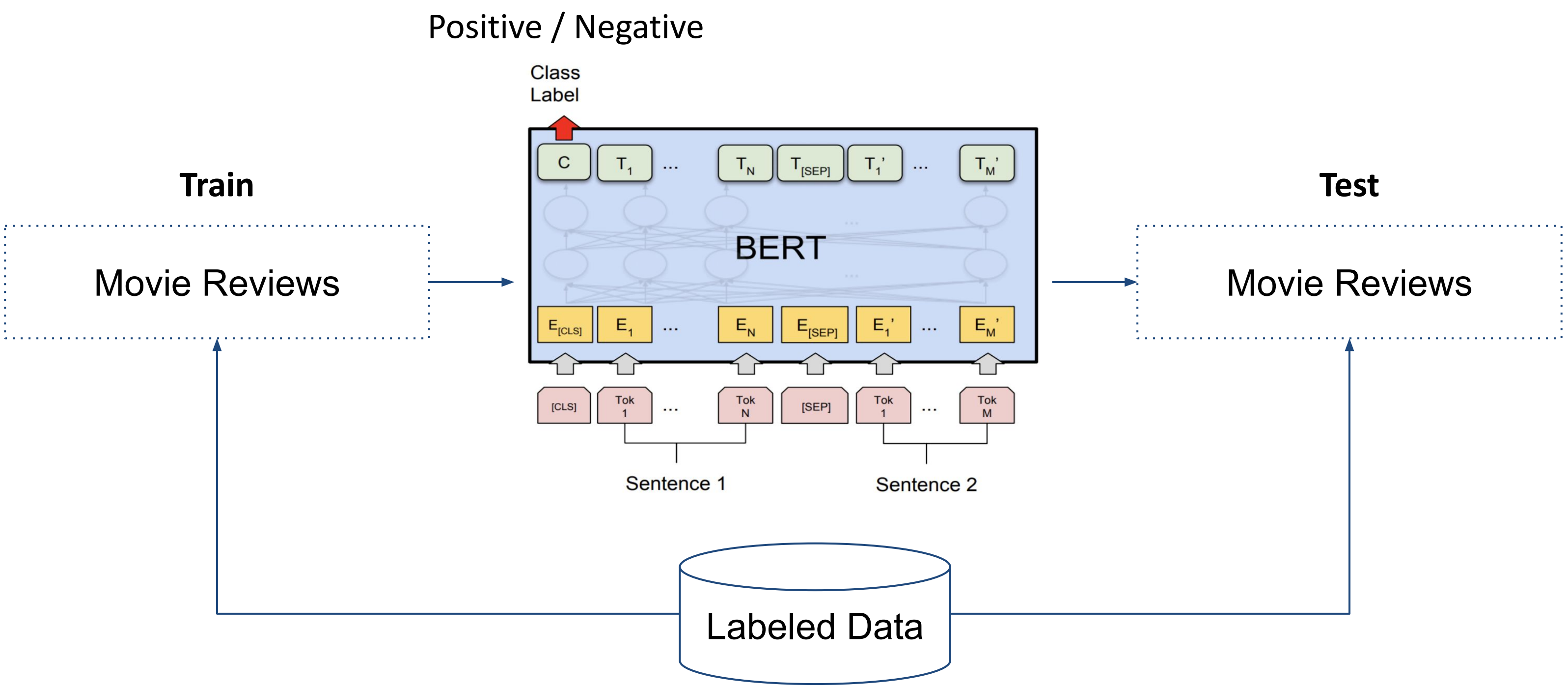
LM Performs well on ID Test set

Positive / Negative



LM Performs well on ID Test set

ID: Identically Distributed



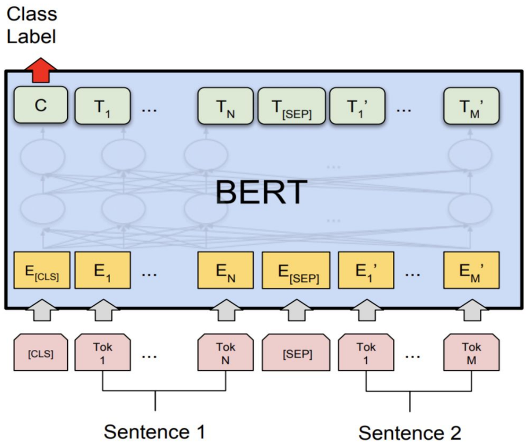
LM Performs well on ID Test set

ID: Identically Distributed

Positive / Negative

SST2 Train

Movie Reviews
(Rotten Tomatoes)



SST2 Test

Movie Reviews
(Rotten Tomatoes)

Model	URL	Score	CoLA	SST-2
Vega v1		91.3	73.8	97.9
Turing NLR v5		91.2	72.6	97.6
DeBERTa + CLEVER		91.1	74.7	97.6
ERNIE		91.1	75.5	97.8
StructBERT + CLEVER		91.0	75.3	97.7
DeBERTa / TuringNLRv4		90.8	71.5	97.5

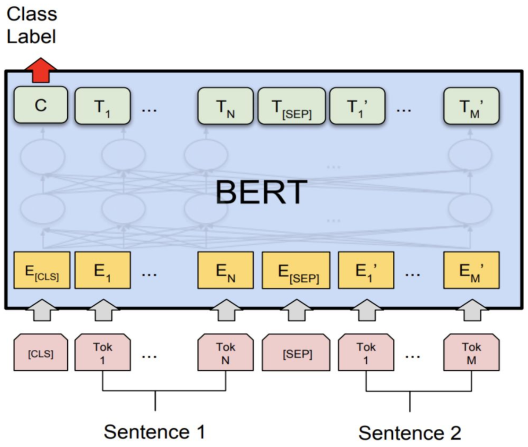
LM Performs well on OOD Test set ?

OOD: Out-of-Distribution

Positive / Negative

SST2 Train

Movie Reviews
(Rotten Tomatoes)



Amazon Reviews

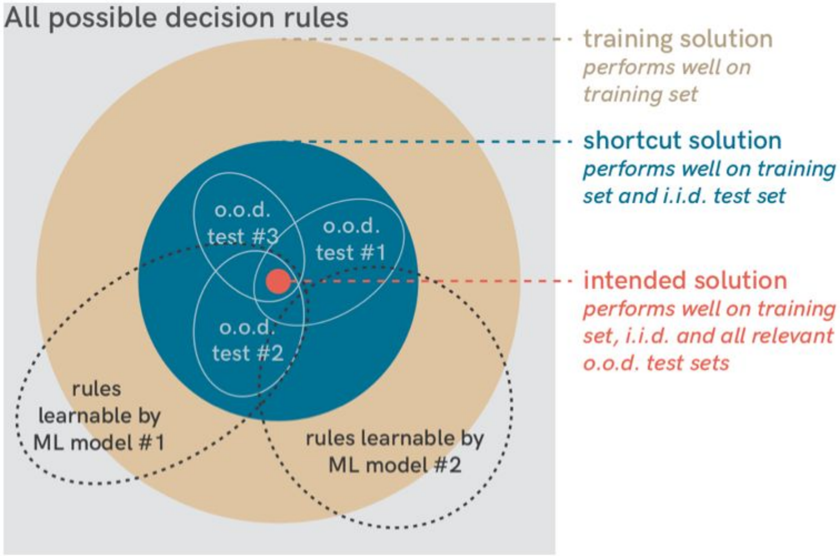
Yelp Reviews

Movie Reviews
(IMDB)

Sentiment Analysis			
In-distribution	Out-of-Distribution		
SST	Amazon	Yelp	Movies
93.4	89.1	89.0	82.0

Bias in NLP Model

Shortcut Learning



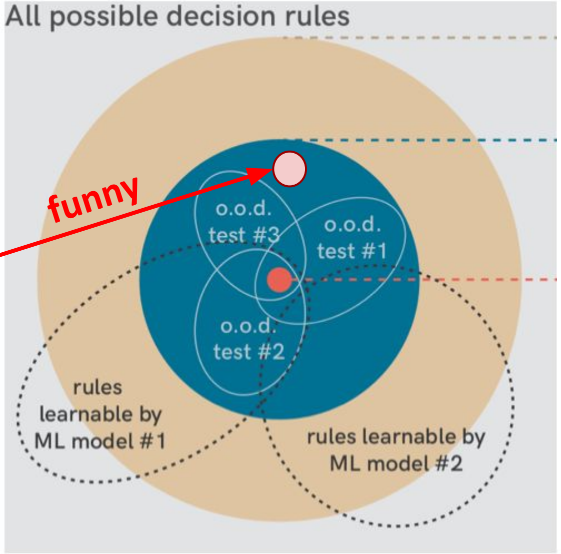
Shortcut Learning in Deep Neural Networks., Geirhos et al., 2020

Bias in NLP Model

Shortcut Learning

Rich veins of **funny** stuff in this movie! (Positive)
Is pretty **funny**. (Positive)
Very **funny** film (Positive)

Movie Reviews
(Rotten Tomatoes)



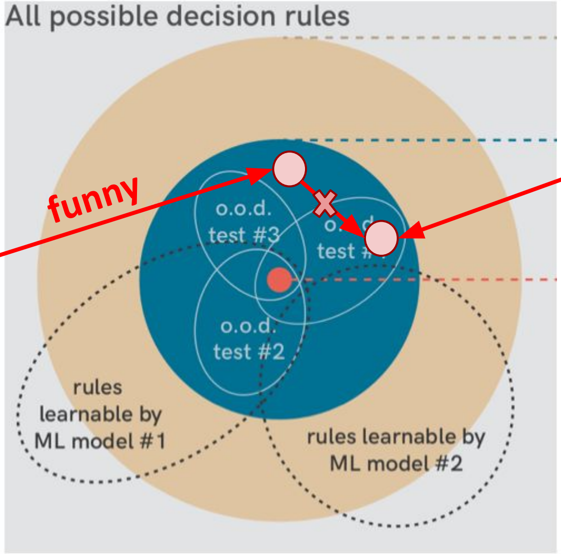
Shortcut Learning in Deep Neural Networks., Geirhos et al., 2020

Bias in NLP Model

Shortcut Learning

Rich veins of **funny** stuff in this movie! (Positive)
Is pretty **funny**. (Positive)
Very **funny** film (Positive)

Movie Reviews
(Rotten Tomatoes)



$40 million of **funny** child movie (Negative)

Movie Reviews
(IMDB)

Shortcut Learning in Deep Neural Networks., Geirhos et al., 2020

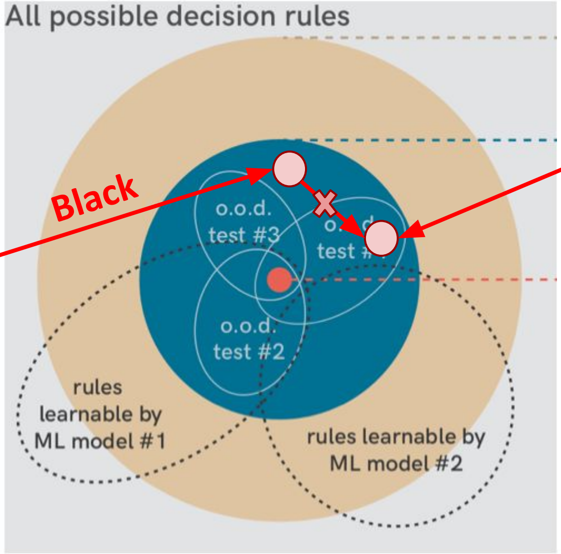
*Slide contains harmful contents

Bias in NLP Model

Shortcut Learning **-> False Positive -> Social Issue**

Whether a scientist of janitor, **black** people are all obedient brutes (Hate)
Blacks have been programmed to love watermelon (Hate)
Black people from the middle ages have always been watermelon-eating (Hate)

Hate Speech Detection Dataset



In the past the only way to get a job for a **black** person was to be a slave which was not fair for the **black** people (Not Hate)

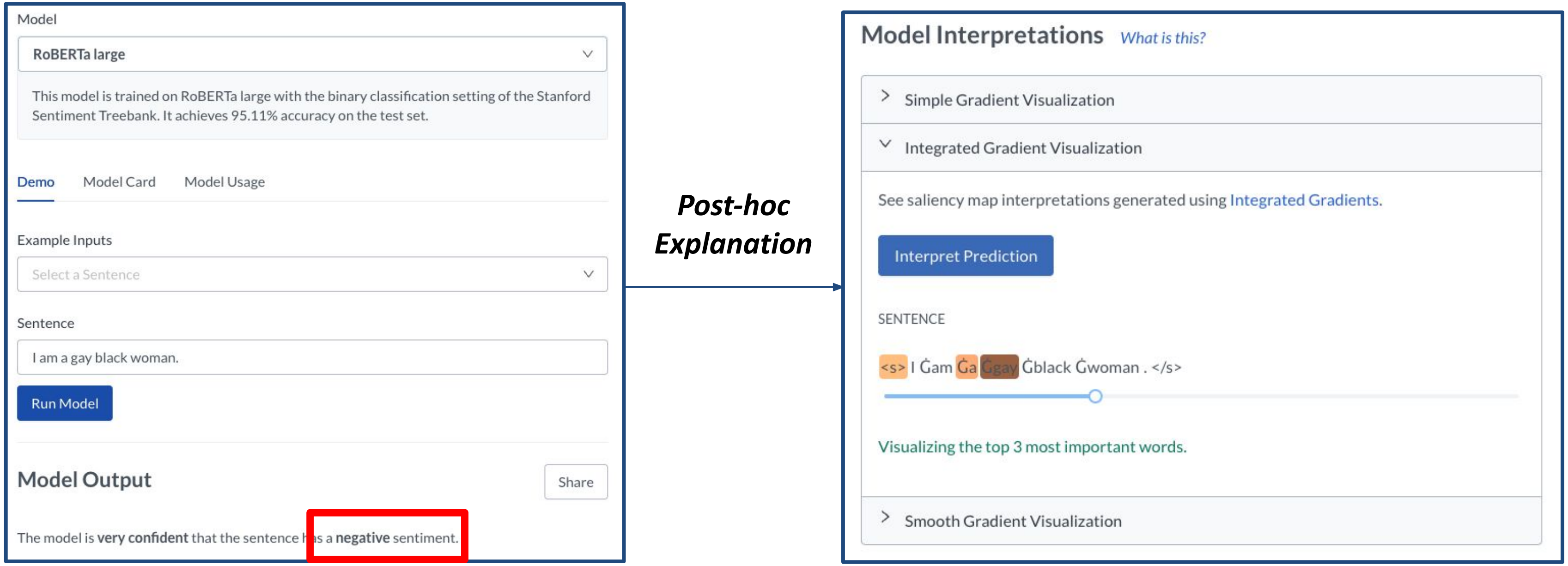
Real-world Case

Shortcut Learning in Deep Neural Networks., Geirhos et al., 2020

*Slide contains harmful contents

Visualize “shortcut” of the current model

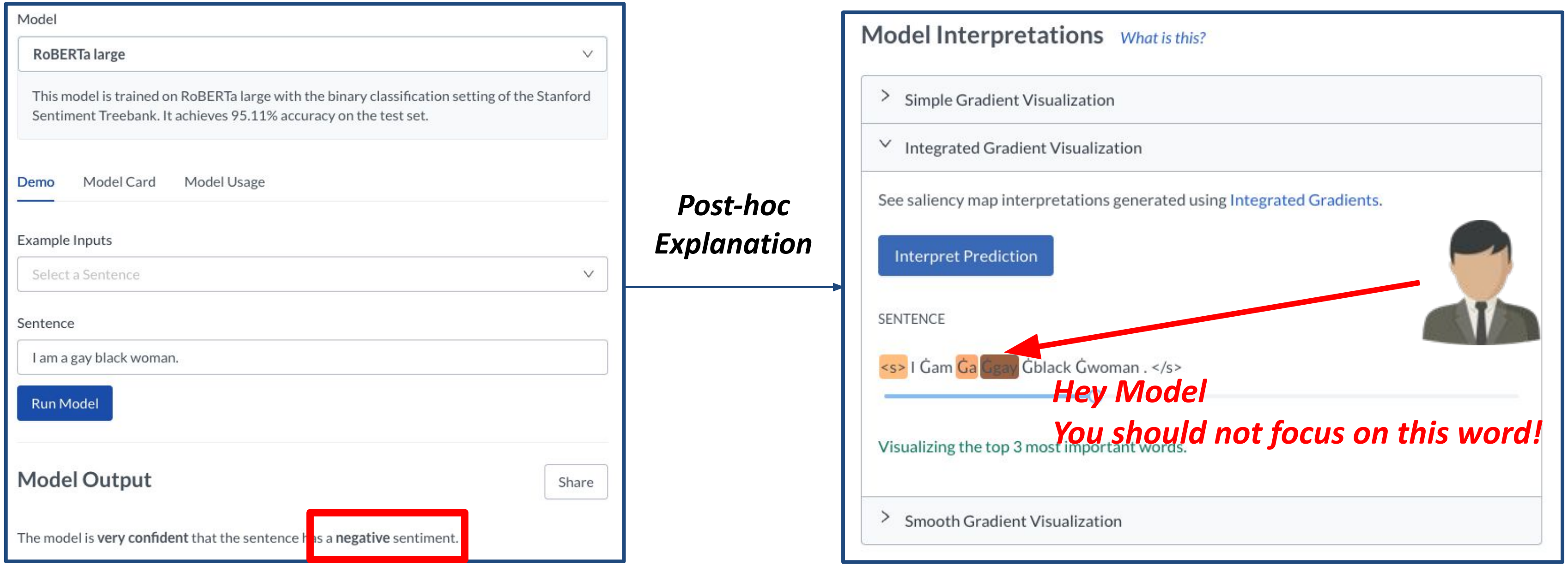
Post-hoc Model Explanation



https://demo.allennlp.org/sentiment-analysis/roberta-sentiment-analysis

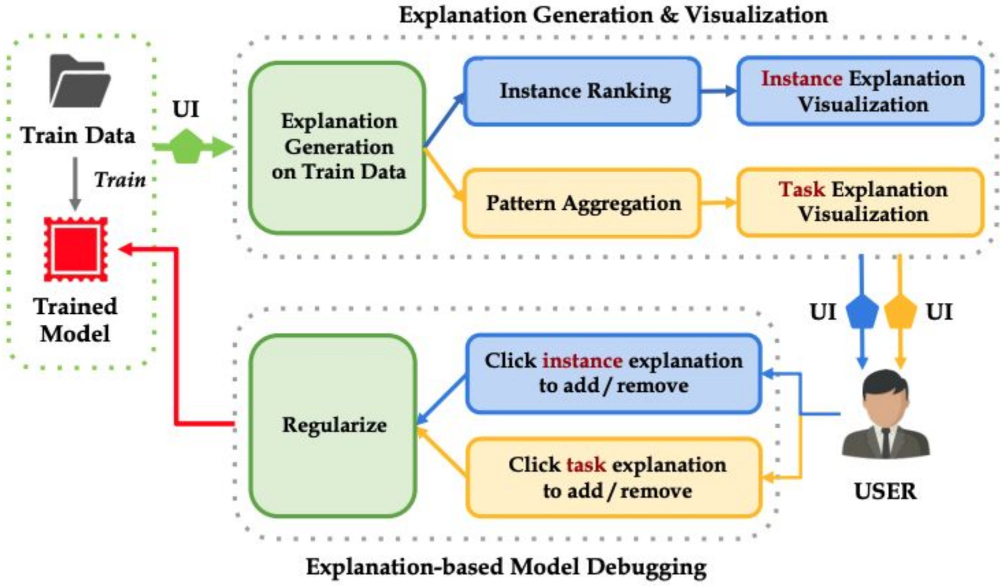
*Slide contains harmful contents

IDEA: Human feedback on Model Explanation



XMD

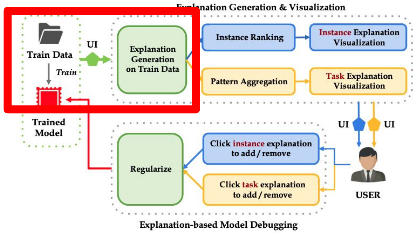
An End-to-End Framework for Interactive Explanation-based Debugging of NLP Models



https://inklab.usc.edu/xmd/

Explanation Generation

Category of Explanation



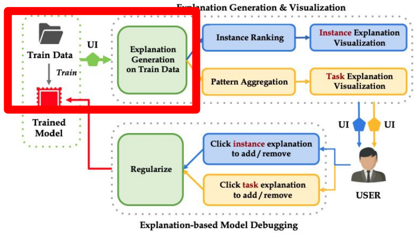
Orthogonal Aspects

Is the explanation for an
- Individual instance?
- AI model?

Is the explanation obtained
- Directly from the prediction?
- Requiring post-processing?

Explanation Generation

Category of Explanation



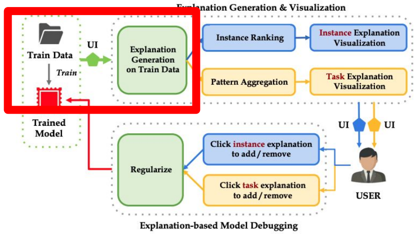
Orthogonal Aspects

Is the explanation for an
- Individual instance?
  - Local Explanation
- AI model?
  - Global Explanation

Is the explanation obtained
- Directly from the prediction?
  - Self-Explanation
- Requiring post-processing?
  - Post-hoc Explanation

Explanation Generation

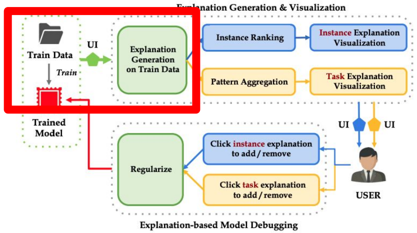
Category of Explanation



Orthogonal Aspects

Is the explanation for an
- Individual instance?
  - **Local Explanation**
- AI model?
  - Global Explanation

Is the explanation obtained
- Directly from the prediction?
  - Self-Explanation
- Requiring post-processing?
  - **Post-hoc Explanation**



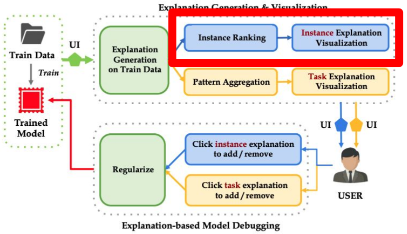
Explanation Generation

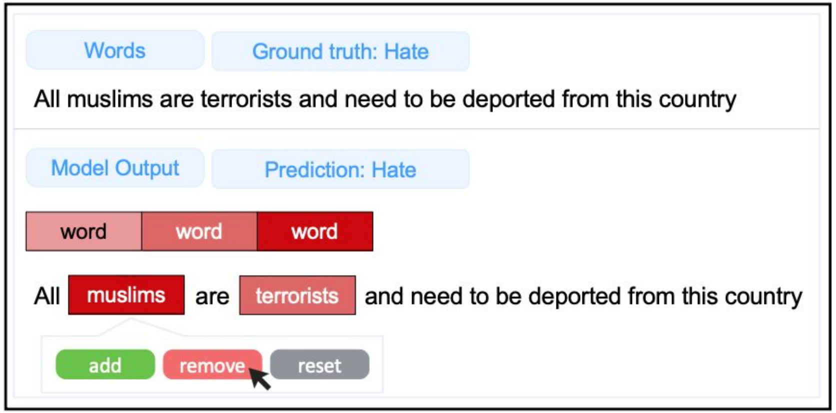
Local Post-hoc Explanation (Integrated Gradients)

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

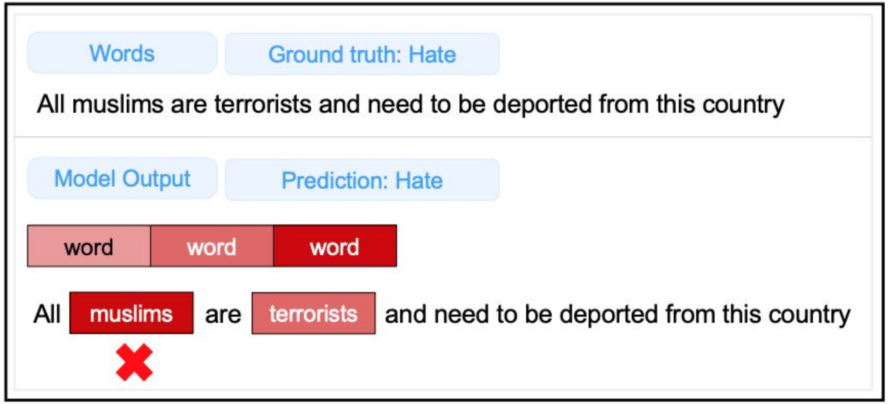
Axiomatic Attribution for Deep Networks., Sundarajan et al., 2017

Instance-level Explanation



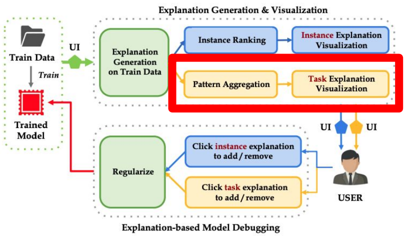


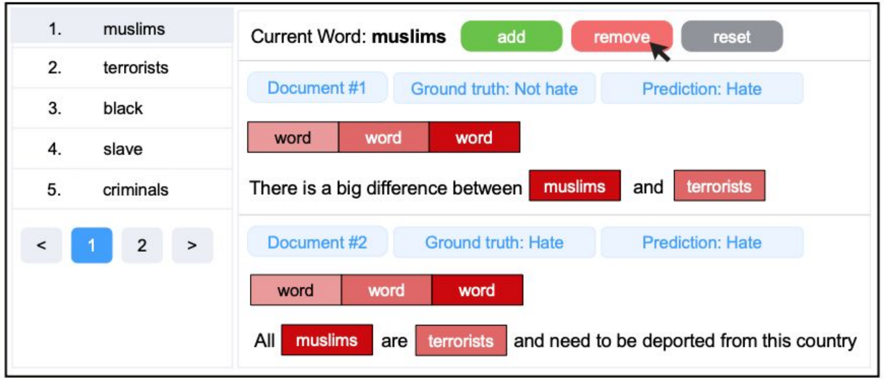
(a) As a user clicks on a word in the sentence, pop-up displaying operation options and a user selects an appropriate operation for that word.



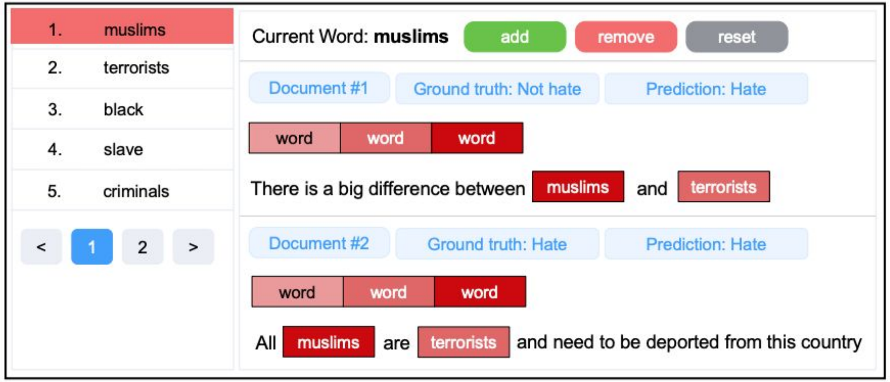
(b) Once the user selects an operation for the selected word, that word in the model output section is marked with an operation symbol (remove: X, add: +).

Task-level Explanation



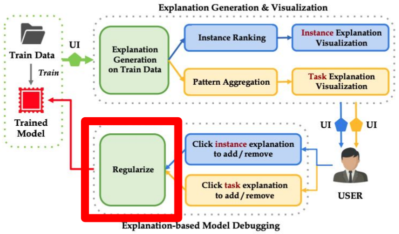


(a) As a user clicks on a word in the list of global explanations in the left panel, examples containing that word are displayed. The user can select the appropriate operation for the word.



(b) After the operation for a word is selected, the word in the left panel is marked with a color of the operation.

Explanation Regularization



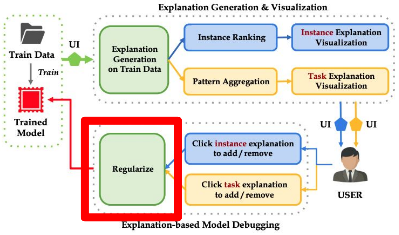
Task: SST-2 / Label Space: [Positive, Negative]

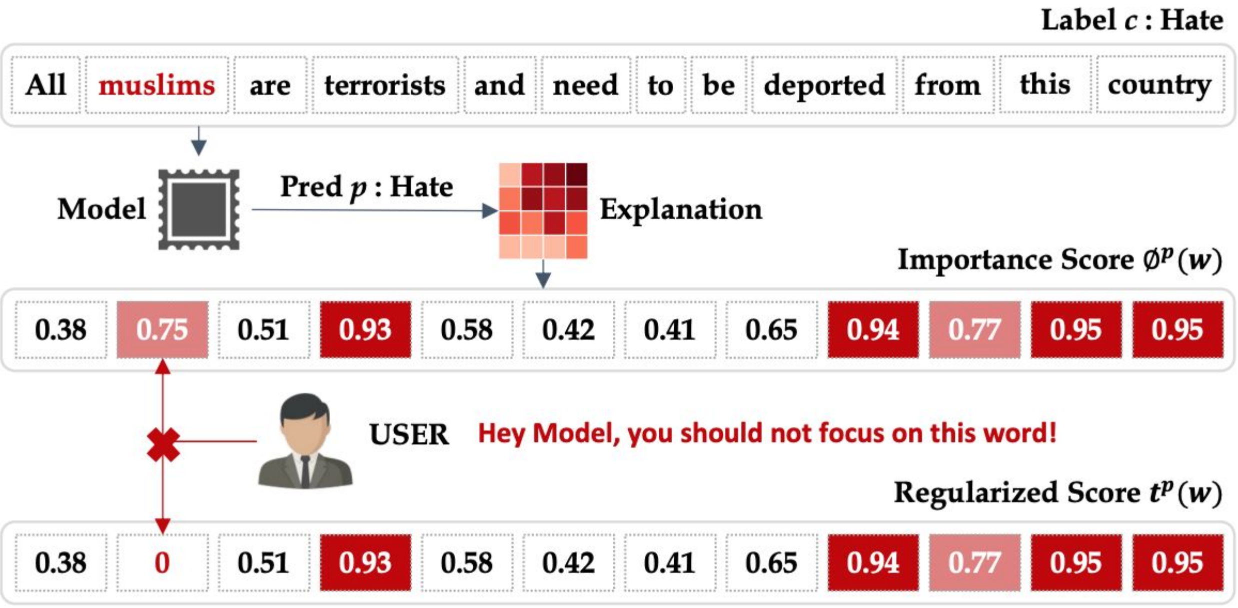
Step								**Pred c**
	Train instance	I	am	a	gay	black	man	Negative
1. Train Model & 2. Run Post-hoc Explanation	Attribution score ∅^c (p) toward "**Prediction**"	0.1	0.05	0.05	0.4	0.3	0.1	
3. Get human feedback	Human selection				del	del		
4. Compute ER term & 5. Re-train Model	Regularized attribution score t_p^c	0.1	0.05	0.05	**0**	**0**	0.1	

Explanation Regularization (ER) Term =
$$L_{ER} = \sum_{p \in x} (\phi^c(p) - t_p^c)^2$$

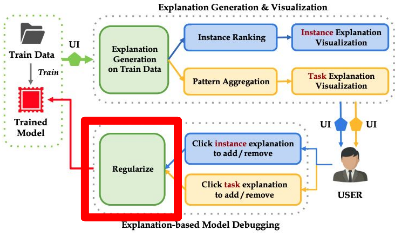
Re-train the model with new loss term =
$$L = L + L_{ER}$$

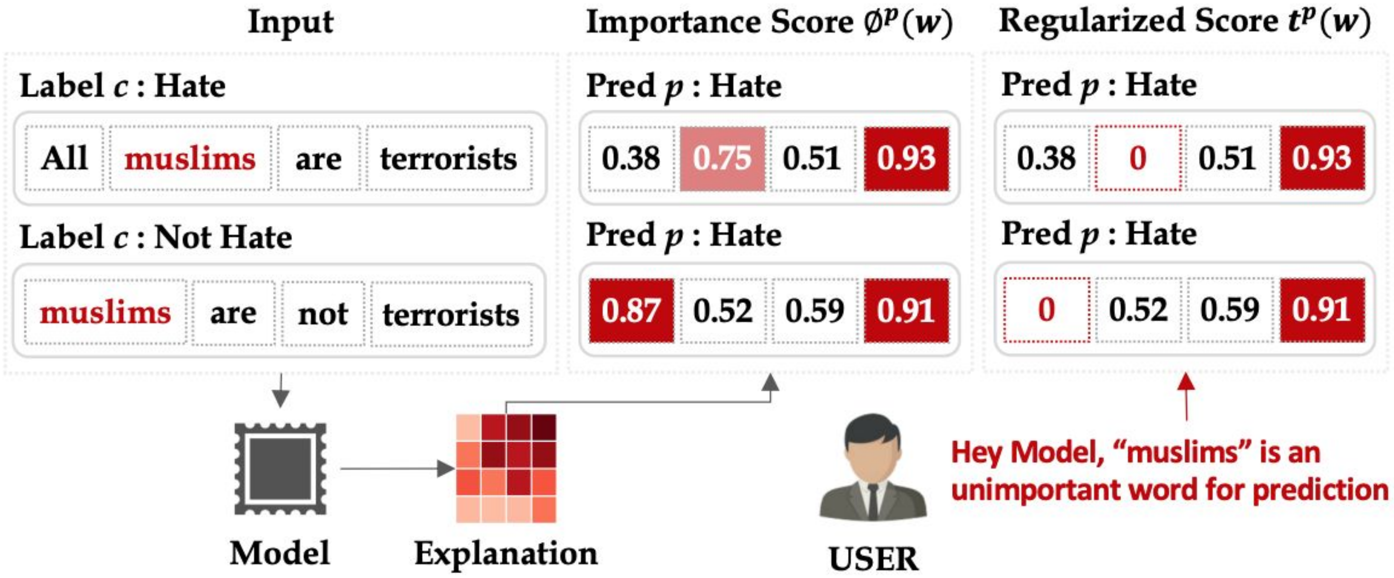
Instance-level Explanation Regularization





Task-level Explanation Regularization





Experimental Results

Regularize	ER Loss	Sentiment Analysis			
		In-distribution	Out-of-Distribution		
		SST	Amazon	Yelp	Movies
None	None	93.4	<u>89.1</u>	89.0	82.0
Correct	MSE	**94.7**	88.4	<u>91.8</u>	**94.5**
	MAE	<u>94.0</u>	**92.3**	**94.4**	<u>94.0</u>

Table 1: **Instance Explanation** ID/OOD Performance (Accuracy). Best models are bold and second best ones are underlined within each metric.

Regularize	ER Loss	Hate Speech Analysis			
		In-distribution	Out-of-Distribution		
		STF	HatEval	GHC	Latent
None	None	89.5	88.2	64.5	67.2
Correct	MSE	89.2	**90.1**	62.3	67.9
	MAE	89.1	**90.1**	59.3	64.9
Incorrect	MSE	88.9	86.3	**67.9**	**70.3**
	MAE	89.3	<u>88.8</u>	64.2	67.6
ALL	MSE	**90.0**	88.4	63.8	67.0
	MAE	<u>89.7</u>	86.9	<u>66.5</u>	<u>70.2</u>

Table 2: **Task Explanation** ID/OOD Performance (Accuracy). Best models are bold and second best ones are underlined within each metric.

Generalize well to Out-of-Distribution data

Q & A