Human-grounded Evaluations of Explanation Methods for Text Classification

Piyawat Lertvittayakumjorn[1] and Francesca Toni

Department of Computing, Imperial College London, UK

Email [1] : pl1515@imperial.ac.uk

2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)

What are explanation methods?

Global → how f works?
Local → why f(x)?

- **Local explanation method** explains an individual prediction

  E.g., to explain why this text is classified as a negative review

  The handles ***didn't fit*** *comfortably in my hand and the silicon tips are* ***hard,*** *not rubbery texture like* ***I'd imagined.***

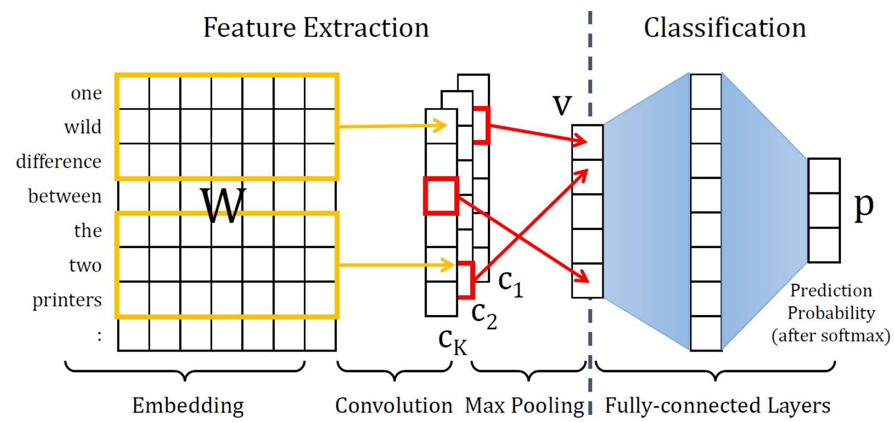
- Explanations for and against the predicted class are called **evidence and counter-evidence**, respectively.

- Previous works proposed several explanation methods which were mostly evaluated using proxy tasks without human involved.

- **In this paper, we propose three human evaluation tasks to evaluate local explanation methods for text classification**

Experimental setup: Datasets and Models

- Two English textual datasets for the three tasks.
  - (1) **Amazon** Review Polarity – (Positive and Negative) (Zhang et al., 2015)
  - (2) **ArXiv** Abstract – “Computer Science”, “Mathematics”, and “Physics”
- Classification models: 1D Convolutional Neural Networks
  - 200-dim GloVe vectors (non-trainable)
  - Three filter sizes [2, 3, 4] x 50 filters
- Performance (Macro-F1)
  - Amazon: 0.90
  - ArXiv: 0.94



Experimental setup: Local explanation methods

Method Name	Approach	Granularity	Models
Random (W)	Random Baselines	Words	Model-agnostic
Random (N)		N-grams	
LIME (Ribeiro et al., 2016)	Perturbation	Words	
LRP (W) (Bach et al., 2015)	Relevance Propagation	Words	Neural Networks
LRP (N)		N-grams	
DeepLIFT (W) (Shrikumar et al., 2017)		Words	
DeepLIFT (N)		N-grams	
Grad-CAM-Text	Gradient	N-grams	1D CNNs
Decision Trees (DTs)	Model Extraction	N-grams	

Newly proposed

Task 1: Revealing the Model Behavior

- Assumption:
  - Good explanation methods should enable humans to notice poor or peculiar behaviours of the model

- Setup:
  - Train two models to make them have different performance on classifying testing examples
  - Use these models to classify an input text
  - Apply the explanation method of interest to explain the predictions
  - Ask humans, based on the explanations from the two models, **which model is more reasonable?**

Experimental setup: Classification Models

- To train worse models for task 1
  - **Amazon**: train using only one epoch → underfitting
  - **ArXiv**: train using more specific topics
    - ‘Computer Science’ → ‘Computation and Language’
    - ‘Mathematics’ → ‘Dynamical Systems’
    - ‘Quantum Physics’ → ‘Physics’

Dataset / Macro-F1	Well-trained	Worse
Amazon	0.90	0.81
ArXiv	0.94	0.85

Piyawat Lertvittayakumjorn and Francesca Toni

Question 9 out of 10: Both Robot C and Robot L classify that the following review has a "**Negative**" sentiment.

Robot C:
Not for the product but to amazon : I order two of those bottles I receiv
waste of money in something that will expire next month ...

Robot L:
Not for the product but to amazon : I order **two of those bottles I received** it today Oct.4 2011 the **expiration date for** those bottles Nov.2011 , waste of money in something that will expire next month ...

Your answer:

- Robot C seems clearly more reasonable than Robot L.
- Robot C seems slightly more reasonable than Robot L.
- I can't say which robot is more reasonable.
- Robot L seems slightly more reasonable than Robot C.
- Robot L seems clearly more reasonable than Robot C.

The chosen input texts must be classified into the same class by both models

We consider both the explanations for correct and incorrect predictions.

(-)1.0	(In)correct, confident
(-)0.5	(In)correct, unconfident
0.0	No preference

Experimental setup: Participants

- **Amazon**: we posted our tasks on Amazon Mechanical Turk (MTurk). Each question is answered by three MTurk workers.
- **ArXiv**: we recruited graduates and post-graduate students in Computer Science, Mathematics, Physics, and Engineering to perform the tasks. Each question is answered by one participant.

Results: Task 1

Explanation Method	Task 1					
	Amazon			ArXiv		
	$\mathcal{A}$	✔	✘	$\mathcal{A}$	✔	✘
Random (W)	.02	.00	.04	-.11	-.05	-.17
Random (N)	.02	.02	.02	-.12	-.16	-.07
LIME (W)	-.02	.02	-.06	.03	.02	**.03**
LRP (W)	.00	-.01	.02	-.03	-.01	-.05
LRP (N)	-.07	-.04	-.09	**.12**	**.24**	-.01
DeepLIFT (W)	.04	.03	.04	.07	.13	.00
DeepLIFT (N)	.06	.06	**.05**	.06	.22	-.10
Grad-CAM-T (N)	**.07**	**.11**	.03	-.03	-.04	-.01
DTs (N)	-.05	-.02	-.08	-.13	-.22	-.03
Fleiss κ (Amazon)	0.050 / 0.054			N/A		

- Amazon: None of the methods can apparently reveal the underfitting CNN
- ArXiv: LRP (N) and DeepLIFT (N) fairly work when both CNNs predicted correctly.
- For two explanations with comparable semantic quality, humans prefer the one with **more evidence texts**.

Task 2: Justifying the Predictions

- Assumption: the evidence texts are truly related to the predicted class and can distinguish it from the other classes, so called class-discriminative
- Setup:
  - Use a well-trained model
  - Select an input example classified by this model with high confidence
    - $\max_c p_c > \tau_h$ where τ_h is a threshold parameter ($\tau_h = 0.9$)
  - Show the top-m evidence text fragments (m = 3) generated by the explanation method of interest and ask humans to guess the class of the document containing the evidence.

Question 6 out of 10.

- enjoyed reading this
- Review : I enjoyed
- the story is fully

Your answer:

- I'm certain that they are from a **Positive** review.
- I'm certain that they are from a **Negative** review.
- I'm not certain but they are likely from a **Positive** review.
- I'm not certain but they are likely from a **Negative** review.
- I can't say.

We consider both the explanations for correct and incorrect predictions.

(-)1.0	(In)correct, confident
(-)0.5	(In)correct, unconfident
0.0	No preference

** Correct = select the class predicted by the model

Results: Task 2

Explanation Method	Task 2					
	Amazon			ArXiv		
	$\mathcal{A}$	✔	✘	$\mathcal{A}$	✔	✘
Random (W)	.06	.10	.02	.07	.09	.04
Random (N)	.12	.13	.12	.29	.32	.25
LIME (W)	**.69**	**.74**	.64	**.70**	**.75**	**.64**
LRP (W)	.13	.26	-.01	.26	.36	.16
LRP (N)	.26	.45	.08	.44	.49	.39
DeepLIFT (W)	.21	.37	.04	.26	.35	.16
DeepLIFT (N)	.23	.47	-.01	.38	.47	.28
Grad-CAM-T (N)	.65	.64	**.66**	.53	.65	.41
DTs (N)	.64	.68	.59	.51	.69	.32
Fleiss κ (Amazon)	0.274 / 0.371			N/A		

- We can use evidence given by LIME, Grad-CAM-Text, and DTs to justify the predictions (regardless of the correctness of the predictions).
- LRP and DeepLIFT
  - Provide good reasons for only correct predictions
  - N-gram version outperforms the word version

Piyawat Lertvittayakumjorn and Francesca Toni

Task 3: Investigating Uncertain Predictions

- Assumption: Good explanations can help human investigate and understand uncertain predictions (predicted by a model with low confidence)

- Set up:
  - Use a well-trained model
  - Select an input example classified by this model with low confidence
    - $\max_c p_c < \tau_l$ where τ_l is a threshold parameter ($\tau_l = 0.7$)
  - Show top-m evidence and top-m counter-evidence texts of the predicted class (m = 3) as well as the predicted class and probability.
  - Ask humans to use all the information to guess the actual class of the input text, without seeing the input text itself

Question 3 out of 10.

- Predicted class: **Positive**

- Evidence for the **Positive** sentiment:
  - Good Sound . :
  - is good . Has
  - . I was happier
- Evidence for the **Negative** sentiment:
  - would not re -
  - cheap foam covers on
  - . I was happier

Your answer:

I'm certain that it is a **Positive** review.
I'm certain that it is a **Negative** review.
I'm not certain but it is probably a **Positive** review.
I'm not certain but it is probably a **Negative** review.

We consider both the explanations for correct and incorrect predictions.

(-)1.0	(In)correct, confident
(-)0.5	(In)correct, unconfident

** Do not provide the "no preference" option as the humans can still rely on the predicted scores when all the explanations are unhelpful

Piyawat Lertvittayakumjorn and Francesca Toni

Results: Task 3

Explanation Method	Task 3					
	Amazon			ArXiv		
	$\mathcal{A}$	✔	✘	$\mathcal{A}$	✔	✘
Random (W)	.05	.53	-.43	.01	.32	-.30
Random (N)	-.01	.54	-.55	.02	.29	**-.25**
LIME (W)	.02	.50	-.45	-.02	.31	-.34
LRP (W)	-.02	.50	-.54	-.06	.33	-.44
LRP (N)	.08	.60	-.43	**.17**	**.60**	-.26
DeepLIFT (W)	-.03	.47	-.53	-.08	.28	-.44
DeepLIFT (N)	.05	.59	-.49	.02	.33	-.30
Grad-CAM-T (N)	.05	.51	-.42	.06	.56	-.45
DTs (N)	**.10**	**.60**	**-.40**	-.11	.29	-.50
Fleiss κ (Amazon)	0.212 / 0.499			N/A		

- DTs performed well only on the Amazon dataset, but not the ArXiv dataset.
- Why did LRP (N) work?
  - good evidence for correct predictions
  - unconvincing evidence for incorrect predictions
- Why didn't LIME work well?
  - good evidence for both the correct and incorrect classes
  - humans become indecisive.

Conclusion

- We proposed three human tasks to evaluate local explanation methods for text classification. We experimented on 1D CNNs and found that
  - LIME is the most class discriminative method, justifying predictions with relevant evidence
  - LRP (N) works fairly well in helping humans investigate uncertain predictions
  - using explanations to reveal model behavior is challenging, and none of the methods achieved impressive results
  - whenever using LRP and DeepLIFT, we should present to humans the most relevant words together with their contexts

- Future work: evaluating on other datasets and other advanced architectures https://github.com/plkumjorn/CNNAnalysis

@plkumjorn @fra_toni

Thank you

Q&A

Piyawat Lertvittayakumjorn[1] and Francesca Toni

Department of Computing, Imperial College London, UK

Email [1] : pl1515@imperial.ac.uk