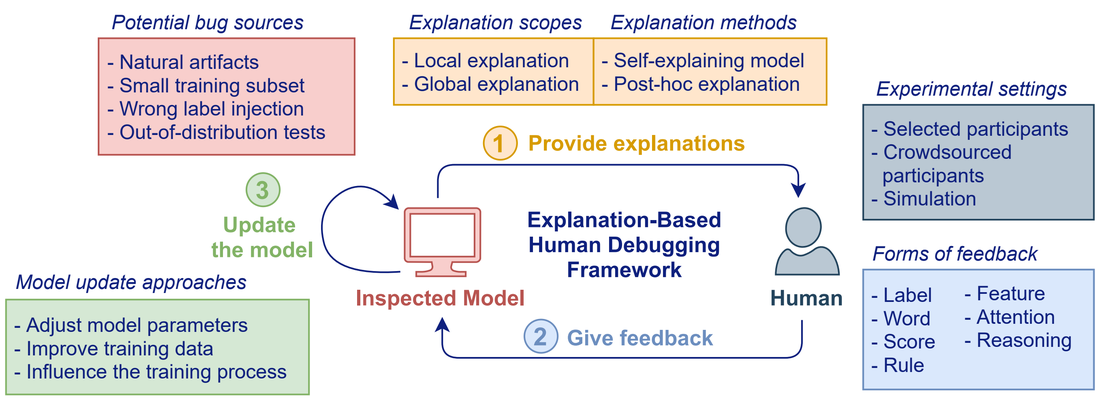
Imperial College London

EXPLANATION-BASED HUMAN DEBUGGING OF NLP MODELS: A SURVEY

Piyawat Lertvittayakumjorn and Francesca Toni – {pl1515, f.toni}@imperial.ac.uk
Department of Computing, Imperial College London, United Kingdom

ABSTRACT

Debugging a machine learning model is hard since the bug usually involves the training data and the learning process. This becomes even harder for an opaque deep learning model if we have no clue about how the model actually works. In this survey, we review papers that exploit explanations to enable humans to give feedback and debug NLP models. We call this problem **explanation-based human debugging (EBHD)**. In particular, we categorize and discuss existing work along three dimensions of EBHD (the bug context, the workflow, and the experimental setting), compile findings on how EBHD components affect the feedback providers, and highlight open problems that could be future research directions.



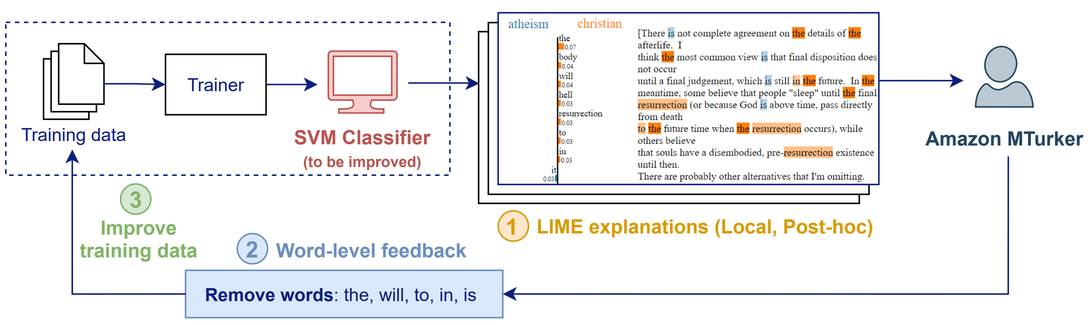
A general framework for EBHD of NLP models

BUG AND DEBUGGING IN ML

- **Bug**: Contamination in the learning and/or prediction pipeline that makes the model produce incorrect predictions or learn error-causing associations, e.g., spurious correlation, labelling errors, and undesirable behavior in OOD testing [1].
- **Debugging**: Identifying the bugs + fixing or mitigating them.
- **Explanation-based human debugging (EBHD)**: The process of fixing or mitigating bugs in a trained model using human feedback given in response to explanations for the model.

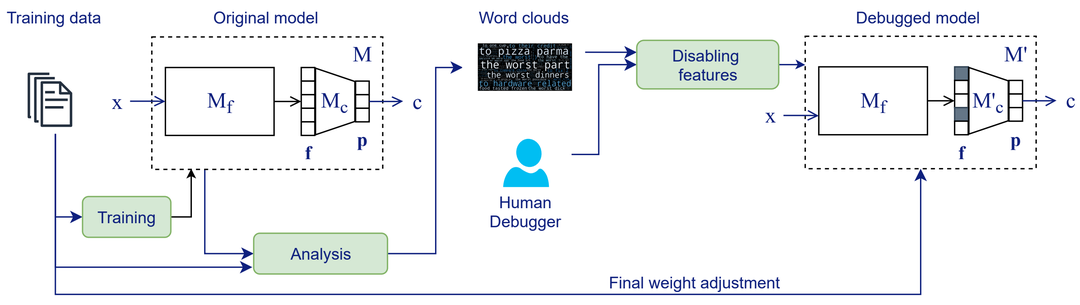
EXAMPLE 1: LIME

- **LIME**: Local Interpretable Model-agnostic Explanations [2]
- **Context**: Text classification; SVM model, trained on *20Newsgroups* (Atheism vs Christianity), tested on *Religion*.



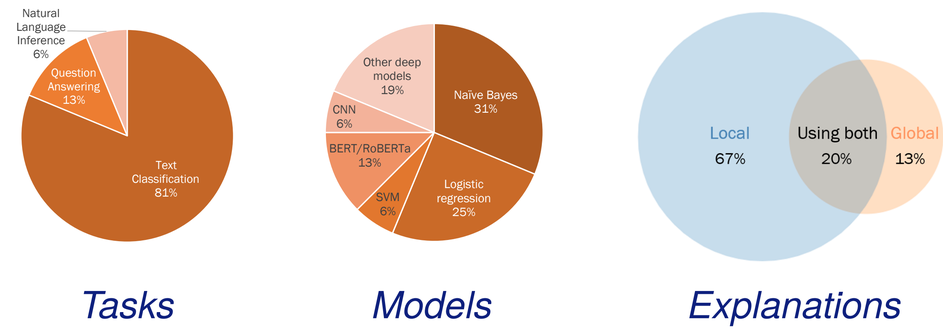
EXAMPLE 2: FIND

- **FIND**: Feature Investigation and Disabling [3]
- **Context**: Text classification (several tasks); 1D CNNs.



STEP 1: PROVIDING EXPLANATIONS

- **Global explanations**: to reveal significant bugs
- **Local explanations**: to reveal fine-grained bugs; Need a strategy to pick examples to explain (e.g., incorrect predictions, non-redundancy, informativeness criteria).



Tasks *Models* *Explanations*

STEP 2: COLLECTING FEEDBACK

Explanation Form	Feedback Method
Rationales, Relevance scores, Hierarchical heat maps	Identify (ir)relevant tokens, Adjust token relevance scores
Influential training examples	Provide correct labels, Provide relevancy scores
Learned features Adversarial rules	Identify (ir)relevant features Check semantic equivalence

STEP 3: UPDATING THE MODEL

- **Directly adjust the model parameters**: Suitable for transparent models; Fast, No retraining required; How can we ensure that human adjustments are indeed good?
- **Improve the training data**: E.g., correcting mislabeled training examples, removing irrelevant words from input texts, adding more training examples to reduce the effects of the artifacts.
- **Influence the (re)training process**: Aiming to make the resulting model behave as the feedback suggests, e.g., disabling features, regularizing the explanations, constraint optimization.

HUMAN FACTORS

- Model understanding, Human feedback characteristics
- Human trust, frustration, expectation

OPEN PROBLEMS

- Beyond English text classification
- Tackling more challenging bugs – dealing with conflicting pieces of feedback, injecting new knowledge to the model
- Analyzing and enhancing efficiency
- Reliable comparison across papers & Towards deployment

REFERENCES

[1] Julius Adebayo, Michael Muelly, Ilaria Liccardi, Been Kim. 2020. *Debugging Tests for Model Explanations*. (NeurIPS'20)

[2] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. 2016. *"Why should i trust you?": Explaining the predictions of any classifier*. (KDD'16)

[3] Piyawat Lertvittayakumjorn, Lucia Specia, Francesca Toni. 2020. *FIND: Human-in-the-Loop Debugging Deep Text Classifiers*. (EMNLP'20)