Latent Concept-based Explanation of NLP Models

Xuemin Yu[1] Fahim Dalvi[2] Nadir Durrani[2]
Marzia Nouri[3] Hassan Sajjad[1]

[1]HyperMatrix, Faculty of Computer Science, Dalhousie University, Canada

[2]Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar

[3]Independent Researcher

Scan ME!

Motivation

- **Problem:** Interpreting and understanding the predictions made by deep learning models poses a formidable challenge due to their inherently opaque nature. Many existing explanations rely heavily on input features, which are less informative due to the discrete nature of words and lack of contextual verbosity.
- **Our Approach:** We propose the **Latent Concept Attribution (LACOAT) method**, which generates explanations using latent concepts that represent the multifaceted roles of words in context.

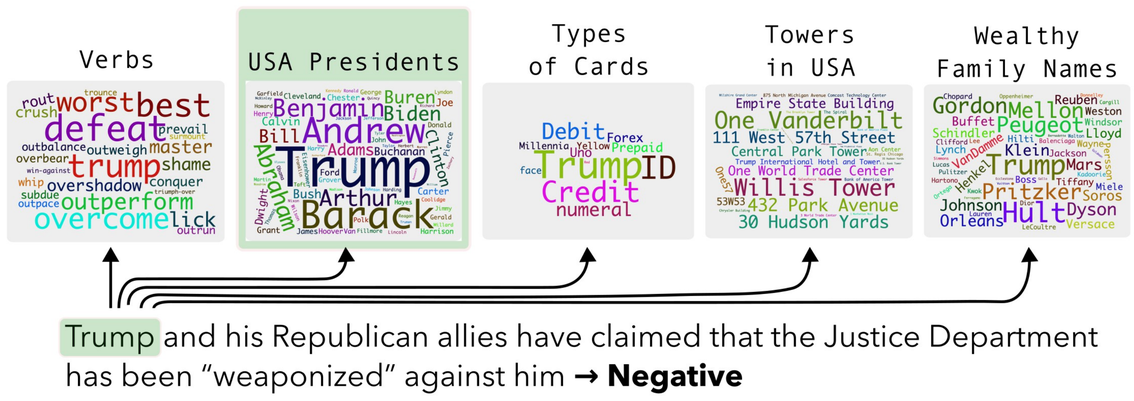


Figure 1. An example of various facets of the word "trump"

Methodology

Key Idea: We hypothesize that, given various contexts of a word in the training data, the model learns these diverse facets during training. For a test instance, depending on the context in which a word appears, the model uses a particular facet of the input words in making the prediction.

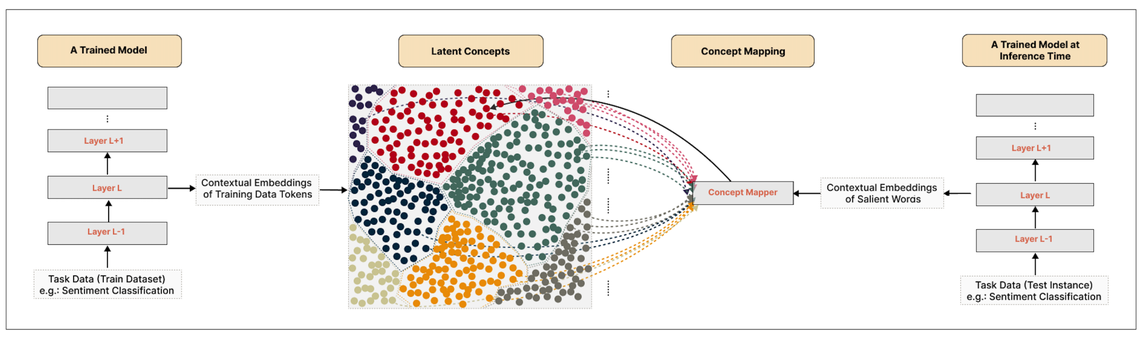


Figure 2. Overview of LACOAT method

LACOAT comprises four modules:

- **ConceptDiscoverer:** discovers latent concepts of a model given a corpus.
- **PredictionAttributor:** selects the most salient words (along with their contextual representations) in a sentence with respect to the model's prediction.
- **ConceptMapper:** maps the representations of the salient words to the latent concepts discovered by **ConceptDiscoverer** and provides a latent concept-based explanation.
- **PlausiFyer:** takes a latent concept explanation as input and generates a plausible and human-understandable explanation of the prediction.

Evaluation Experimental Setup

- **Task:** Parts-of-Speech tagging, sentiment classification, toxicity classification, and natural language inference
- Models:
  - **Fine-tuned 12-layered pre-trained models**: BERT-base-cased, RoBERTa-base, and XLM-Roberta
  - **Based model without fine-tuning with zero-shot prompting**: Llama-2-7b-chat-hf
- Evaluations:
  - **Qualitative evaluation:** evaluates the usefulness of the latent concept-based explanation and the generated plausible explanation.
  - **Human Evaluation:** evaluates the usefulness of LACOAT's explanation in understanding a prediction and compares LACOAT with other explanation methods.
  - **Module Specific Evaluation:** evaluates the efficacy of the modules.
  - **Faithfulness Evaluation:** evaluates the faithfulness of the latent concept to the prediction.

Results

- **Qualitative Evaluation:** It shows that LACOAT explanations are insightful in explaining predictions and comparing explanations between models. Additionally, it demonstrates how the concepts evolve across layers of the model. We found that the latent concepts in the last layer are best aligned with the input instance and its prediction and are the most informative explanation of the prediction.

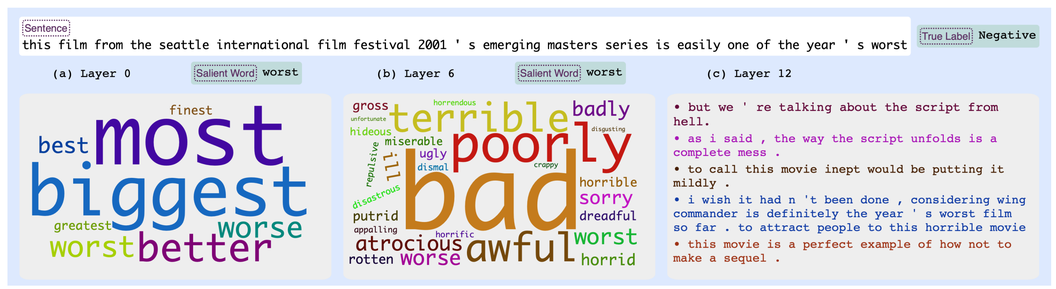
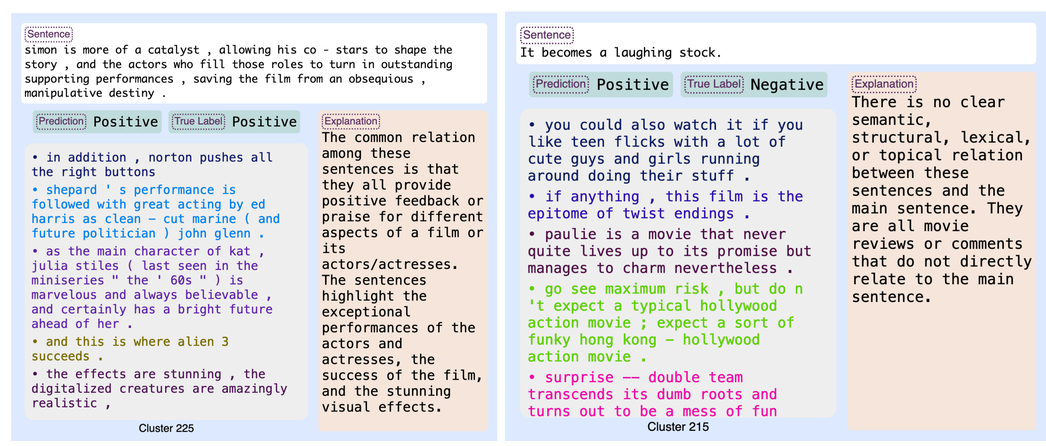


Figure 3. Sentiment task: Latent concepts of the most attributed words in Layers 0, 6, and 12



(a) A positive labeled test instance correctly predicted by the model.

(b) A negatively labeled test instance that is incorrectly predicted as positive.

Figure 4. A few examples of LACOAT explanations for BERT using Sentiment tasks

- Human Evaluation:
  - **LACOAT Effectiveness:** LACOAT's explanations were generally helpful when the model made correct predictions. In cases of incorrect predictions, while the explanations may not always be as informative, they often highlight potential issues in the model's reasoning.
  - **Comparison with other Methods:** The results showed that LACOAT explanation is more useful for understanding predictions compared to other methods.

Top	Labels	Correct Samples	Incorrect Samples	All Samples	
				Annotation	Fleiss κ
Q1	Yes/No	28 / 0	20 / 2	48 / 2	0.35
Q2	Helps/Neutral/Hinders	27 / 1 / 0	17 / 5 / 0	44 / 6 / 0	0.41
Q3	Helps/Neutral/Hinders	16 / 10 / 2	1 / 19 / 2	17 / 29 / 4	0.61
Q4	Yes/No	17 / 11	5 / 17	22 / 28	0.47
Q5	Yes/No	17 / 11	6 / 16	23 / 27	0.80

Bottom	A1	A2	A3	A4	Consolidated	Average Cohen's κ
LACOAT ↑	85%	72%	77%	87%	89%	0.37

Table 1. **Top**: Consolidated label distribution for Q1-Q5. Fleiss' κ scores are computed by averaging each annotator with the consolidated annotation. The consolidated labels and agreement scores are shown for all the samples, as well as partitioned into those where the model prediction was correct/incorrect. **Bottom**: Percentage of samples where LACOAT is ranked similar or better than other methods. A* represents the average preference of LACOAT per annotator.

- **Faithfulness Evaluation:** The results showed that manipulating the [CLS] token representation using the LACOAT vector led to significant performance drops and prediction changes across all datasets, confirming the faithfulness of the latent concepts.

		Faithfulness Metrics	
Dataset	Setting	Accuracy	% Label Flip
Sentiment	Original	96.31	-
	LACOAT	55.91	43.98
	Random	96.09	0.14
Toxicity	Original	91.55	-
	LACOAT	51.78	46.44
	Random	91.93	0.13
MNLI	Original	87.69	-
	LACOAT	82.08	8.83
	Random	88.12	0.55

Table 2. Faithfulness evaluation using the RoBERTa model. *Original* refers to the model's performance without any manipulation, *LACOAT* represents the performance after subtracting the most salient latent concept vector from the [CLS] vector, and Random is the average performance of the model after subtracting five random vectors from the [CLS] vector.