AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models

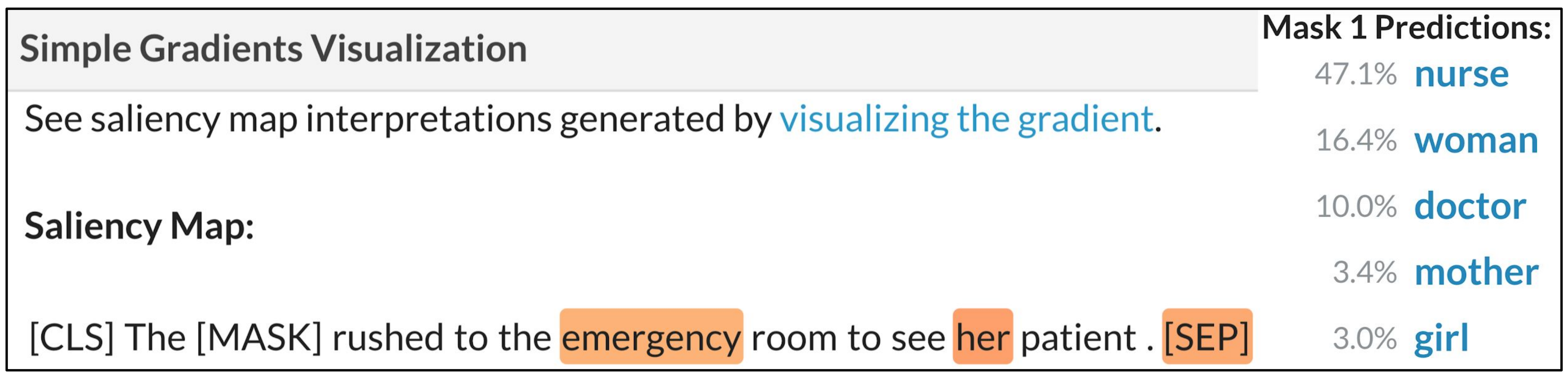
Eric Wallace, Jens Tuyls, Junlin Wang,
Sanjay Subramanian, Matt Gardner, Sameer Singh
AI2, UC Irvine

Goal: Explain *why* NLP models make certain predictions

We provide interactive interpretations for *any* AllenNLP model

- ✓ Saliency maps, adversarial attacks, input reduction, ...
- ✓ SoTA models (BERT, GPT-2, QANet, BiDAF, LSTM-CRF, ...)
- ✓ Complex task formats (QA, language modeling, NER)
- ✓ Easy to add your own model and interpretation

Example Uses



Simple Gradients Visualization

See saliency map interpretations generated by visualizing the gradient.

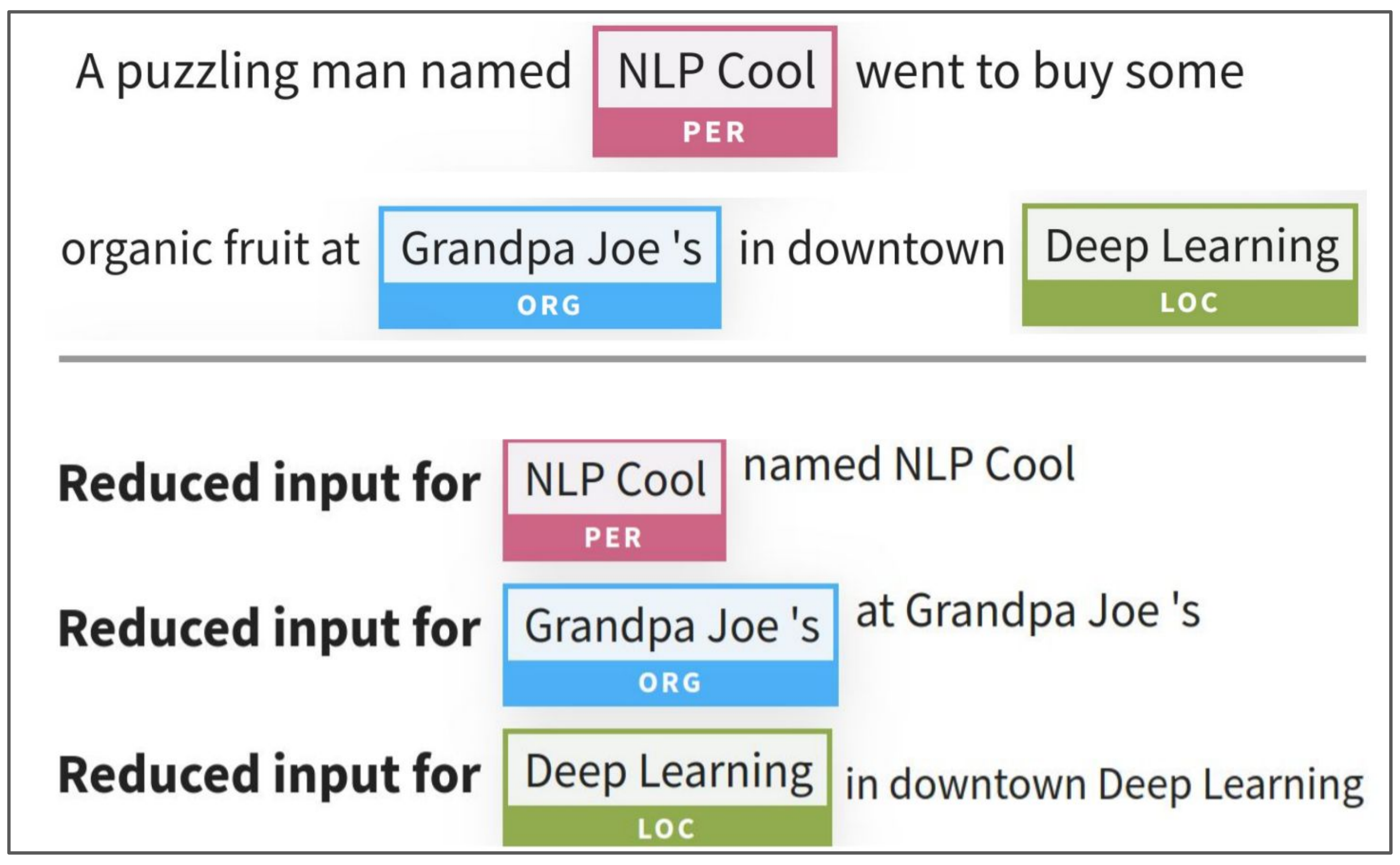
Saliency Map:

[CLS] The [MASK] rushed to the emergency room to see her patient . [SEP]

Mask 1 Predictions:

47.1%	nurse
16.4%	woman
10.0%	doctor
3.4%	mother
3.0%	girl

Saliency Map (BERT)



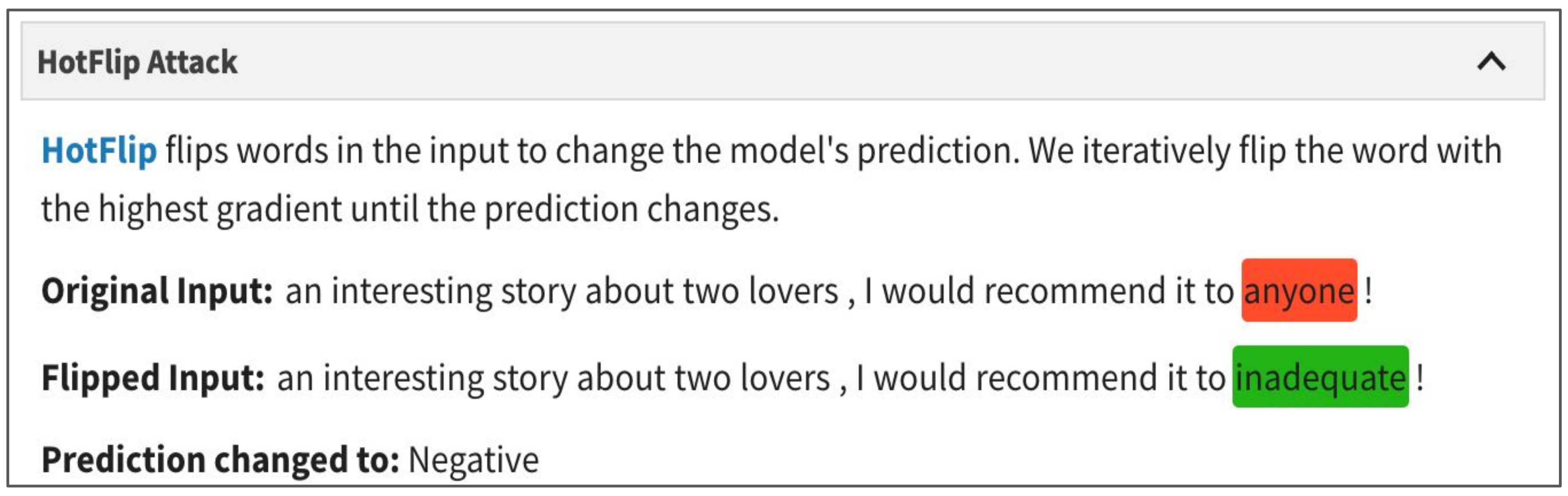
A puzzling man named NLP Cool PER went to buy some organic fruit at Grandpa Joe 's ORG in downtown Deep Learning LOC

Reduced input for NLP Cool PER named NLP Cool

Reduced input for Grandpa Joe 's ORG at Grandpa Joe 's

Reduced input for Deep Learning LOC in downtown Deep Learning

Input Reduction (NER)



HotFlip Attack

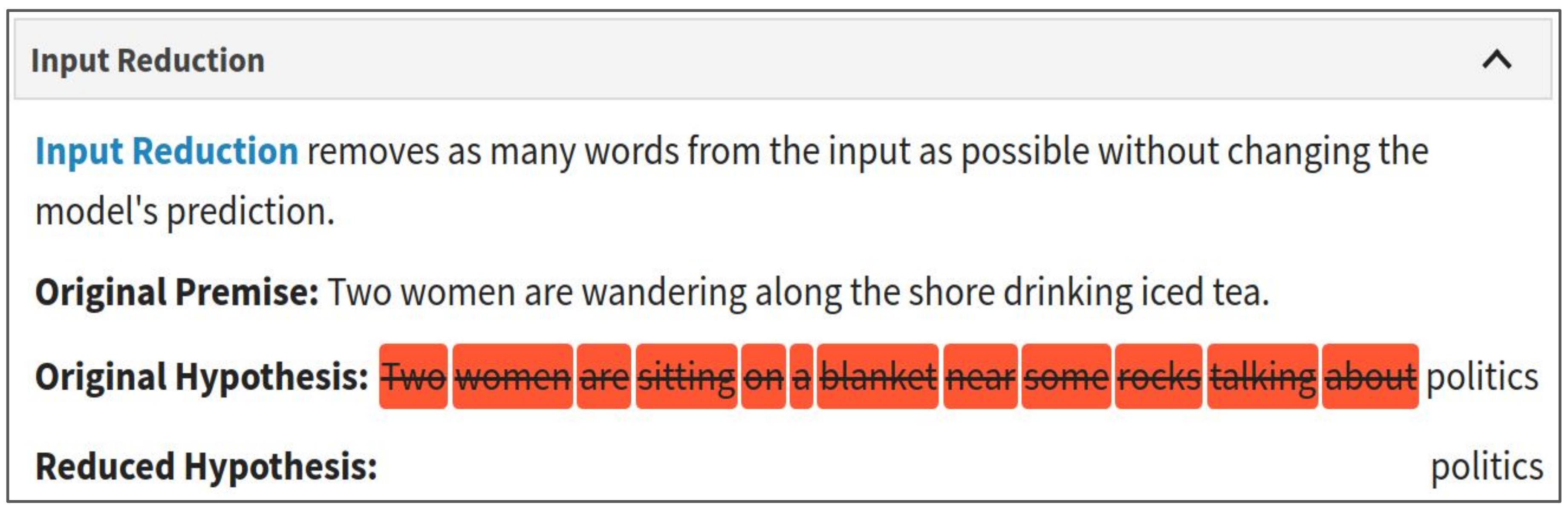
HotFlip flips words in the input to change the model's prediction. We iteratively flip the word with the highest gradient until the prediction changes.

Original Input: an interesting story about two lovers , I would recommend it to anyone !

Flipped Input: an interesting story about two lovers , I would recommend it to inadequate !

Prediction changed to: Negative

Adversarial Attack (Sentiment)



Input Reduction

Input Reduction removes as many words from the input as possible without changing the model's prediction.

Original Premise: Two women are wandering along the shore drinking iced tea.

Original Hypothesis: ~~Two~~ ~~women~~ ~~are~~ ~~sitting~~ ~~on~~ ~~a~~ ~~blanket~~ ~~near~~ ~~some~~ ~~rocks~~ ~~talking~~ ~~about~~ politics

Reduced Hypothesis: politics

Input Reduction (Entailment)

https://allennlp.org/interpret