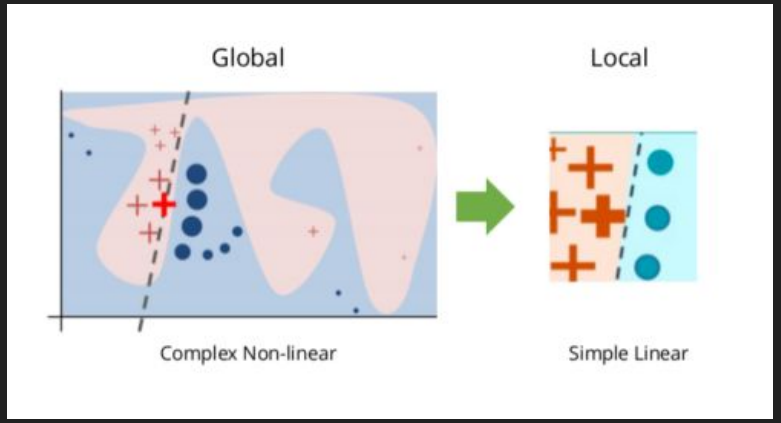
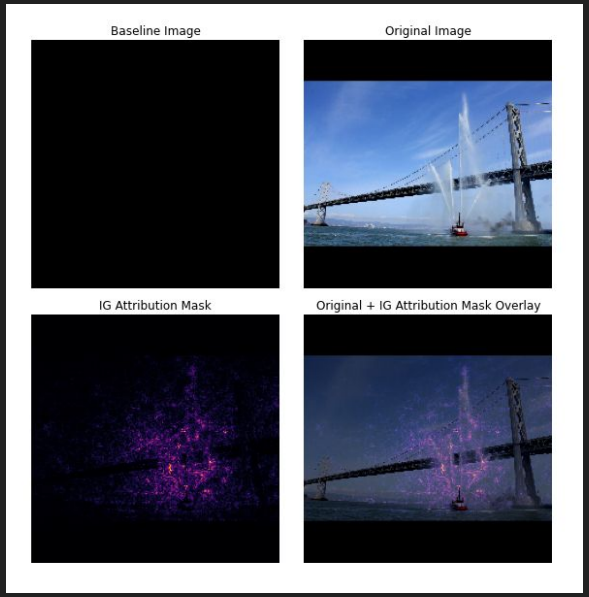
Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing, 2021 Sinha et al

XAI MIMUW 2022/2023
Michał Krutul

“Why should I trust you?” Problems with XAI methods in NLP

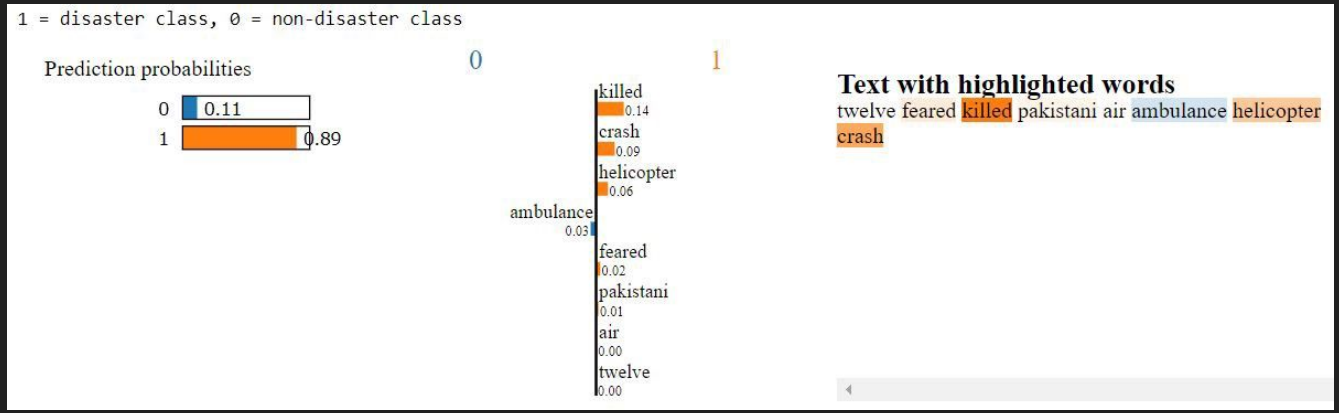


Lime

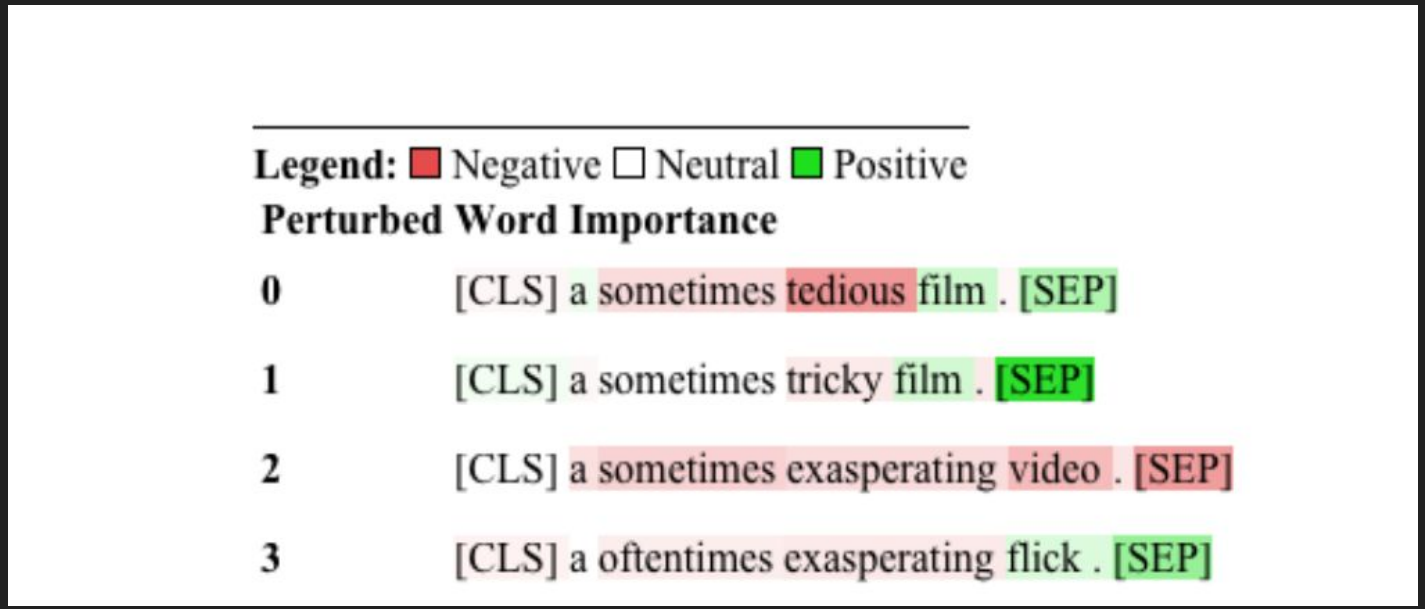


Integrated Gradients

Explanations in NLP



Words Perturbations



Metrics for different interpretations

$$LOM(I) = \frac{\sum_{t=0}^{t=n-1}(i_t * t)}{\sum_{t=0}^{t=n-1} i_t}$$

$$\Delta LOM(I_1, I_2) = |LOM(I_1) - LOM(I_2)|$$

Delta LOM

$$L2Norm(I_1, I_2) = \|I_1 - I_2\|_2$$

L2Norm

How to perturbate?

1. Order a words by Leave-one out approach

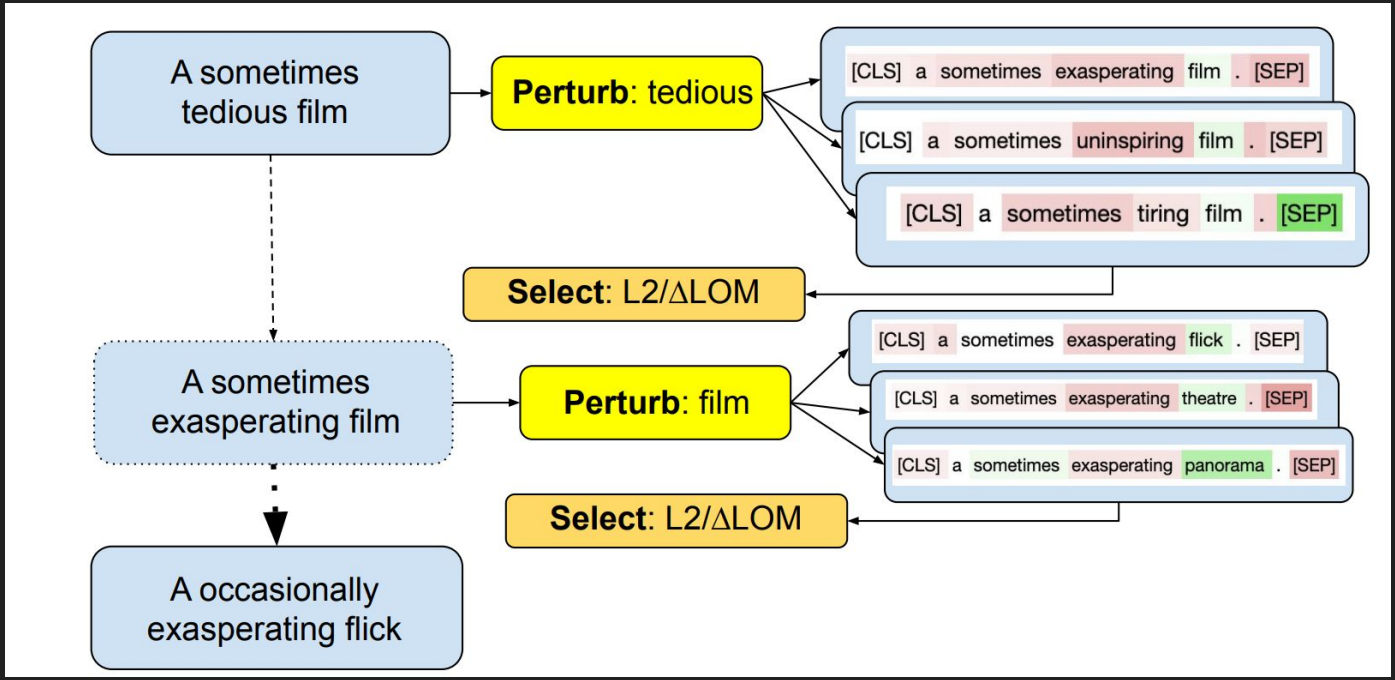
2. Start from most important word and substitute it with k NN

3. For each ford check interpretation

Perturbation constraints

- Repeat Modification
- Stop Word Modification
- Word Embedding Distance
- Part of Speech
- Sentence Embedding

Algorithm schema



ExplainFooler algorithm

Result: A - list of candidate sentences ordered by number of words perturbed from original
For each sentence in dataset
$A \leftarrow$empty
$S \leftarrow$original sentence
$I_0 \leftarrow$ InterpretMethod(S)
$P \leftarrow$ordered list of important words (LOO)
while *<=50% of words perturbed from P* **do**
  $w \leftarrow P[0]$
  $C \leftarrow$empty
  while *Possible perturbations exist* **do**
    $c \leftarrow$Perturb S and get candidate
    if *constraints pass* ***and*** *prediction label is same as S* **then**
      $I \leftarrow$ InterpretMethod(c)
      $\Delta diff \leftarrow diff(I_0, I)$,
      $C \leftarrow C \cup (\Delta diff, c)$
    else
      continue
  $A \leftarrow A \cup$ c where max(diff)
  $P \leftarrow$ remove $P[0]$

Algorithm 1: The "ExplainFooler" algorithm

Datasets

- SST-2 The Stanford Sentiment Treebank-2
- AG News
- IMDB

Example

True Label	Predicted Label	Confidence	Predicted Class	Words Perturbed	LOM Score	Word Importance
0	0	0.98	Negative	0	0.00	[CLS] a sometimes tedious film . [SEP]
0	0	0.97	Negative	1	0.82	[CLS] a sometimes exasperating film . [SEP]
0	0	0.86	Negative	2	0.92	[CLS] a sometimes exasperating flick . [SEP]
0	0	0.57	Negative	3	0.95	[CLS] a occasionally exasperating flick . [SEP]

Example DistilBERT_uncased, IG

True Label	Predicted Label	Confidence	Predicted Class	Words Perturbed	LOM Score	Word Importance
1	1	1.00	Positive	0	0.00	[CLS] it ' s a charming and often affecting journey . [SEP]
1	1	1.00	Positive	1	0.31	[CLS] it ' s a alluring and often affecting journey . [SEP]
1	1	0.98	Positive	2	0.80	[CLS] it ' s a beautifull and often afflicted journey . [SEP]
1	1	0.72	Positive	3	1.22	[CLS] it ' s a beautifull and often afflicted rook . [SEP]
0	0	0.98	Negative	0	0.00	[CLS] unflinchingly bleak and desperate [SEP]
0	0	0.89	Negative	1	0.72	[CLS] unflinchingly baleful and desperate [SEP]
0	0	0.66	Negative	2	0.86	[CLS] unflinchingly dusky and depressive [SEP]
0	0	0.99	Negative	0	0.00	[CLS] it ' s slow - - very , very slow . [SEP]
0	0	0.99	Negative	1	0.25	[CLS] it ' s slow - - crucially , very slow . [SEP]
0	0	0.99	Negative	2	0.91	[CLS] it ' s slow - - crucially , vitally slow . [SEP]
0	0	0.97	Negative	3	1.44	[CLS] it ' s slow - - crucially , highly lent . [SEP]
0	0	0.91	Negative	4	1.61	[CLS] it ' s sluggish - - crucially , highly lent . [SEP]

Same example using RoBERTa-base

True Label	Predicted Label	Confidence	Predicted Class	Words Perturbed	LOM Score	Word Importance
1	1	1.00	Positive	0	0.00	it 's a charming and often affecting journey . #/s
1	1	1.00	Positive	1	0.73	it 's a mignon and often affecting journey . #/s
1	1	0.99	Positive	2	1.19	it 's a dreamy and often plaguing journey . #/s
1	1	0.98	Positive	3	1.12	it 's a dreamy and often effect nomad . #/s
0	0	1.00	Negative	0	0.00	unflinchingly bleak and desperate #/s
0	0	1.00	Negative	1	0.45	unflinchingly dreary and desperate #/s
0	0	0.90	Negative	2	0.76	unflinchingly sombre and frenetic #/s
0	0	1.00	Negative	0	0.00	it 's slow -- very , very slow . #/s
0	0	1.00	Negative	1	0.78	it 's slow -- very , very lent . #/s
0	0	0.97	Negative	2	0.73	it 's slow -- perfectly , very slowness . #/s
0	0	1.00	Negative	3	0.31	it 's slow -- immeasurably , immeasurably slowness . #/s
0	0	1.00	Negative	4	0.39	it 's slowest -- immeasurably , immeasurably slowness . #/s

SST-2									
	DistilBERT			RoBERTa			BERT-adv		
Ratio	L2	ΔLOM	Random	L2	ΔLOM	Random	L2	ΔLOM	Random
0-0.1	0.65	0.78	0.8	0.64	0.76	0.81	0.53	0.6	0.73
0.1-0.2	0.53	0.65	0.64	0.57	0.61	0.69	0.43	0.43	0.52
0.2-0.3	0.42	0.55	0.59	0.51	0.59	0.6	0.3	0.33	0.42
0.3-0.4	0.36	0.48	0.48	0.47	0.47	0.55	0.35	0.3	0.43
0.4-0.5	0.31	0.42	0.47	0.42	0.43	0.48	0.14	0.24	0.36

Table 1: Change in average rank-order correlation using metrics - L2 Norm, LOM and random selection conmputed using the interpretability method: INTEGRATED GRADIENT, for dataset- SST-2 over 3 models - DistilBERT, RoBERTa and BERT-adv.

SST-2									
	DistilBERT			RoBERTa			BERT-adv		
Ratio	L2	ΔLOM	Random	L2	ΔLOM	Random	L2	ΔLOM	Random
0-0.1	0.77	0.78	0.81	0.75	0.76	0.81	0.75	0.76	0.79
0.1-0.2	0.71	0.71	0.73	0.71	0.71	0.74	0.68	0.68	0.7
0.2-0.3	0.67	0.68	0.68	0.68	0.69	0.7	0.63	0.64	0.65
0.3-0.4	0.65	0.65	0.65	0.66	0.67	0.67	0.61	0.61	0.64
0.4-0.5	0.6	0.62	0.62	0.63	0.63	0.65	0.59	0.56	0.63

Table 2: Change in average Top-50% intersection using metrics - L2 Norm, LOM and random selection conmputed using the interpretability method: INTEGRATED GRADIENT, for dataset- SST-2 over 3 models - DistilBERT, RoBERTa and BERT-adv.

SST-2									
	DistilBERT			RoBERTa			BERT-adv		
Ratio	L2	ΔLOM	Random	L2	ΔLOM	Random	L2	ΔLOM	Random
0-0.1	0.64	0.7	0.79	0.59	0.66	0.76	0.57	0.68	0.72
0.1-0.2	0.52	0.58	0.65	0.58	0.63	0.7	0.37	0.52	0.59
0.2-0.3	0.46	0.51	0.56	0.52	0.58	0.62	0.34	0.47	0.54
0.3-0.4	0.39	0.43	0.46	0.48	0.54	0.58	0.31	0.36	0.36
0.4-0.5	0.23	0.29	0.46	0.55	0.55	0.54	0.28	0.2	0.24

Table 3: Change in average rank-order correlation using metrics - L2 Norm, LOM and random selection conmputed using the interpretability method: LIME, for dataset- SST-2 over 3 models - DistilBERT, RoBERTa and BERT-adv.

SST-2									
	DistilBERT			RoBERTa			BERT-adv		
Ratio	L2	ΔLOM	Random	L2	ΔLOM	Random	L2	ΔLOM	Random
0-0.1	0.64	0.7	0.79	0.59	0.66	0.76	0.57	0.68	0.72
0.1-0.2	0.52	0.58	0.65	0.58	0.63	0.7	0.37	0.52	0.59
0.2-0.3	0.46	0.51	0.56	0.52	0.58	0.62	0.34	0.47	0.54
0.3-0.4	0.39	0.43	0.46	0.48	0.54	0.58	0.31	0.36	0.36
0.4-0.5	0.23	0.29	0.46	0.55	0.55	0.54	0.28	0.2	0.24

Table 4: Change in average Top-50% intersection using metrics - L2 Norm, LOM and random selection conmputed using the interpretability method: LIME, for dataset- SST-2 over 3 models - DistilBERT, RoBERTa and BERT-adv.

Thank you :)