PERTURBING INPUTS FOR FRAGILE INTERPRETATIONS

Michał Tyrolski, Emilia Wiśnios

CONTEXT OF WORK

Model Manipulations - defined as a model fine-tuning step that aims to radically alter the explanations without hurting the accuracy of the original models

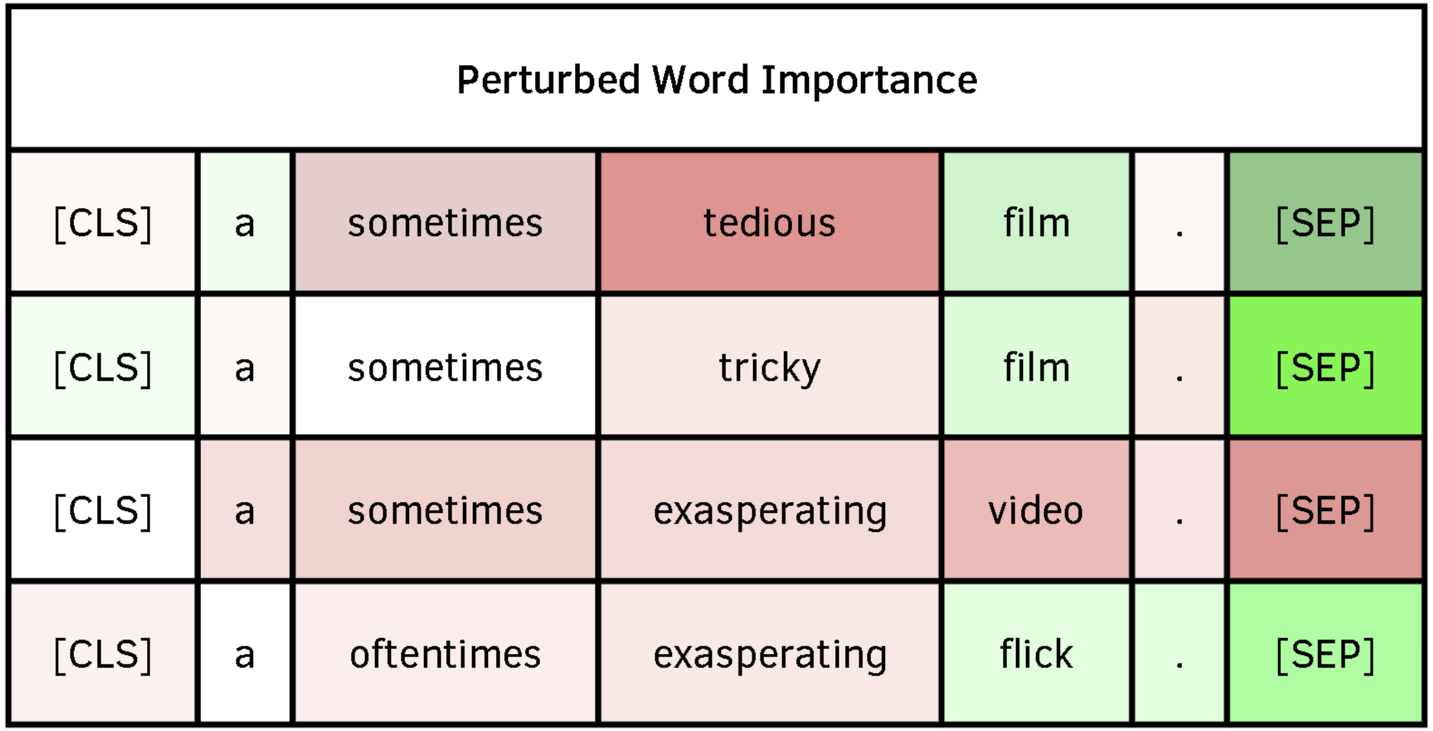
Input Manipulations - defined as a process of modifying the input data in order to get different explanations

CONTEXT OF WORK

Model Manipulations - defined as a model fine-tuning step that aims to radically alter the explanations without hurting the accuracy of the original models

Input Manipulations - defined as a process of modifying the input data in order to get different explanations [DEEP NLP]

IDEA



Perturbed Word Importance						
[CLS]	a	sometimes	tedious	film	.	[SEP]
[CLS]	a	sometimes	tricky	film	.	[SEP]
[CLS]	a	sometimes	exasperating	video	.	[SEP]
[CLS]	a	oftentimes	exasperating	flick	.	[SEP]

CONTRIBUTIONS

The need for robust interpretations in high-stakes areas such as medicine or finance for trustworthy NLP applications

Demonstration of how interpretations can be manipulated through simple word perturbations on input text

The importance of understanding the potential for manipulation in NLP models for responsible and ethical use in high-stakes areas.

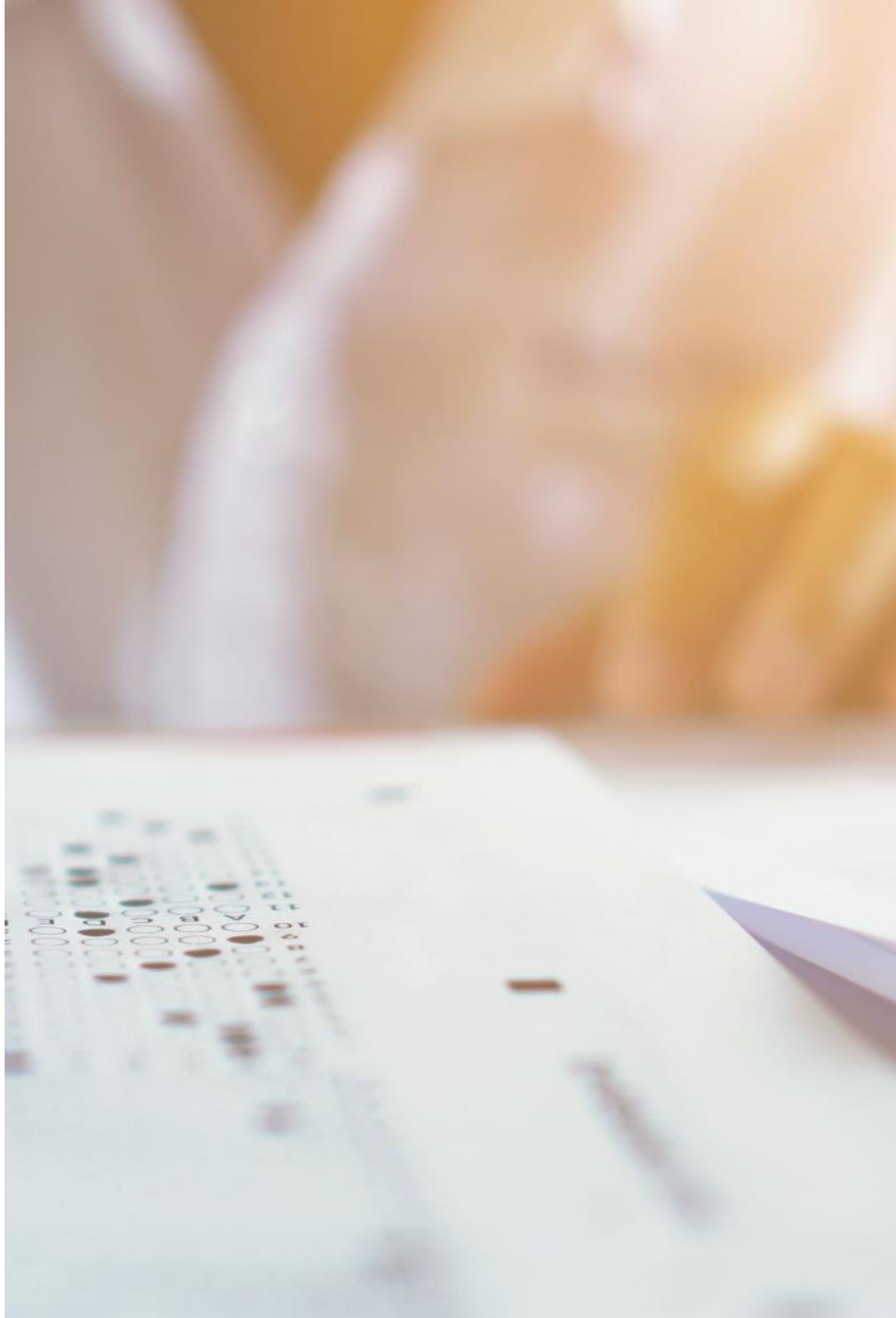
EXPLAINFOOLER

Input

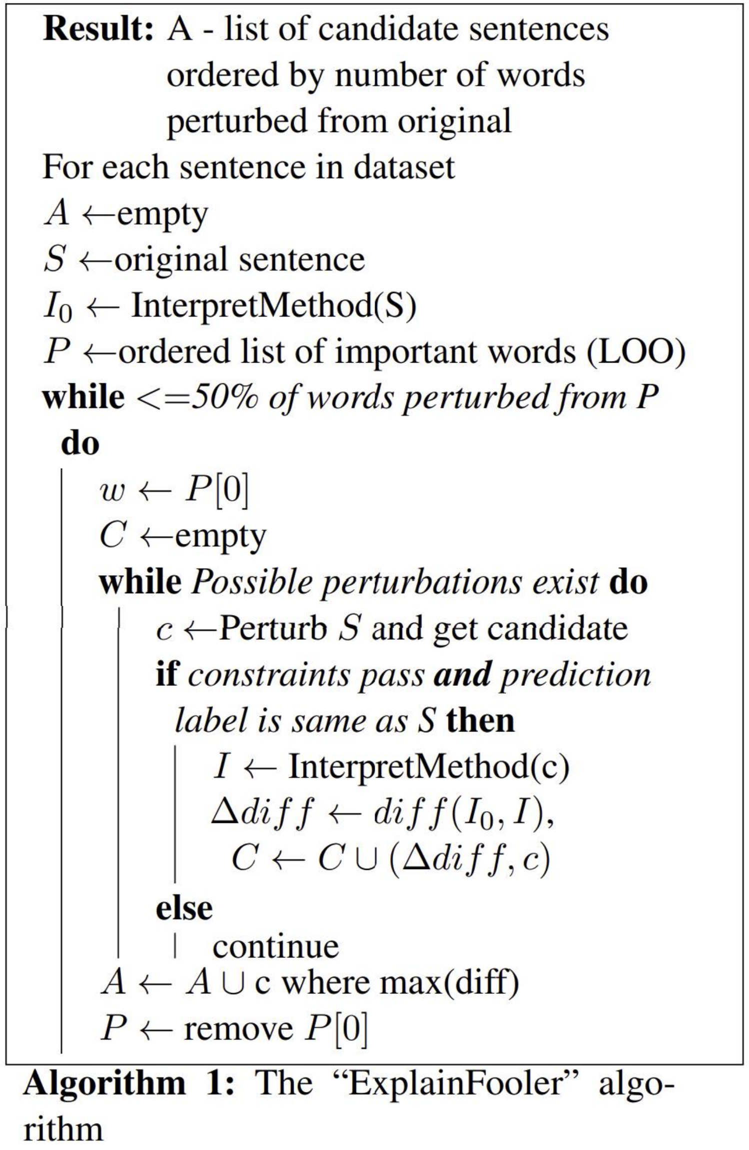
- Orginal Sentence
- Ordered list of important words
- Interpretation method

Output:

- List of candidate sentences ordered by number of words perturbed from original text



ALGORITHM



```
Result: A - list of candidate sentences
        ordered by number of words
        perturbed from original
For each sentence in dataset
$A$ ←empty
$S$ ←original sentence
$I_0$ ← InterpretMethod(S)
$P$ ←ordered list of important words (LOO)
while <=50% of words perturbed from P
 do
    $w \leftarrow P[0]$
    $C$ ←empty
    while Possible perturbations exist do
        $c$ ←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

FINDING IDEAL CANDIDATE

1

Obtaining all candidates and their metric scores for every candidate achieving the same prediction label as the original

2

Storing ideal candidates with each 'm' number of words perturbed, giving a list of candidates for each level of word perturbation

3

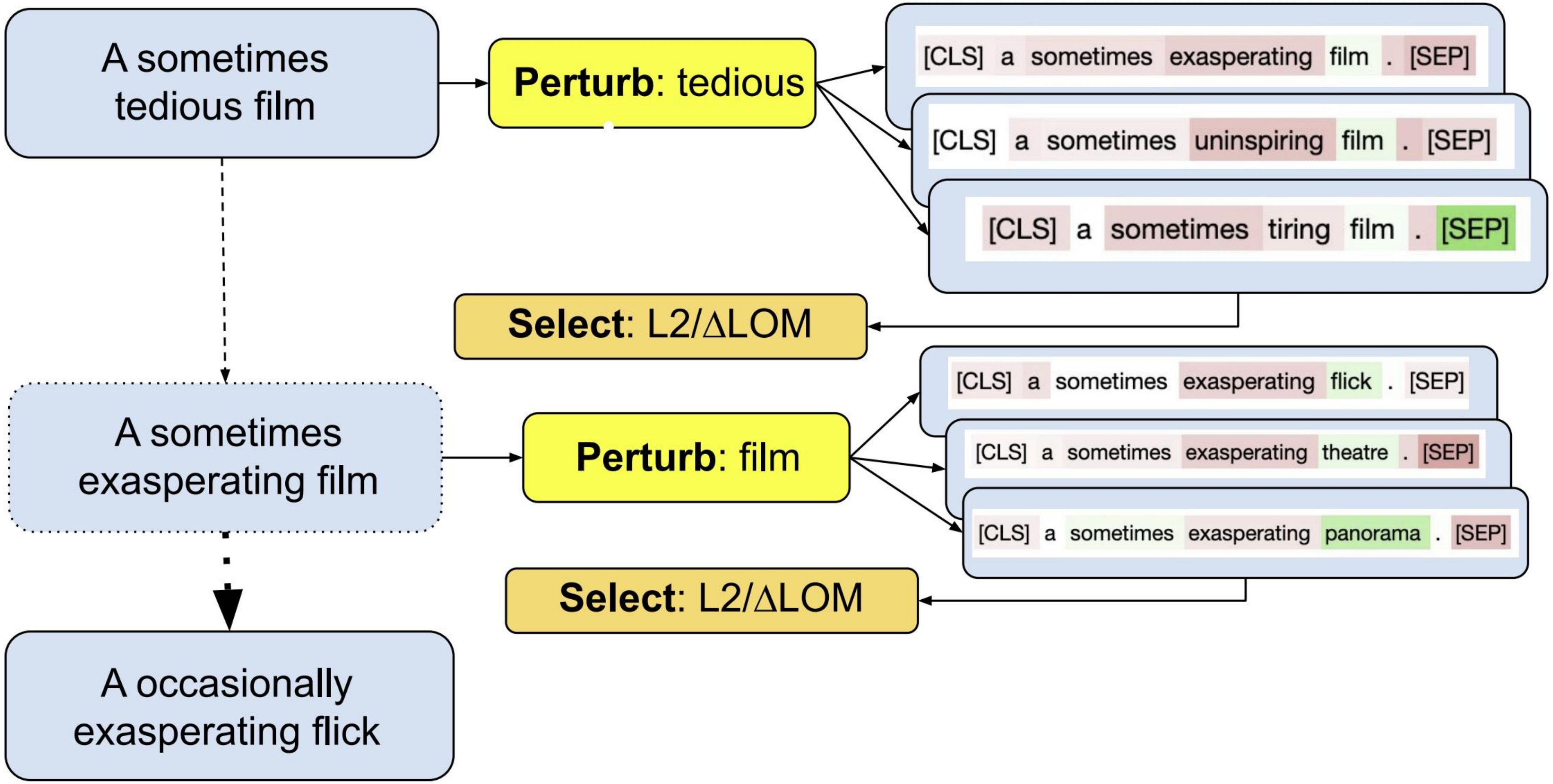
Choosing the candidate with the highest metric score against the original for each level

4

Converting the number of perturbed words into a ratio with respect to the input’s length to take into account varying sentence lengths and get a normalized measure

5

Limiting the ratio to 50% to avoid losing semantic meaning when more than half of the words are perturbed



A sometimes tedious film

Perturb: tedious

[CLS] a sometimes exasperating film . [SEP]

[CLS] a sometimes uninspiring film . [SEP]

[CLS] a sometimes tiring film . [SEP]

Select: L2/ΔLOM

A sometimes exasperating film

Perturb: film

[CLS] a sometimes exasperating flick . [SEP]

[CLS] a sometimes exasperating theatre . [SEP]

[CLS] a sometimes exasperating panorama . [SEP]

Select: L2/ΔLOM

A occasionally exasperating flick

HOW TO COMPARE TWO INTERPRETATIONS?

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

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

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

Proposing two objective metrics, "Delta LOM" and "L2 Norm" to quantify the difference between two interpretations

Metrics are divergent, meaning that the higher the metric, the more different the interpretations are.

EVALUATION METRICS

$$Intersection = \frac{\bigcap(argsort(I_1), argsort(I_2))}{0.5 * length(I_1)} \tag{5}$$

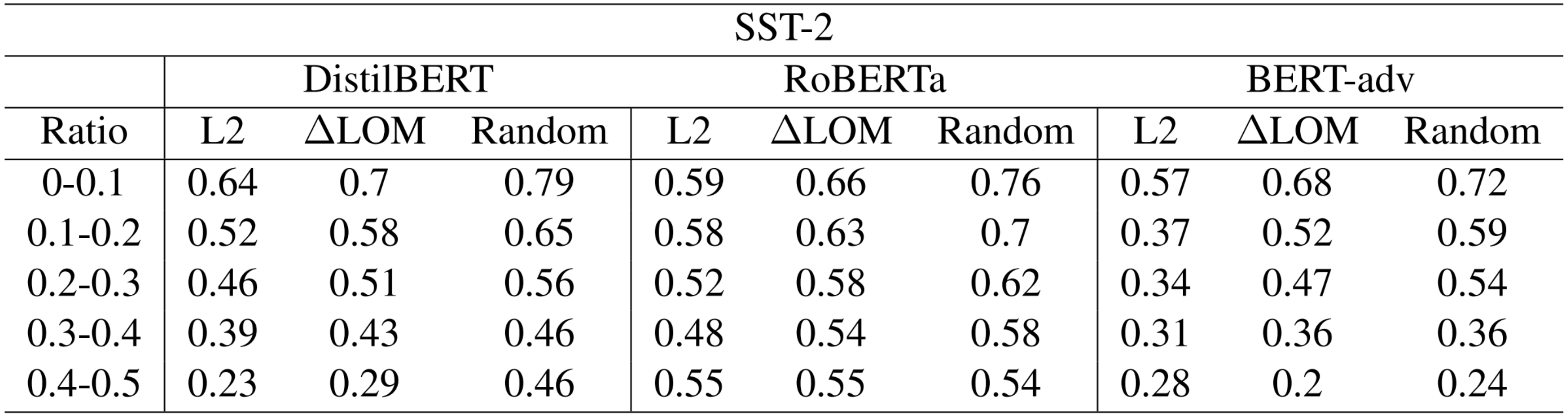
where argsort returns the indices of the top-50% of the words in a sentence with highest attributions.

$$R - Correlation = \max(0, Spearman(I_1, I_2))$$

To compare the correlation between interpretations of 2 sentences, we use the Spearman rank correlation metric. The more the ranks of the interpretations agree with each other, the higher the rank correlations.

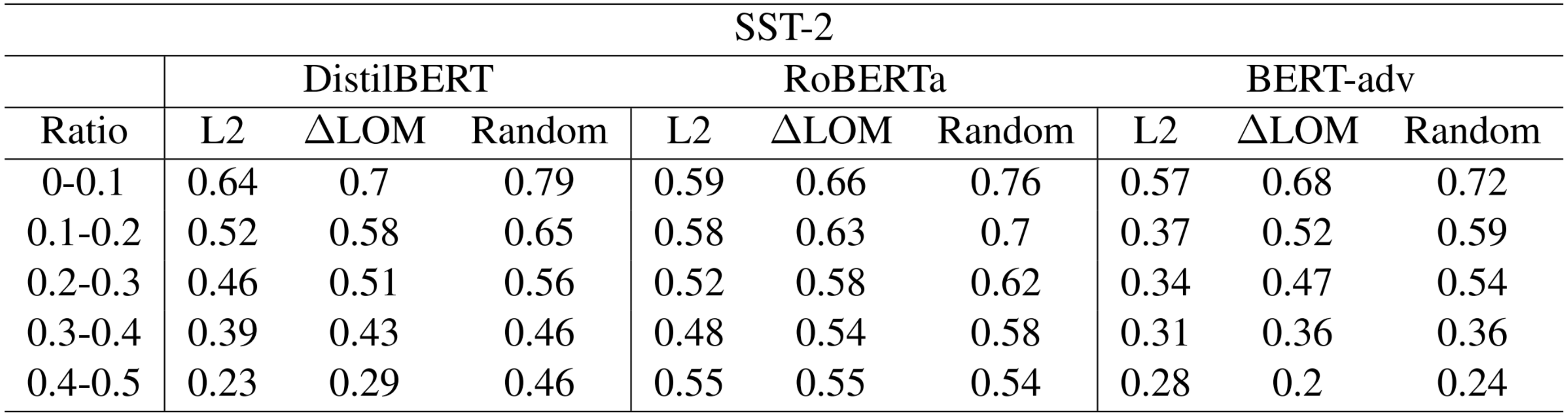
To compare the extent to which the words with the highest attributions are correctly predicted by both interpretation methods, we use the Top-k% intersection metric.

RESULTS



	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.

SUMMARY

Growing emphasis on interpretation techniques for explaining NLP model predictions in literature

Novel algorithm for generating perturbed inputs that provide evidence of fragile interpretations

Effectiveness of the approach demonstrated across three different models, including one adversarially trained

Results show it is possible to attack interpretations using simple input-level word swaps under certain constraints

Both black and white-box interpretability approaches (LIME and INTEGRATED GRADIENT) shown to be fragile in derived interpretations

Findings can pave way for future studies on defending against the problem of fragile interpretations in NLP.