Adversarial Text in NLP: Improving robustness via XAI

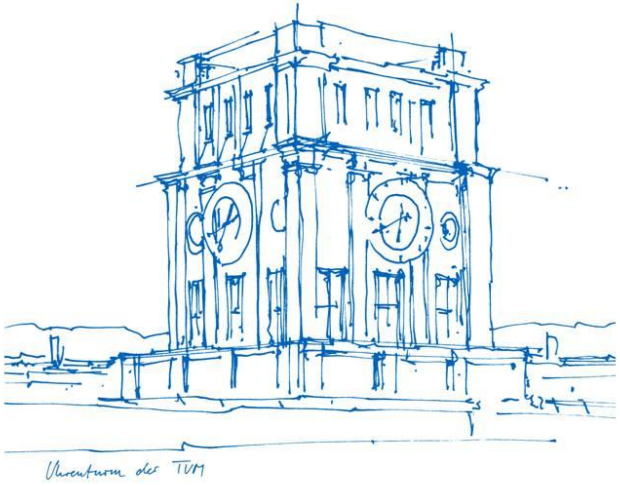
Technische Universität München

Faculty of Informatics

XAI Lab Course, SS2022

Final Presentation

10.08.2022



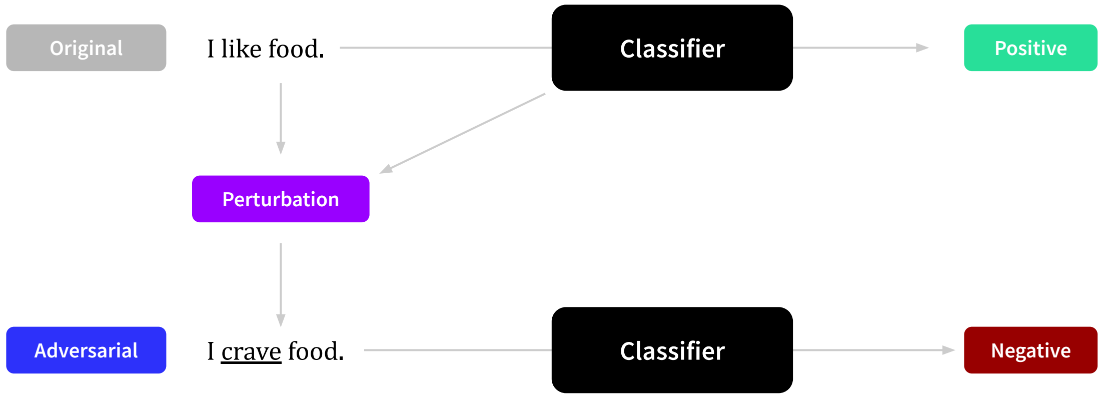
Niklas Hölterhoff

Karl Richter

Yifeng Dong

Fooling the Classifier: Adversarial attacks

Fooling the Classifier: Adversarial attacks



Types of adversarial-attacks

Character-level

Original: I would put this at the top of my list of films in the category of unwatchable trash!

Adversarial: I would put this at the top of my list of films in the category of unWatchable trash!

Pruthi
Deepwordbug
TextBugger

Word-level

Original: I would put this at the top of my list of films in the category of unwatchable trash!

Adversarial: I would put this at the top of my list of films in the category of insatiable trash!

PWWS
TextFooler
Alzantot

Sentence-level

Original: I would put this at the top of my list of films in the category of unwatchable trash!

Adversarial: This is the first piece of work that falls into the category of trash that I can't stand to look at!

StyleAdv
SCPN
GAN

Types of adversarial attacks

Character-level

Original	I would put this at the top of my list of films in the category of unwatchable trash!
Adversarial	I would put this at the top of my list of films in the category of unWatchable trash!

No attack: 89% Acc

Pruthi 1-char attack: 60% Acc [BERT on MRPC]

[Pruthi et al. 2019 @ Carnegie Mellon, https://arxiv.org/pdf/1905.11268.pdf]

Word-level

Original	I would put this at the top of my list of films in the category of unwatchable trash!
Adversarial	I would put this at the top of my list of films in the category of insatiable trash!

PWWS: 78.8% ASR [BERT, SST-2]

[Zeng et al. 2019 @ Tsingua, https://arxiv.org/pdf/2009.09191.pdf]

TextFooler: 90% ASR [BERT, SST-2]

[Zeng et al. 2019 @ Tsingua, https://arxiv.org/pdf/2009.09191.pdf]

Sentence-level

Original	I would put this at the top of my list of films in the category of unwatchable trash!
Adversarial	This is the first piece of work that falls into the category of trash that I can't stand to look at!

GAN: 26-47% ASR

[Zhao et al. 2018 @ UCI, https://arxiv.org/abs/1710.11342]

SCPN: 52-64% ASR

[Iyyer et al. 2018 @ Stanford, https://arxiv.org/abs/1804.06059]

StyleAdv: 91-96% ASR

[Qi et al. 2021 @ Tsinghua, https://arxiv.org/abs/2110.07139]

Types of adversarial defenses

- ❑ There are three options to mitigate the risk of adversarial attacks
  - ❑ [Training] Make the model more robust
  - ❑ [Inference] Pre-process Inputs
  - ❑ **[Inference] Detect adversarial samples**

Research Plan

Experiment 1 **RQ1:** How do word-level defense mechanisms perform on sentence-level attacks?

Experiment 1: WDR Background

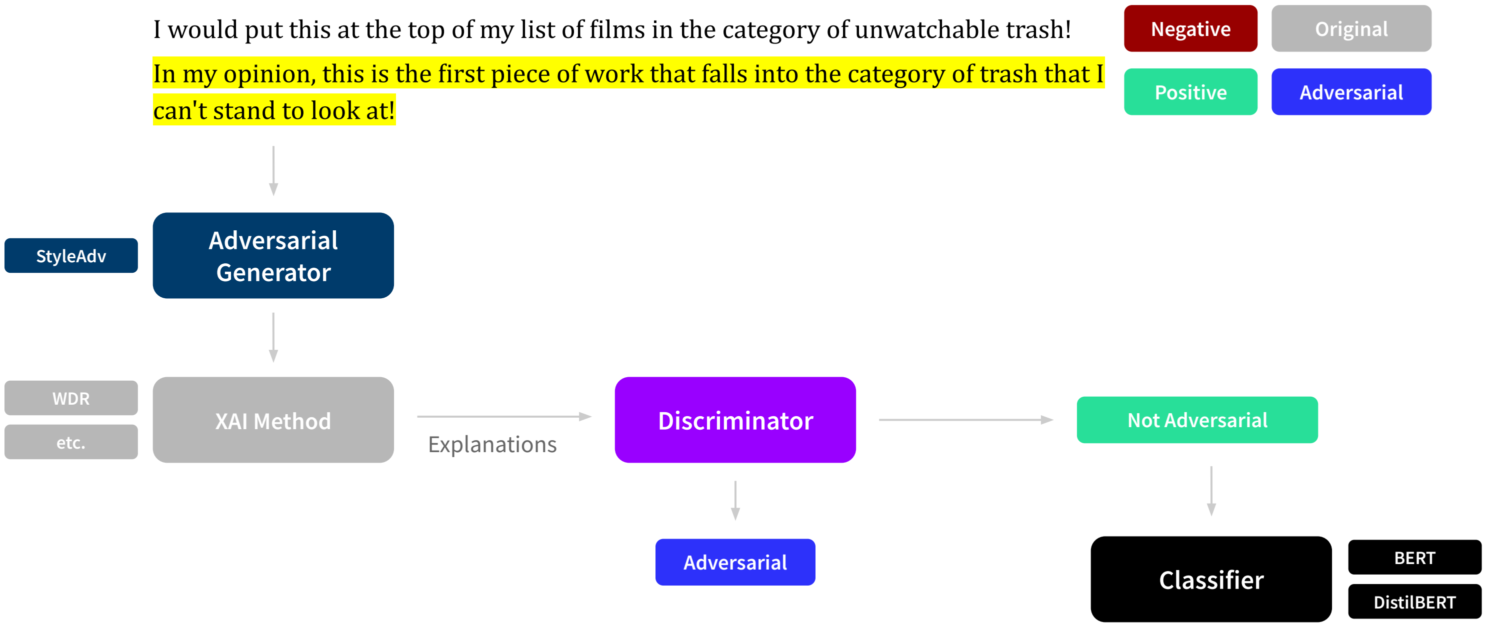
- WDR compares model predictions between original and when a word is removed:

```
['one ', 'of ', 'the ', 'best', 'of ', 'the ', 'year']

['one ', 'of ', 'the ', <unk> , 'of ', 'the ', 'year']
```

Experiment 1: Design

Benchmark the **WDR-based detector** on **StyleAdv** sentence-level attacks.



Experiment 1: Results

Benchmarking of WDR on word- and sentence-level attacks

Discriminator trained on word-level attacks (PWWS)

Attack	Type	Dataset	Encoder	F1-Score
PWWS	Word-level	AG-News	DistilBERT	**0.936**
BAE	*Word-level*	*AG-News*	*DistilBERT*	***0.864***
TextFooler	*Word-level*	*AG-News*	*DistilBERT*	***0.957***
StyleAdv	Sentence-level	AG-News	DistilBERT	**0.707**
StyleAdv	Sentence-level	SST-2	BERT	**0.706**
Pruthi et al.	Character-level	SST-2	BERT	**0.676**

= Training configuration

Experiment 1: Results

Benchmarking of WDR on word- and sentence-level attacks

Discriminator trained on sentence-level attacks (StyleAdv)

Attack	Type	Dataset	Encoder	F1-Score
PWWS	Word-level	AG-News	DistilBERT	**0.806**
TextFooler	Word-level	IMDB	DistilBERT	**0.883**
StyleAdv	Sentence-level	SST-2	BERT	**0.721**
Pruthi et al.	Char-level	SST-2	BERT	**0.699**

Discriminator trained on char-level attacks (Pruthi et al.)

Attack	Type	Dataset	Encoder	F1-Score
PWWS	Word-level	AG-News	DistilBERT	**0.838**
TextFooler	Word-level	IMDB	DistilBERT	**0.894**
StyleAdv	Sentence-level	SST-2	BERT	**0.671**
Pruthi et al.	Char-level	SST-2	BERT	**0.700**

= Training configuration

Research Plan

Experiment 1

RQ1: How do word-level defense mechanisms perform on sentence-level attacks?

Result: WDR works reasonable for StyleAdv, but significantly worse compared to word-level attacks.

Research Plan

Experiment 1 **RQ1:** How do word-level defense mechanisms perform on sentence-level attacks?

Result: WDR works reasonable for StyleAdv, but significantly worse than for word-level attacks.

Experiment 2 **RQ2:** How can word-level defenses be improved to perform better on a sentence-level?

Experiment 2: Explore sentence-level defenses

- ❏ Approach 1: Extend WDR by masking multiple consecutive tokens
- ❏ Approach 2: Extend WDR by using a ML to fill masked tokens

Experiment 2: Mask multiple tokens

- WDR compares model predictions between original and when a word is removed:

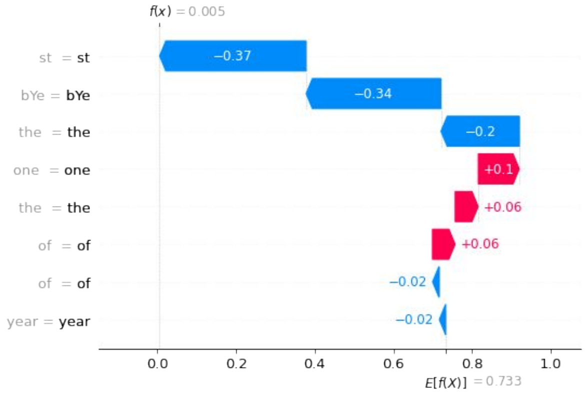
```
['one ', 'of ', 'the ', 'best', 'of ', 'the ', 'year']
                          ↓
['one ', 'of ', 'the ', <unk> , 'of ', 'the ', 'year']
```

Challenge: Tokenizer

- Character-level attacks are often fooling the tokenizer

Original "One of the best of the year" → ['one ', 'of ', 'the ', **'best '**, 'of ', 'the ', 'year']

Adversarial "One of the bYest of the year" → ['one ', 'of ', 'the ', **'bYe'**, **'st '**, 'of ', 'the ', 'year']



Challenge: Tokenizer

- ❏ Character-level attacks are often fooling the tokenizer
- ❏ Sentence-level attacks may be

Experiment 2: Mask multiple tokens

- ❏ Sentence-level attacks might be distributing perturbation across entire phrases
- ❏ → Multiple token masking

Experiment 2: Mask multiple tokens

- Masking multiple consecutive tokens:

```
['one ', 'of ', 'the ',  'best', 'of ', 'the ', 'year']

                              ↓

['one ', 'of ', 'the ',      <unk>    , 'the ', 'year']
```

Experiment 2: Fill tokens using LM

- Use a large language model (such as BERT) to fill in the gaps

```
['one ', 'of ', 'the',  'best', 'of ', 'the ', 'year']
```

↓

```
['one ', 'of ', 'the ',   'greatest  , 'of ', 'the ', 'year']
```

Experiment 2: Fill tokens using LM

- Using `<unk>` might introduce a distributional bias
- Use a large language model (e.g. BERT) to fill masks.

Experiment 2: Results

- WDR Performance with multiple token masking and BERT-masking
- Training and testing on sentence-level attack:

Attack	Defense	Type	Dataset	Model	F1
StyleAdv	*WDR*	*Sentence-level*	*SST-2*	*BERT*	*0.721*
StyleAdv	**WDR-Mask 1-2**	Sentence-level	SST-2	BERT	**0.728**
StyleAdv	**WDR-Mask 1-3**	Sentence-level	SST-2	BERT	**0.734**
StyleAdv	**WDR-Mask 1-4**	Sentence-level	SST-2	BERT	**0.734**
StyleAdv	**WDR+BERT Mask**	Sentence-level	SST-2	BERT	**0.714**

Experiment 2: Results

- WDR Performance with multiple token masking and BERT-masking
- **Generalization** performance when trained on sentence-level attack:

Attack	Defense	Type	Dataset	Model	F1
StyleAdv	**WDR-Mask 1-3**	Sentence-level	SST-2	BERT	**0.734**
Pruthi	***WDR***	*Character-level*	*SST-2*	*BERT*	***0.699***
Pruthi	**WDR-Mask 1-3**	Character-level	SST-2	BERT	**0.720**
PWWS	***WDR***	*Word-level*	*AG-News*	*DistilBERT*	***0.819***
PWWS	**WDR-Mask 1-3**	Word-level	AG-News	DistilBERT	**0.856**

= Training configuration

Experiment 2: Results

- ❑ WDR Performance with multiple token masking and BERT-masking
- ❑ Training and testing on character-level attack:

Attack	Defense	Type	Dataset	Encoder	F1-Score
Pruthi	*WDR*	*Character-level*	*SST-2*	*BERT*	*0.700*
Pruthi	**WDR-Mask 1-3**	Character-level	SST-2	BERT	**0.697**
Pruthi	**WDR +BERT Mask**	Character-level	SST-2	BERT	**0.676**

Experiment 2: Analysis

- Multiple token masking helps with sentence level attacks, but not with character-level
- BERT-masking is not helpful

Research Plan

Experiment 1

RQ1: How do word-level defense mechanisms perform on sentence-level attacks?

Result: WDR works reasonable for StyleAdv, but significantly worse than for word-level attacks.

Experiment 2

RQ2: How can word-level defenses be improved to perform better on a sentence-level?

Result: Multiple token masking with WDR can improve performance against sentence-level attacks.

Research Plan

Experiment 1

RQ1: How do word-level defense mechanisms perform on sentence-level attacks?

Result: WDR works reasonable for StyleAdv, but significantly worse than for word-level attacks.

Experiment 2

RQ2: How can word-level defenses be improved to perform better on a sentence-level?

Result: Multiple token masking with WDR can improve performance against sentence-level attacks.

Experiment 3

RQ3: Which layers are differently activated? How well does a discriminator detect adversarial examples?

Adversarial Layer Attribution

BERT

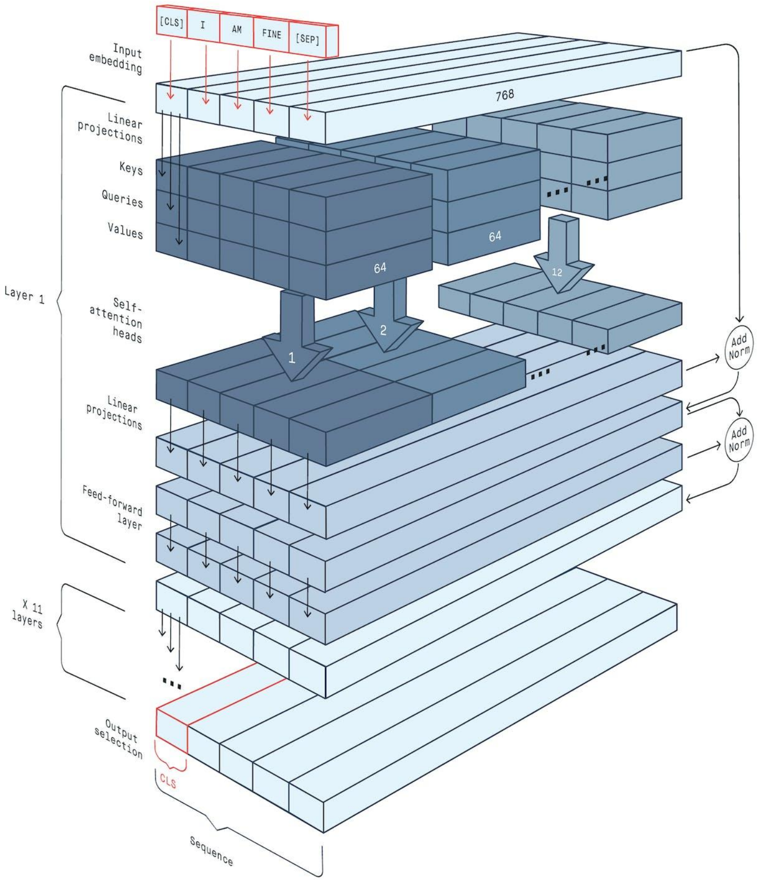
- ❏ 12 layers + 1 embedding layer
- ❏ Each layer has a hidden size of 768

Layer Attribution

- ❏ attribution of a token in the input on the each neuron in a layer

Attribution handling

- ❏ Concat all attributions (128 x 768)
- ❏ Sum of attributions (128 x 1)
- ❏ Neuron with max. diff. between orig. and adv. samples (128 x 1)



Classifier performance by Attribution method

Configuration

Model: BERT

Dataset: AG-News

Layer: Embedding

Classifier: RandomForest

Samples: 1000

Neuron activation diff.

$\mathrm{attr}_t = \sum_n (\mathrm{attr}_{\mathrm{orig},t,n} - \mathrm{attr}_{\mathrm{adv},t,n})^2$

t tokens

n neurons

Method	**Attributions**	**F1-Score**
Layer Activation	Sum	0.62
Layer Gradient X Activation	Sum	0.60
Layer Integrated Gradients	Sum	0.59
Layer Activation	Max. neuron	0.63 (Layer 9, neuron 381)
Layer Gradient X Activation	Max. neuron	0.62 (Layer 9, neuron 308)
Layer Integrated Gradients	Max. neuron	0.62 (Layer 9, neuron 308)
Layer Activation	Concat	0.65
Layer Gradient X Activation	Concat	0.68
Layer Integrated Gradients	Concat	0.66

Classifier performance by Attribution method

Configuration

Model: BERT

Dataset: AG-News

Layer: Embedding

Classifier: RandomForest

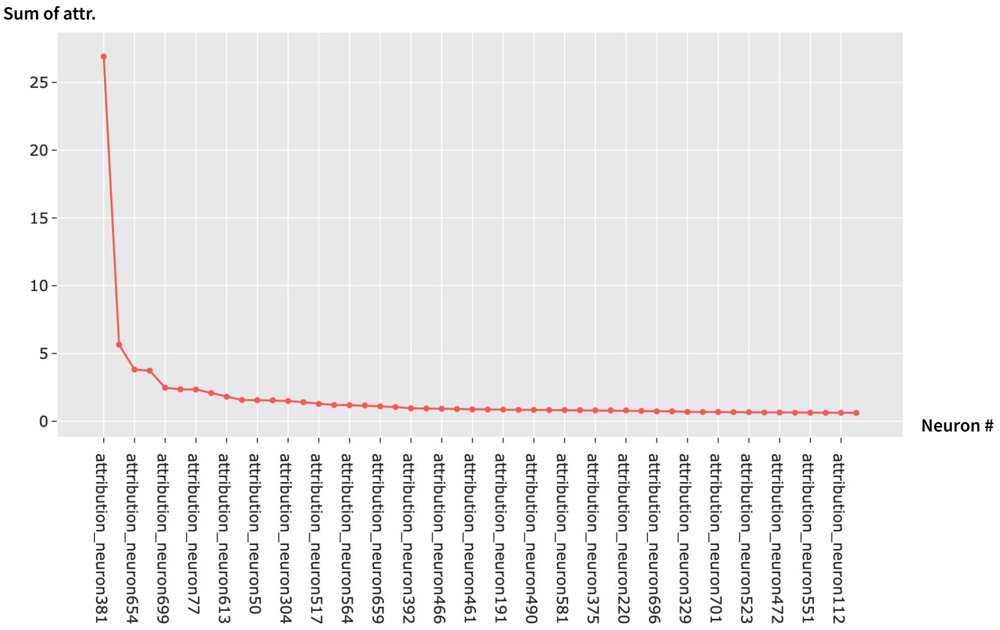
Samples: 1000

Neuron activation diff.

$\mathrm{attr}_t = \sum_n (\mathrm{attr}_{\mathrm{orig},t,n} - \mathrm{attr}_{\mathrm{adv},t,n})^2$

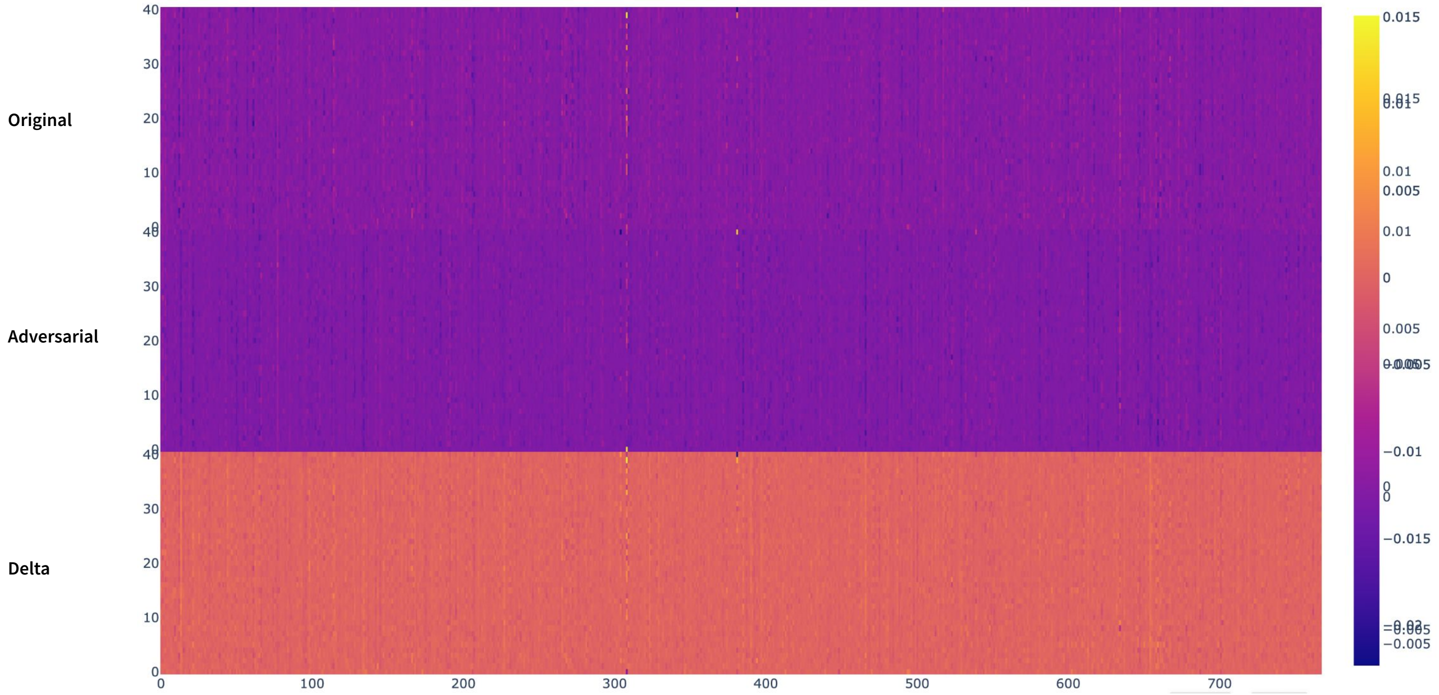
t tokens

n neurons



The difference between the activation of neuron 381 in Layer Activation and neuron 308 in Layer Integrated Gradients is the strongest when comparing original and adv. examples.

Classifier performance by Attribution method



Original: Privacy questions arise as RFID hits stores BALTIMORE--Proponents of radio frequency identification used to have a quick and easy response to consumer advocates charging that the technology posed an alarming threat to privacy.
Adversarial: secrecy inquiry uprise as RFID striking stores BALTIMORE--Proponents of radio frequency identification practice to have a quick and easy response to consumer advocates shoot that the technology posed an alarming threat to privacy.

Classifier performance by Attribution layer

Configuration

Model: BERT

Dataset: AG-News

Method: Layer Integrated Gradients

Attributions: Concat

Classifier: RandomForest

Samples: 1000

LM	Layer	F1-Score
BERT	Embedding Layer	0.65
BERT	Layer 1	0.67
BERT	Layer 2	0.72
BERT	Layer 3	0.68
BERT	Layer 4	0.72
BERT	Layer 5	0.62
BERT	Layer 6	0.65

LM	Layer	F1-Score
BERT	Layer 7	0.68
BERT	Layer 8	0.64
BERT	Layer 9	0.72
BERT	Layer 10	0.68
BERT	Layer 11	0.64
BERT	Layer 12	0.61

Classifier performance by classifier type

Configuration

Model: BERT

Dataset: AG-News

Method: Layer Integrated Gradients

Attributions: Concat

Samples: 1000

Classifier	F1-Score
Gaussian Naive Bayes	0.55
Bernoulli Naive Bayes	0.59
SGD Classifier	0.60
KNN Classifier	0.62
C-Support Vector Classifier	0.62
Random Forest Classifier	0.72
Extra Trees Classifier	0.73
Nu-Support Vector Classifier	0.74

Random Forest performed the most reliable classifications, however, depending on the configuration, we could find classifiers with better classification metrics.

Classifier performance by attack method on SST-2

Configuration

Model: BERT
Dataset: SST-2
Layer: Embedding
Classifier: RandomForest
Samples: 1000

Method	Dataset	Attack	F1-Score
Layer Activation	SST-2	Pruthi et al.	0.63
Layer Gradient X Activation	SST-2	Pruthi et al.	0.65
Layer Integrated Gradients	SST-2	Pruthi et al.	0.64
Layer Activation	SST-2	StyleAttack	0.77
Layer Gradient X Activation	SST-2	StyleAttack	0.64
Layer Integrated Gradients	SST-2	StyleAttack	0.67

The difference between the activation of neuron 381 in Layer Activation and neuron 308 in Layer Integrated Gradients is the strongest when comparing original and adv. examples.

Classifier performance by dataset and attack method

- **Generalization** performance

Attack	Defense	Type	Dataset	Model	F1
PWWS	**Layer Integrated Gradients**	Word-level	AG-News	BERT	**0.66**
Pruthi	**Layer Integrated Gradients**	Character-level	SST-2	BERT	**0.47**

Attack	Defense	Type	Dataset	Model	F1
StyleAdv	**Layer Integrated Gradients**	Sentence-level	SST-2	BERT	**0.67**
Pruthi	**Layer Integrated Gradients**	Character-level	SST-2	BERT	**0.47**

= Training configuration

Research Plan

Experiment 1

RQ1: How do word-level defense mechanisms perform on sentence-level attacks?

Result: WDR works reasonable for StyleAdv, but significantly worse than for word-level attacks.

Experiment 2

RQ2: How can word-level defenses be improved to perform better on a sentence-level?

Result: Multiple token masking with WDR can improve performance against sentence-level attacks.

Experiment 3

RQ3: Which layers are differently activated? How well does a discriminator detect adversarial examples?

Result: The attribution of layers can be used to detect adversarial examples at a decent rate, however, the discriminator is very dataset and model specific.

Results

Experiment 1

RQ1: How do word-level defense mechanisms perform on sentence-level attacks?

Result: WDR works reasonable for StyleAdv, but significantly worse than for word-level attacks.

Experiment 2

RQ2: How can word-level defenses be improved to perform better on a sentence-level?

Result: Multiple token masking with WDR can improve performance against sentence-level attacks.

Experiment 3

RQ3: Which layers are differently activated? How well does a discriminator detect adversarial examples?

Result: The attribution of layers can be used to detect adversarial examples at a decent rate, however, the discriminator is very dataset and model specific.

Challenges & Future Work

- ❑ Attack validity of sentence-level attacks is still poor
  - ❑ Needs more work to create good attack methods / adversarial datasets
  - ❑ Explore paraphrasing as defense method
- ❑ Smarter masking / perturbation generation for calculating attributions
- ❑ Try using WDR on deeper layers
- ❑ Experiment with Layer Attributions in other BERT architectures

Q&A

Appendix

Layer Attribution Methods

Layer Activation

Layer Activation is a simple approach for computing layer attribution, returning the activation of each neuron in the identified layer.

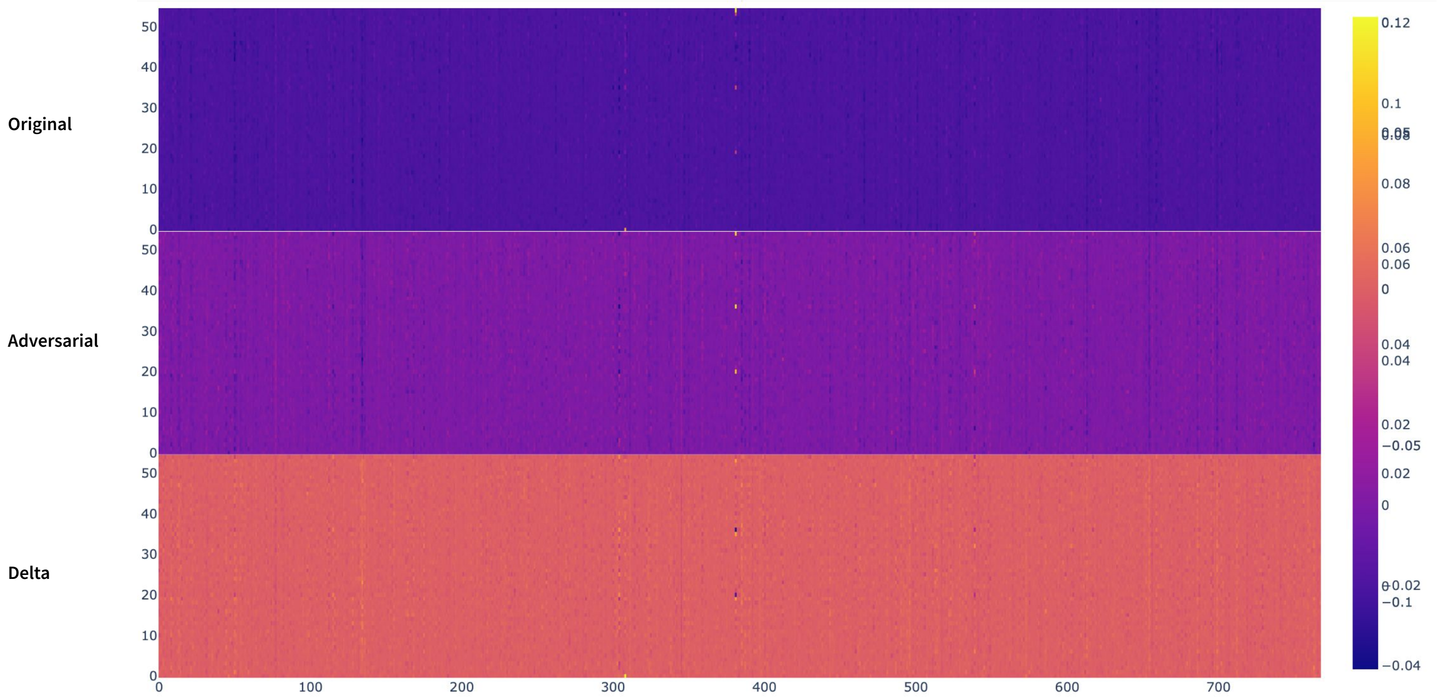
Layer Gradient X Activation

Layer Gradient X Activation is the analog of the Input X Gradient method for hidden layers in a network. It element-wise multiplies the layer's activation with the gradients of the target output with respect to the given layer.

Layer Integrated Gradients

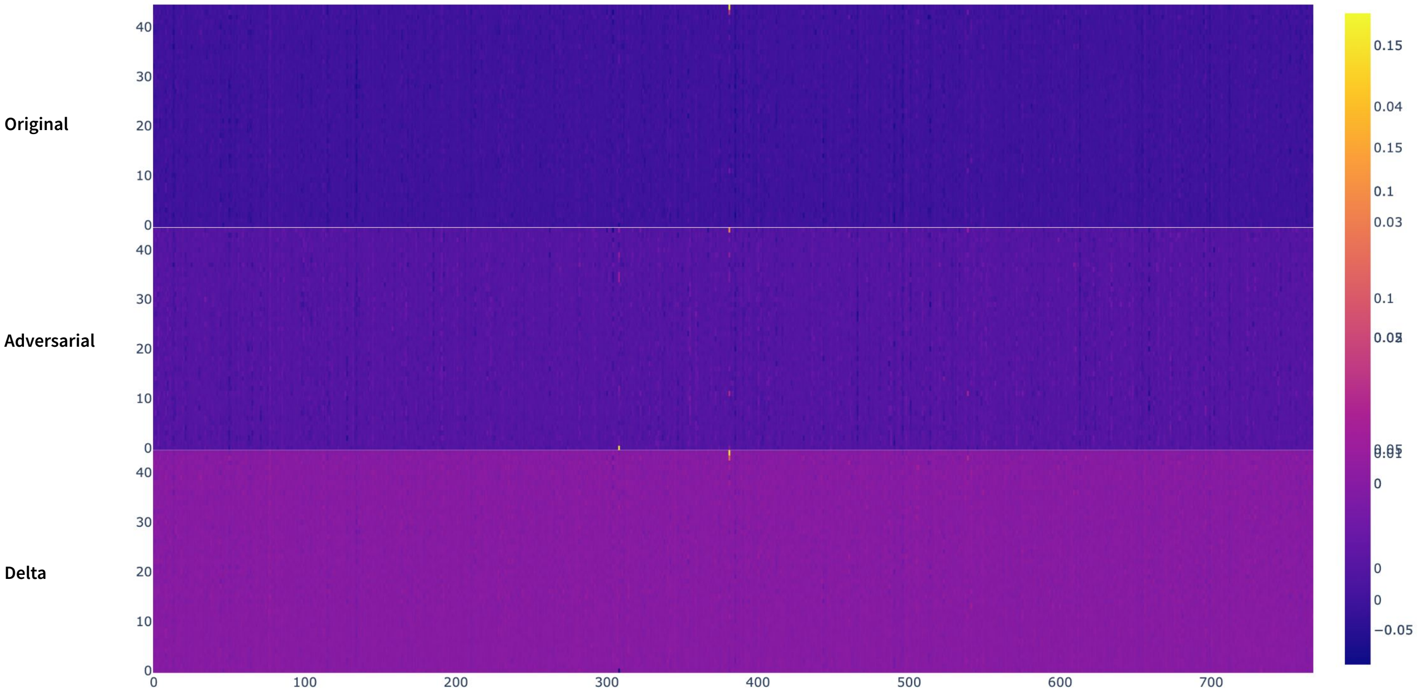
Layer integrated gradients represents the integral of gradients with respect to the layer inputs / outputs along the straight-line path from the layer activations at the given baseline to the layer activation at the input.

Experiment - Classifier Performance by BERT model



Original: Fed lifts rates a further quarter point By Andrew Balls in Washington and Jennifer Hughes in New York. The US Federal Reserve on Tuesday raised interest rates by a quarter point to 2.25 per cent and signalled there had been no change in its assessment of economic conditions.
Adversarial: course lifts grass a further quarter taper past Andrew Balls in Washington and Jennifer Hughes in New York. The US Federal hold on Tuesday raised interest grass by a quarter point to 2.25 per cent and signalled there had been no change in its assessment of economic check.

Experiment - Classifier Performance by BERT model



Original: Seven-wicket Kumble destroys Australia India #39;s spin king Anil Kumble grabbed seven wickets for 25 runs to skittle world champions Australia for 235 in a dramatic start to the second Test on Thursday.
Adversarial: Seven-wicket Kumble destroys Australia India #39;s spin mogul Anil Kumble grabbed seven hoop for 25 incline to skittle global booster Australia for 235 in a dramatic start to the endorsement quiz on Thursday.

Character Level: Attack

Replace/remove/add individual chars to change the classification of a sentence.

No attack: 89% Acc

1-char attack: 60% Acc [BERT on MRPC]

2-char attack: 31% Acc [BERT on MRPC]

[Pruthi et al. 2019 @ Carnegie Mellon, https://arxiv.org/pdf/1905.11268.pdf]

I would put this at the top of my list of films in the category of unwatchable trash!	Negative	Original
I would put this at the top of my list of films in the category of unWatchable trash!	Positive	Adversarial

Character Level: Defense

Most approaches are based on pre-processing. Achieve good results.

Neutral Word: 82.5% Acc [BERT on MRPC]

Spell Correction: 61.6% Acc [BERT on MRPC]

Pass-Through: 81.5% Acc [BERT on MRPC]

Background Model: 82.5% Acc [BERT on MRPC]

[Pruthi et al. 2019 @ Carnegie Mellon, https://arxiv.org/pdf/1905.11268.pdf]

I would put this at the top of my list of films in the category of unwatchable trash! — Negative — Original

I would put this at the top of my list of films in the category of unWatchable trash! — Positive — Adversarial

Pre-processing: Correct typos & grammar

I would put this at the top of my list of films in the category of unwatchable trash! — Processed

Classifier

Word Level: Attack

Replace/remove/add as little words as possible to change the classification of a sentence.

PWWS: 78.8% ASR [BERT, SST-2]
[Zeng et al. 2019 @ Tsingua, https://arxiv.org/pdf/2009.09191.pdf]

TextFooler: 90% ASR [BERT, SST-2]
[Zeng et al. 2019 @ Tsingua, https://arxiv.org/pdf/2009.09191.pdf]

BERT-ATTACK: 87% ASR [BERT, SST-2]
[Zeng et al. 2019 @ Tsingua, https://arxiv.org/pdf/2009.09191.pdf]

PWWS: 96.6% ASR [Word-CNN, IMDb]
[Wang et al. 2019 @ Huazhong, https://openreview.net/pdf?id=BJl_a2VYPH]

I would put this at the top of my list of films in the category of unwatchable trash!	Negative	Original
I would put this at the top of my list of films in the category of insatiable trash!	Positive	Adversarial

Word Level: Defense

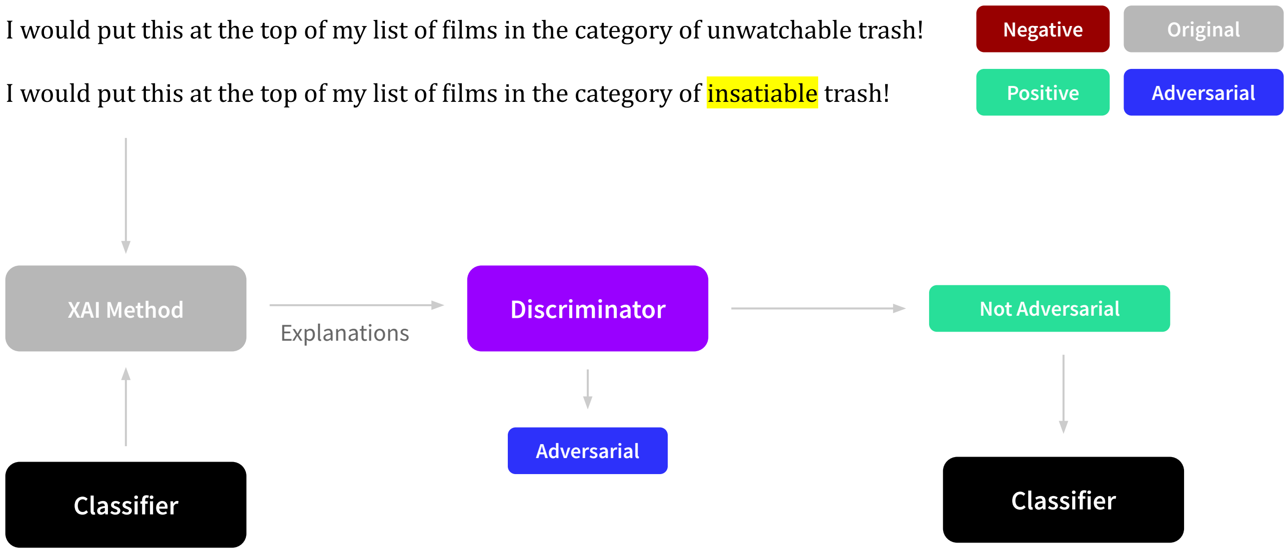
Currently most researched, approaches reach satisfying results.

DISP: 75.4 F1 [PWWS, IMDb]
[Zhou et al. 2019 @ UCLA, https://arxiv.org/abs/1909.03084]

FGWS: 89.5 F1 [PWWS, IMDb]
[Mozes et al. 2020 @ UCL, https://arxiv.org/abs/2004.05887]

SHAP: 90 F1 [PWWS, IMDb]
[Mosca et al. 2021 @ TUM]

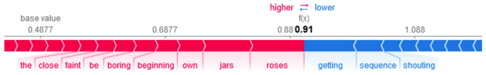
WDR: 92.1 F1 [PWWS, IMDb]
[Mosca et al. 2021 @ TUM, https://arxiv.org/abs/2204.04636]



Word Level: Defense

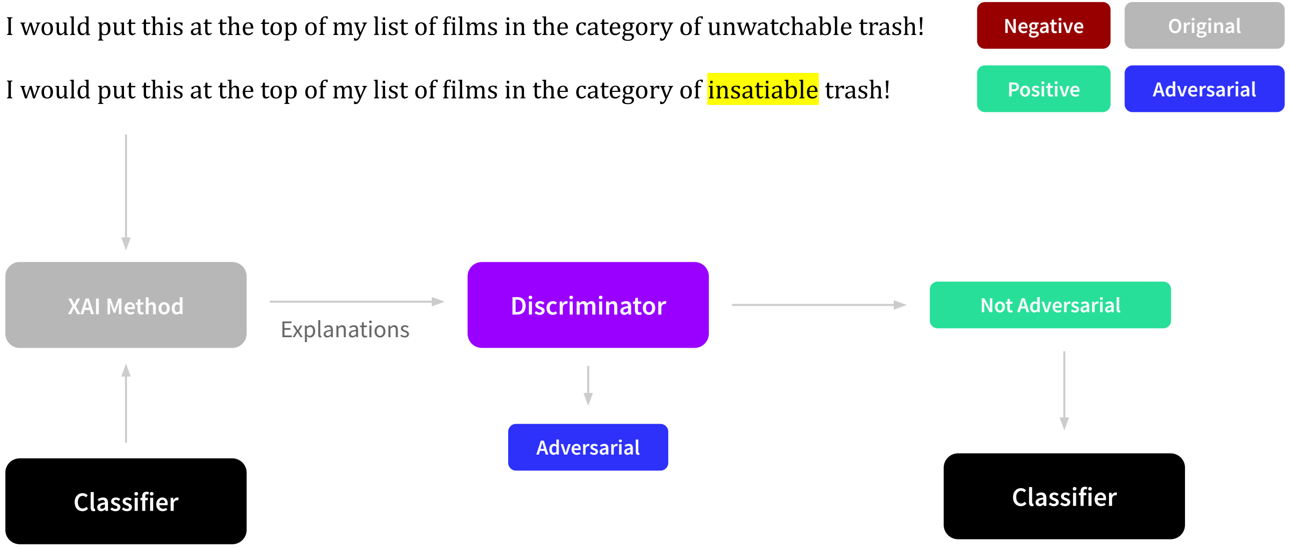
Currently most researched, approaches reach satisfying results.

SHAP



WDR

Replaced word		Logit class 0		Logit class 1		Logits difference	
∅	∅	**3.435**	-1.850	-3.456	**2.174**	6.891	4.024
worst	tough	**1.684**	**1.684**	-1.745	-1.745	3.43	-3.43
(arghhh!)	the	**3.343**	**1.378**	-3.4158	-1.374	6.76	-2.75
feature,	absolutely	**3.402**	-0.31	-3.453	**0.475**	6.86	0.79
This	I	**3.411**	-1.07	-3.474	**1.355**	6.89	2.43



Sentence Level: Attack

I would put this at the top of my list of films in the category of unwatchable trash!

Negative | Original

In my opinion, this is the first piece of work that falls into the category of trash that I can't stand to look at!

Positive | Adversarial

GAN: 26-47% ASR

[Zhao et al. 2018 @ UCI, https://arxiv.org/abs/1710.11342]

SCPN: 52-64% ASR

[Iyyer et al. 2018 @ Stanford, https://arxiv.org/abs/1804.06059]

StyleAdv: 91-96% ASR

[Qi et al. 2021 @ Tsinghua, https://arxiv.org/abs/2110.07139]

Dataset	Victim	BERT			ALBERT			DistilBERT		
	Attacker	ASR	PPL	GE	ASR	PPL	GE	ASR	PPL	GE
SST-2	GAN	26.42	4643.5	3.34	39.40	1321.7	9.26	47.53	752.3	3.93
	SCPN	52.84	553.2	3.20	59.98	432.9	3.43	64.73	479.0	3.29
	StyleAdv	**91.47**	**228.7**	**1.15**	**95.51**	**191.9**	**1.16**	**96.21**	**180.7**	**1.13**
HS	SCPN	6.56	**223.1**	3.37	7.56	358.2	4.10	1.36	652.8	3.38
	StyleAdv	**51.25**	263.3	**1.26**	**59.03**	**267.0**	**1.32**	**31.00**	**254.8**	**1.39**
AG's News	SCPN	32.98	343.7	4.51	30.91	261.8	4.39	51.04	294.7	5.26
	StyleAdv	**58.36**	**338.8**	**3.14**	**80.70**	**259.2**	**2.59**	**89.54**	**232.6**	**2.86**

[Qi et al. 2021, https://arxiv.org/abs/2110.07139]

Sentence Level: Defense

Defenses for sentence-level attacks are mostly unexplored.

I would put this at the top of my list of films in the category of unwatchable trash!	Negative	Original
In my opinion, this is the first piece of work that falls into the category of trash that I can't stand to look at!	Positive	Adversarial