

ADVERSARIAL EXAMPLES AGAINST A BERT ABSA MODEL

FOOLING BERT WITH L33T, MISSPELLIGN, AND PUNCTUATION,

**N. HOFER, P. SCHÖTTLE, A. RIETZLER, S. STABINGER
AUGUST, 2021**

Motivation

Adversarial Examples Against a BERT ABSA Model



Bidirectional Encoder Representations from Transformers - BERT

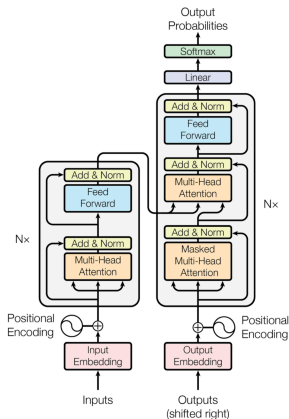


Abbildung: Transformer Model Architecture (Vaswani et al., 2017)

Motivation

Adversarial Examples Against a BERT ABSA Model

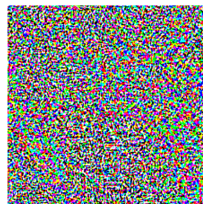


x

“panda”

57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

$=$



$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”

99.3 % confidence

Abbildung: Adversarial Examples in Computer Vision (Goodfellow et al, ICLR 2015)

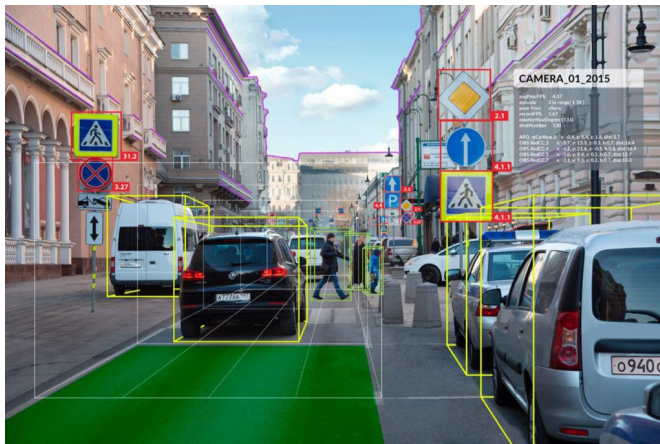


Abbildung: Object detection in autonomous driving (Source: becominghuman.ai)



n0r4
@n0r42



Covid-19 can be treated by gargling with salt water!



[Get the facts about COVID-19](#)

5:40 PM · Sep 1, 2020 · [Twitter Web App](#)

 View Tweet activity



Abbildung: Tweet containing misleading information regarding Covid-19, detected and labeled correctly



n0r4
@n0r42



Covid-19 can be treated by gargling with salt water!



Get the facts about COVID-19

5:40 PM · Sep 1, 2020 · [Twitter Web App](#)

||| [View Tweet activity](#)



Abbildung: Tweet containing misleading information regarding Covid-19, detected and labeled correctly



n0r4
@n0r42

C0v1d-19 can be treated by gargling with salt water!

5:42 PM · Sep 1, 2020 · [Twitter Web App](#)

 View Tweet activity



Abbildung: Tweet containing misleading information regarding Covid-19. Potential problems due to the use of Leet Speak.

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA

Aspect-based Sentiment Analysis

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA

Aspect-based Sentiment Analysis

Dataset

- Labels contain a set of **Entity - Attribute - Sentiment**
- 23 Entities - 9 Attributes - 3 Sentiments (POS, NEG, NEU)
- **Entity:** reviewed entity
- **Attribute:** particular attribute of an entity
- **Sentiment:** polarity towards the entity and its attribute

Entity Labels	
1. LAPTOP	13. BATTERY
2. DISPLAY	14. GRAPHICS
3. KEYBOARD	15. HARD DISK
4. MOUSE	16. MULTIMEDIA DEVICES
5. MOTHERBOARD	17. HARDWARE
6. CPU	18. SOFTWARE
7. FANS& COOLING	19. OS
8. PORTS	20. WARRANTY
9. MEMORY	21. SHIPPING
10. POWER SUPPLY	22. SUPPORT
11. OPTICAL DRIVES	23. COMPANY
Attribute Labels	
A. GENERAL	E. USABILITY
B. PRICE	F. DESIGN& FEATURES
C. QUALITY	G. PORTABILITY
D. OPERATION& PERFORMANCE	H. CONNECTIVITY
	I. MISCELLANEOUS

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA

Aspect-based Sentiment Analysis

The computer is excellent for gaming but I think it is way too expensive!!

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA

Aspect-based Sentiment Analysis

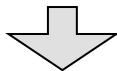
The computer is excellent for gaming but I think it is way too expensive !!

Aspect: Gaming, Sentiment: POS

Aspect: Price, Sentiment: NEG

Adversarial Attacks against BERT for ABSA

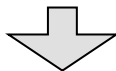
1. Fine-Tuning BERT base for ABSA



2. Identify Important Word

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA

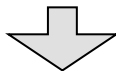


2. Identify Important Word

Leave-One-Out Method

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA



2. Identify Important Word

Leave-One-Out Method

The computer is excellent for gaming but I think it is way too expensive!! Gaming - POS; Price - NEG

***The** - computer is excellent for gaming but I think it is way too expensive!!* Gaming - POS; Price - NEG

***computer** - The is excellent for gaming but I think it is way too expensive!!* Gaming - POS; Price - NEG

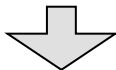
***is** - The computer excellent for gaming but I think it is way too expensive!!* Gaming - POS; Price - NEG

***excellent** - The computer is for gaming but I think it is way too expensive!!* Price - NEG

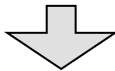
***for** - The computer is excellent gaming but I think it is way too expensive!!* Gaming - POS; Price - NEG

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA



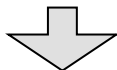
2. Identify Important Word



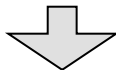
3. Modification of Important Words

Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA



2. Identify Important Word



3. Modification of Important Words

Leetspeak

Misspellings

Punctuation

Objectives

- Semantic Meaning
- Inconspicuousness
- Relevance

Adversarial Attacks against BERT for ABSA

1. Leetspeak

The computer is excellent for gaming but I think it is way too expensive!!

Aspect: Gaming, Sentiment: POS

Aspect: Price, Sentiment: NEG

*Original important word: **excellent***

*Modified important word: **exce11ent***

The computer is exce11ent for gaming but I think it is way too expensive!!

Aspect: Gaming, **Sentiment: NEG**

Aspect: Price, Sentiment: NEG

Adversarial Attacks against BERT for ABSA

2. Misspellings

The computer is excellent for gaming but I think it is way too expensive!!

Aspect: Gaming, Sentiment: POS

Aspect: Price, Sentiment: NEG

*Original important word: **excellent***

*Modified important word: **ecxcellent***

The computer is ecxcellent for gaming but I think it is way too expensive!!

Aspect: Price, Sentiment: NEG

Adversarial Attacks against BERT for ABSA

3. Punctuation

The computer is excellent for gaming but I think it is way too expensive!!

Aspect: Gaming, Sentiment: POS

Aspect: Price, Sentiment: NEG

*Original important word: **excellent***

*Modified important word: **excellent,***

The computer is excellent, for gaming but I think it is way too expensive!!

Aspect: Laptop (general), Sentiment: NEG

Aspect: Gaming, Sentiment: NEG

Aspect: Price, Sentiment: NEG

Perturbation Method	Leetspeak	Misspellings	Punctuation
Dataset A - # of original sentences	943	943	943
Dataset B - # of modifiable original sentences	897	369	943
Dataset C - # of adversarial sentences	2232	1354	2555
Dataset D - # of changed predictions total	1066	420	382
Dataset E - # of changed predictions per sentence	790	259	253
Overall Success Rate	47.76%	31.01%	14.95%
Distinct Success Rate	88.07%	70.19%	26.83%

Tabelle: Comparison of the success rates of the three attack methods.

Summary

- BERT can be fooled by input modifications
- Three attack methods:
 - Leetspeak
 - Misspellings
 - Falsly placed Punctuation

Next Steps

- Transferability between Transformer Models
- Using generated adversarial datasets for Adversarial Training

ADVERSARIAL EXAMPLES AGAINST A BERT ABSA MODEL

FOOLING BERT WITH L33T, MISSPELLIGN, AND PUNCTUATION,

N. HOFER, P. SCHÖTTLE, A. RIETZLER, S. STABINGER
AUGUST, 2021