

# **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

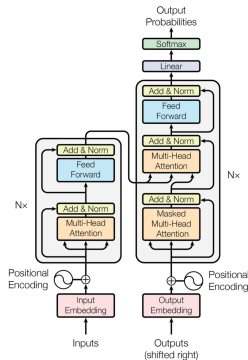


Abbildung: Transformer Model Architecture (Vaswani et al., 2017)



# Motivation

## Adversarial Examples Against a BERT ABSA Model

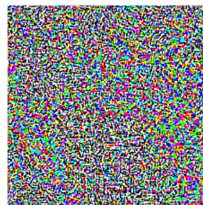


$x$

“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)

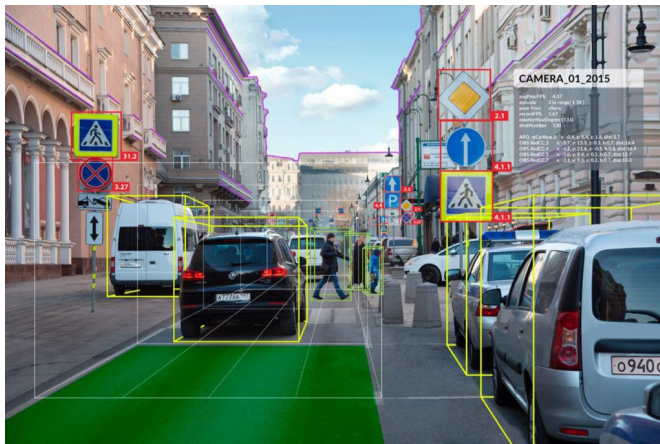


Abbildung: Object detection in autonomous driving (Source: becominghuman.ai)



**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.

*“We’re starting to deploy machine learning as a technology that can fail so we need to have some checks in place.” Nicolas Papernot, 2017*



## 1. Fine-Tuning BERT base for ABSA

# 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

*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

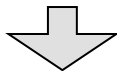
*The computer is excellent for gaming but I think it is way too expensive !!*

Aspect: Gaming, Sentiment: POS

Aspect: Price, Sentiment: NEG

# Adversarial Attacks against BERT for ABSA

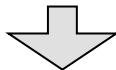
1. Fine-Tuning BERT base for ABSA



2. Identify Important Word

# Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA

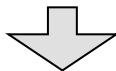


2. Identify Important Word

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

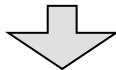


# Adversarial Attacks against BERT for ABSA

1. Fine-Tuning BERT base for ABSA



2. Identify Important Word



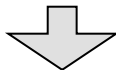
3. Modification of Important Words

# 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

## Qualitative Results



tba

*Our reliance on deep learning based language models for real-world (security-relevant) applications is questionable.*



# ADVERSARIAL EXAMPLES AGAINST A BERT ABSA MODEL

FOOLING BERT WITH L33T, MISSPELLIGN, AND PUNCTUATION,

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