Adversarial Examples Against a BERT ABSA Model

Fooling Bert With L33T, Misspellign, and Punctuation,

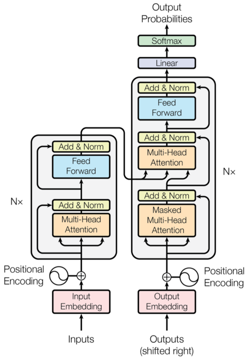
N. Hofer, P. Schöttle, A. Rietzler, S. Stabinger

August, 2021

Motivation

Adversarial Examples Against a BERT ABSA Model

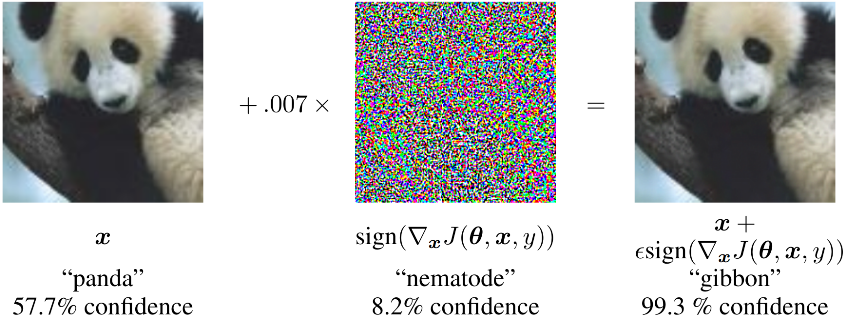
Bidirectional Encoder Representations from Transformers - BERT



Transformer Model Architecture (Vaswani et al., 2017)

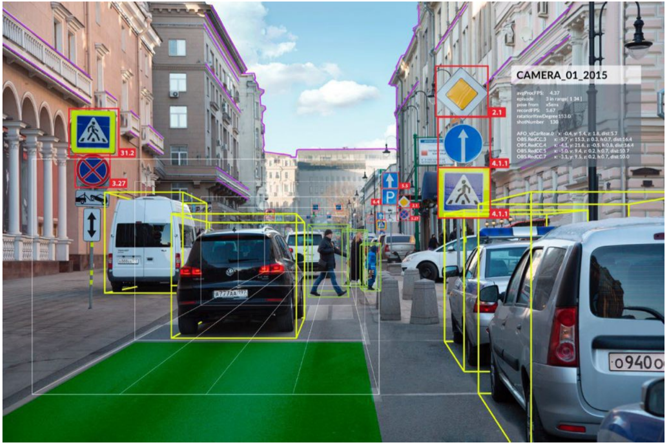
Motivation

Adversarial Examples Against a BERT ABSA Model


$\boldsymbol{x}$
"panda"
57.7% confidence
$+ .007 \times$
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$=$
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

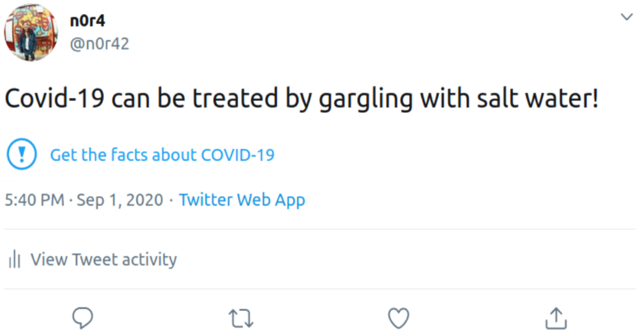
Adversarial Examples in Computer Vision (Goodfellow et al, ICLR 2015)

Motivation


31.2
3.27
CAMERA_01_2015
avgProcFPS: 4.37
episode 3 in range[1 34]
pose from xSens
recordFPS: 5.67
ratationYawDegree 153.0
shotNumber 130
AFO_vjCarRear.0 x: -0.4, y: 5.4, z: 1.8, dist:5.7
OBS.RedCC.3 x: -5.7, y: 15.3, z: 0.3, h:0.7, dist:16.4
OBS.RedCC.4 x: -4.1, y: 21.6, z: -0.5, h:0.8, dist:16.4
OBS.RedCC.5 x: -5.0, y: 9.4, z: 0.2, h:0.7, dist:10.7
OBS.RedCC.7 x: -3.1, y: 9.5, z: 0.2, h:0.7, dist:10.0
2.1
4.1.1
4.1.1

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

Motivation



n0r4
@n0r42

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

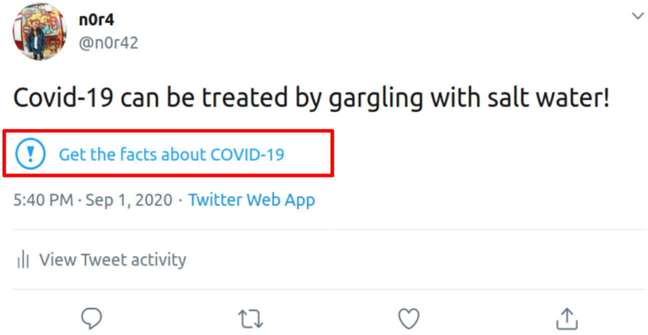
Get the facts about COVID-19

5:40 PM · Sep 1, 2020 · Twitter Web App

View Tweet activity

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

Motivation



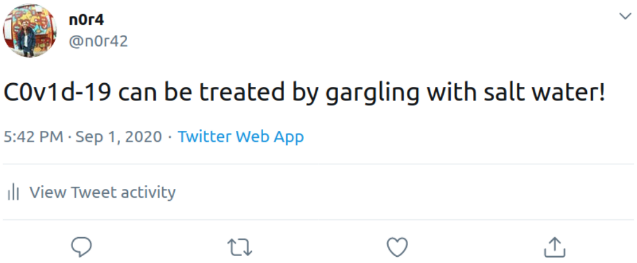
n0r4
@n0r42

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

5:40 PM · Sep 1, 2020 · Twitter Web App

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

Motivation



n0r4
@n0r42

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

5:42 PM · Sep 1, 2020 · Twitter Web App

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

Adversarial Attacks

“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

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

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

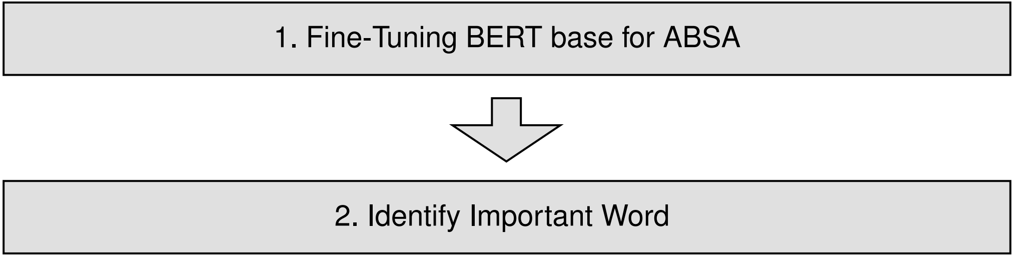
1. Fine-Tuning BERT base for ABSA

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

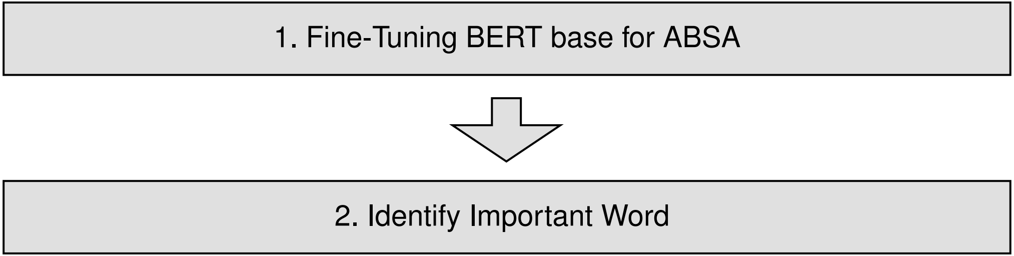
Adversarial Attacks against BERT for ABSA



1. Fine-Tuning BERT base for ABSA

2. Identify Important Word

Adversarial Attacks against BERT for ABSA

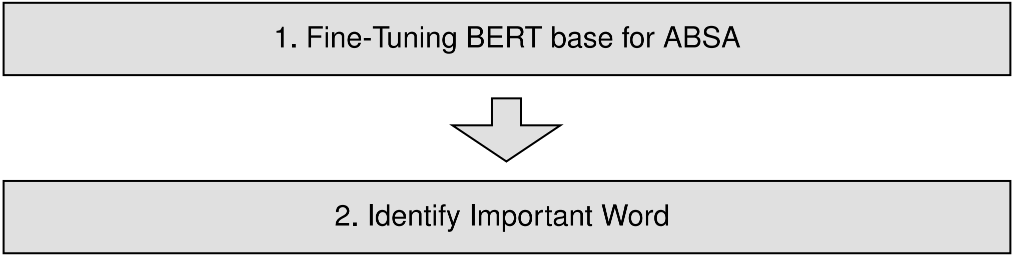


1. Fine-Tuning BERT base for ABSA

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

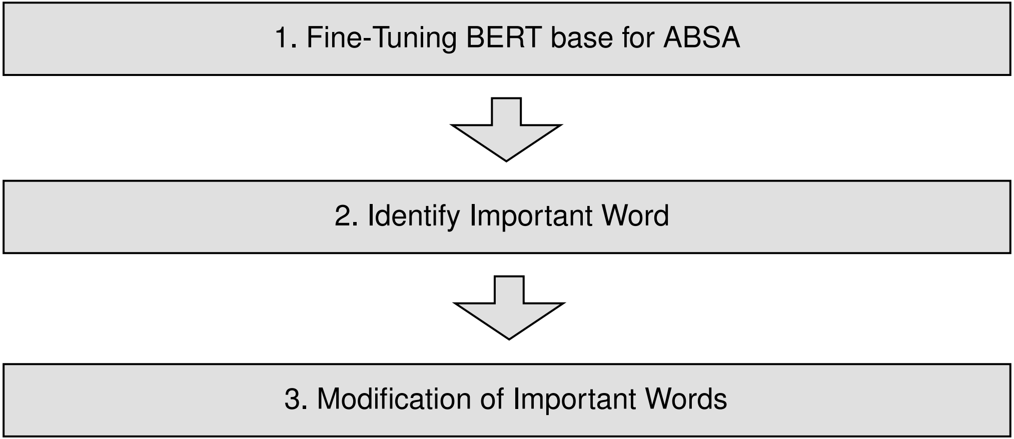
computer - *The is excellent for gaming but I think it is way too expensive!!* Gaming - POS; Price - NEG

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

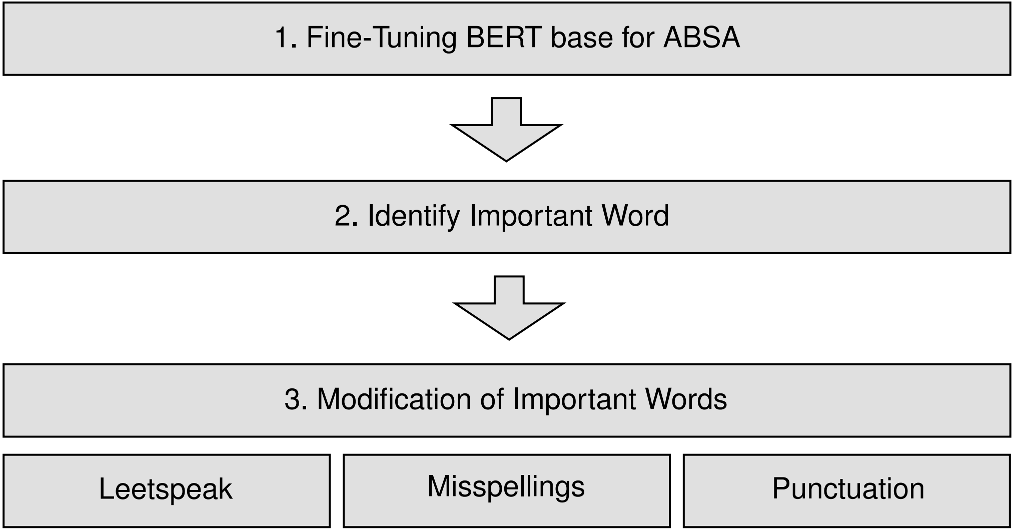


1. Fine-Tuning BERT base for ABSA

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

Adversarial Attacks

Objectives

- Semantic Meaning
- Inconspicousness
- 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: ***ecxellent***

The computer is ecxellent 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

Conclusion

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