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

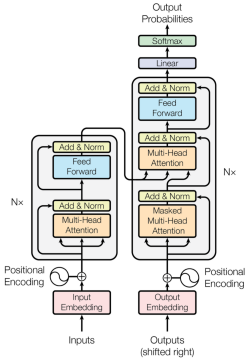


Abbildung: Transformer Model Architecture (Vaswani et al., 2017)

Motivation

Adversarial Examples Against a BERT ABSA Model

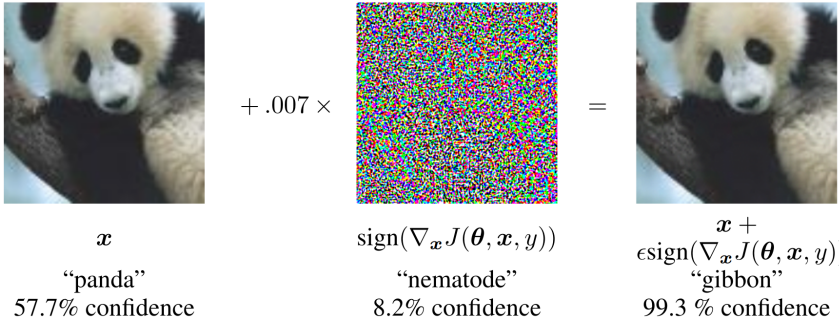


Abbildung: Adversarial Examples in Computer Vision (Goodfellow et al, ICLR 2015)

Motivation

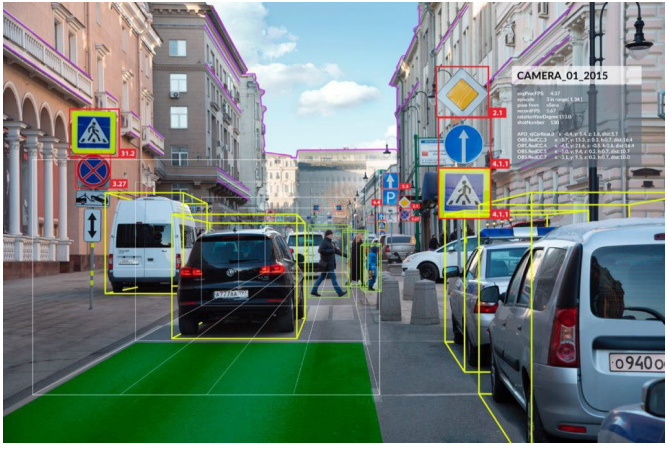


Abbildung: Object detection in autonomous driving (Source: becominghuman.ai)

Motivation

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

Motivation

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

Motivation

Abbildung: 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-Tunint BERT base for ABSA

Adversarial Attacks

bsp block

bla bla blagfdsfds
bla bla

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

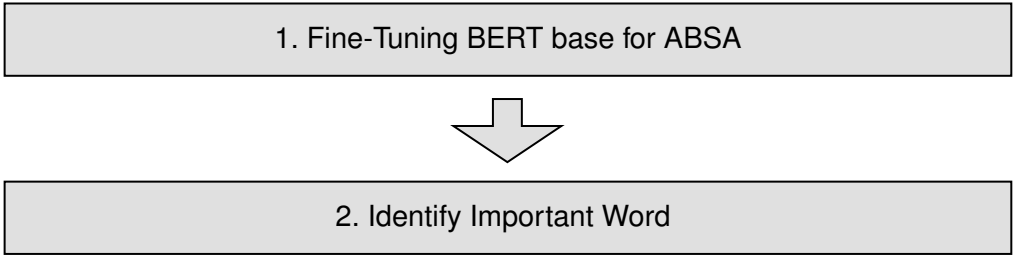
Aspect-based Sentiment Analysis

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

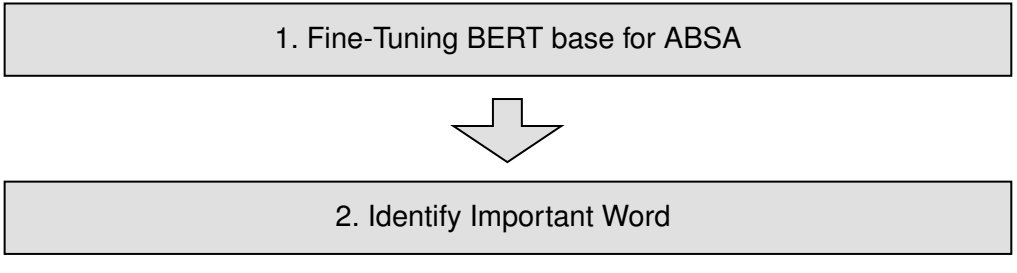
Aspect: Gaming, Sentiment: POS

Aspect: Price, Sentiment: NEG

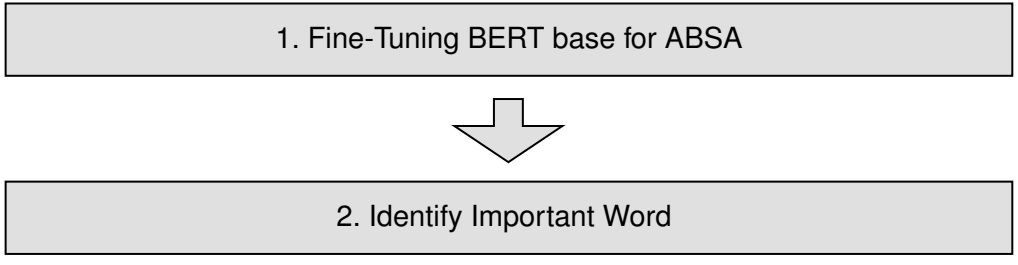
Adversarial Attacks against BERT for ABSA



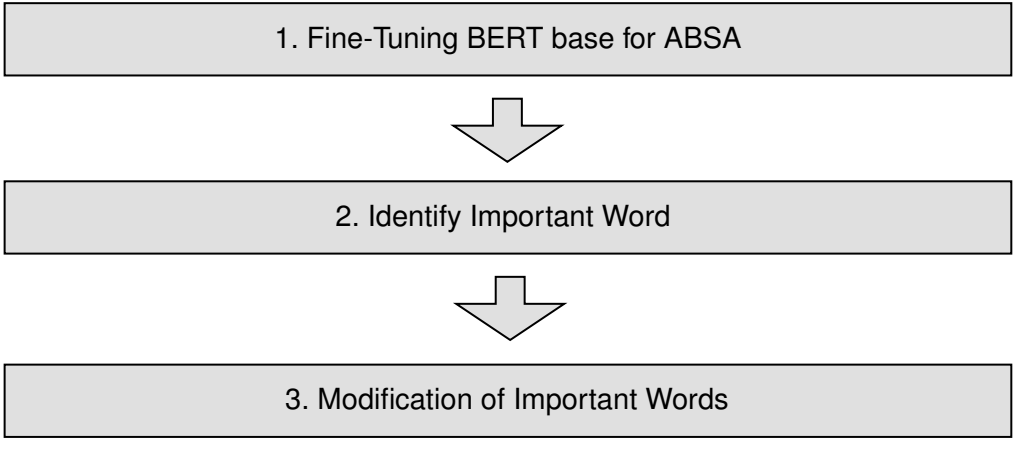
Adversarial Attacks against BERT for ABSA



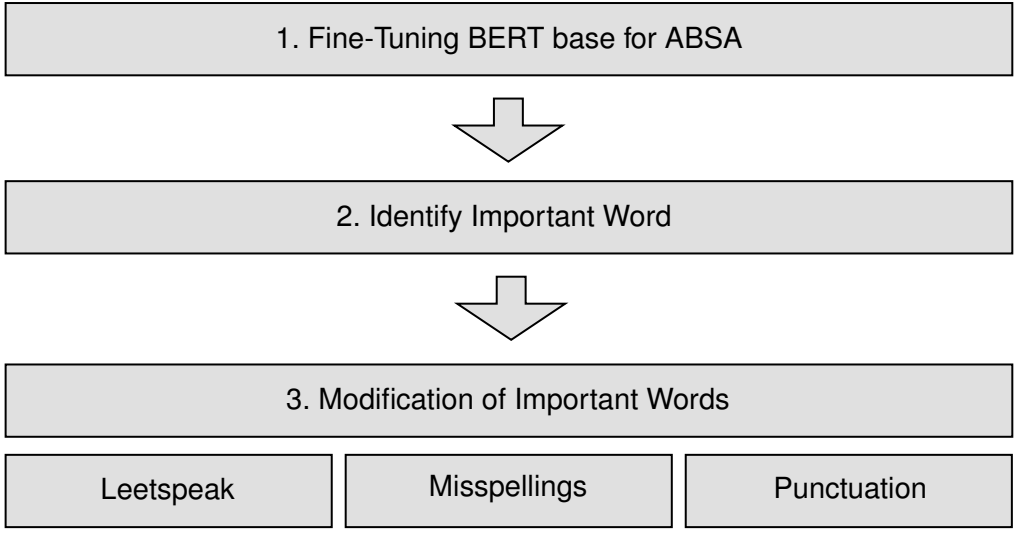
Adversarial Attacks against BERT for ABSA



Adversarial Attacks against BERT for ABSA



Adversarial Attacks against BERT for ABSA



Objectives

- Semantic Meaning
- Inconspicousness
- Relevance

Adversarial Attacks against BERT for ABSA

1. Leetspeak

Adversarial Attacks against BERT for ABSA

2. Misspellings

Adversarial Attacks against BERT for ABSA

3. PUnctuation

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, Misspelligh, and Punctuation,

N. Hofer, P. Schöttle, A. Rietzler, S. Stabinger
August, 2021