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 Machine Learning

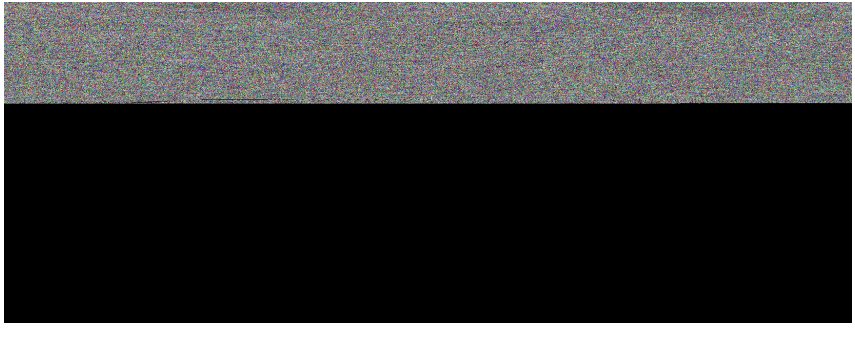


Figure: Adversarial Examples in Computer Vision (Goodfellow et al, ICLR 2015)

Motivation

Figure: Tweet containing misleading information regarding COVID-19.

Motivation

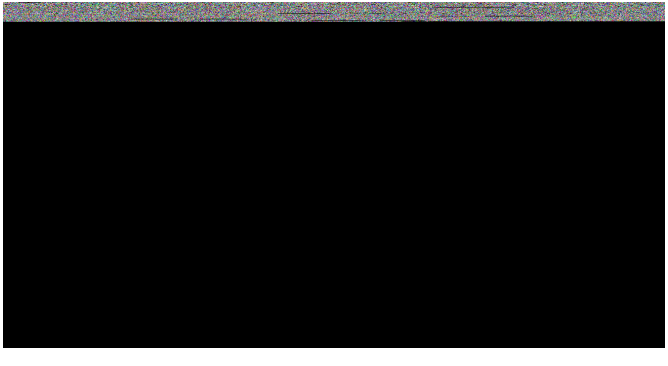


Figure: Tweet containing misleading information regarding COVID-19, detected and labeled correctly.

Motivation

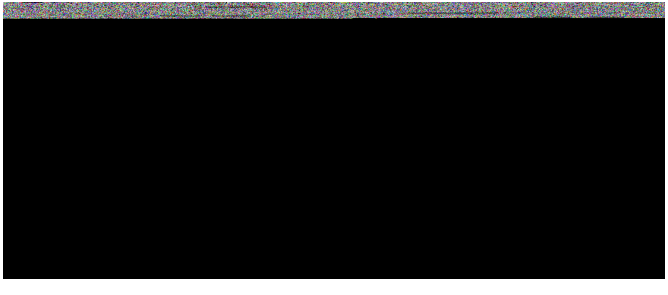


Figure: Tweet containing misleading information regarding COVID-19, undetected due to the use of leetspeak.

Experimental Setup

Fine-Tuning BERT base for ABSA

Experimental Setup

Fine-Tuning BERT base for ABSA

Aspect-**b**ased **S**entiment **A**nalysis

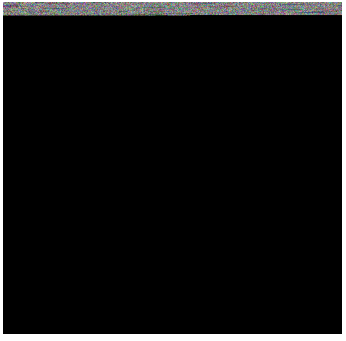
Experimental Setup

Fine-Tuning BERT base for ABSA

Aspect-**b**ased **S**entiment **A**nalysis

Dataset: SemEval-2015 Task 12

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



Experimental Setup

Fine-Tuning BERT base for ABSA

Aspect-**b**ased **S**entiment **A**nalysis

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

Experimental Setup

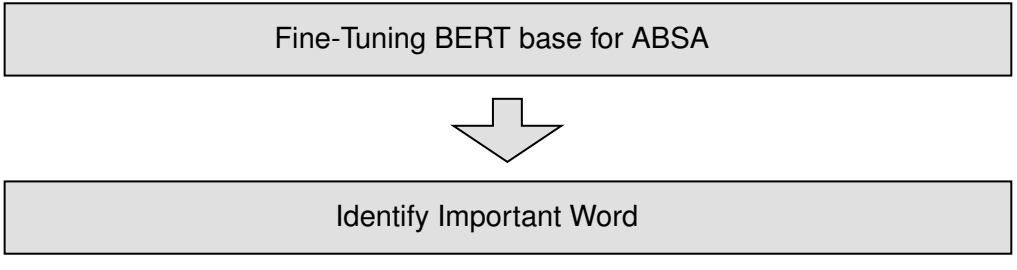
Fine-Tuning BERT base for ABSA

Aspect-based Sentiment Analysis

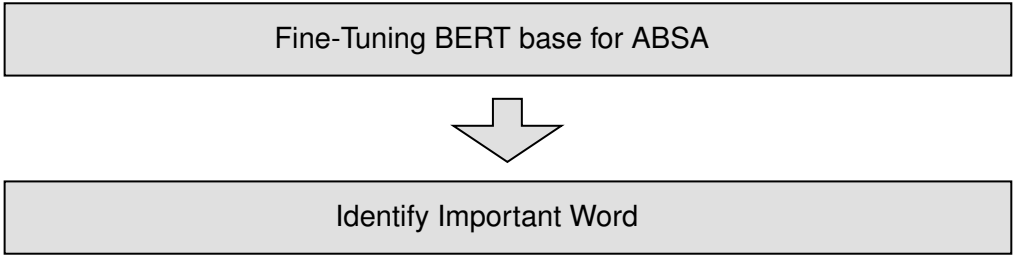
The computer is excellent for gaming but I think it is way too expensive !!

Aspect: Gaming, Sentiment: POS
Aspect: Price, Sentiment: NEG

Experimental Setup

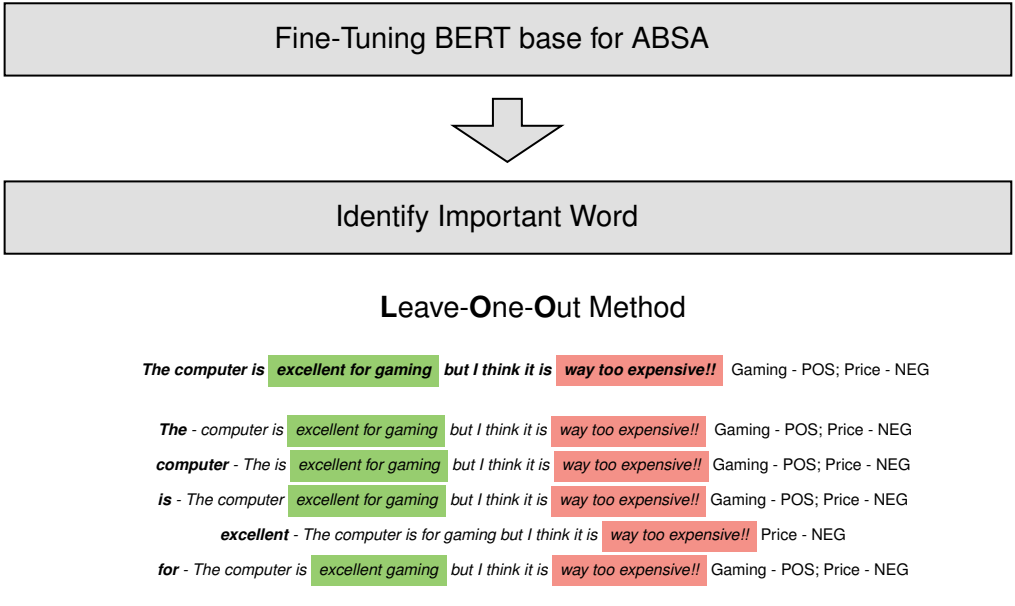


Experimental Setup

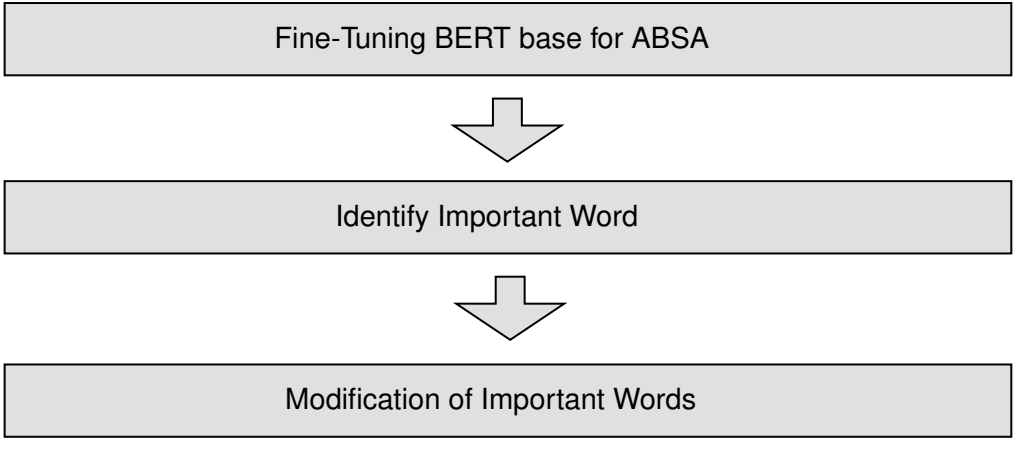


Leave-**O**ne-**O**ut Method

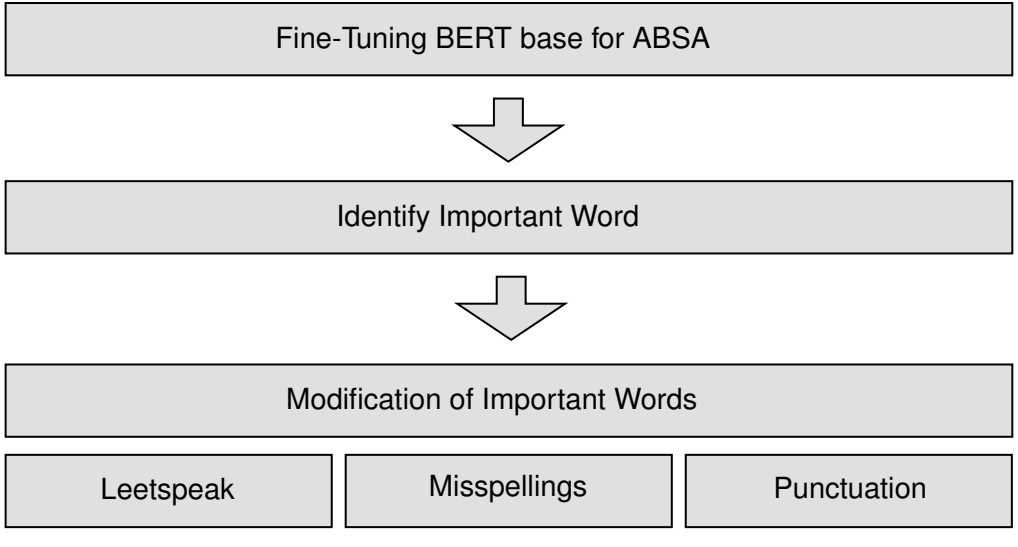
Experimental Setup



Proposed Attacks



Proposed Attacks



Design Criteria

Objectives

- Keeping the **semantic meaning** of the input data
- **Inconspicousness** to a human observer
- Relevance in a **real-world scenario**

Attack Methods

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

Attack Methods

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

Attack Methods

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

Results

Perturbation Method	Leetspeak	Misspellings	Punctuation
Dataset A - # of original sentences	943	943	943

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

Results

Perturbation Method	Leetspeak	Misspellings	Punctuation
Dataset A - # of original sentences	943	943	943
Dataset B - # of modifiable original sentences	897	369	943

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

Results

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

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

Results

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

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

Conclusion & Further Steps

Summary

- BERT can be fooled by input modofications
- Three attack methods:
  - Leetspeak
  - Misspellings
  - Misplaced Punctuation

Next Steps

- **Transferability** between Transformer Models
- Using generated adversarial datasets for **adversarial training**

Thank you!

Adversarial Examples Against A BERT ABSA Model -

Fooling BERT with L 33T , Misspelli gn , and Punctuation ,

GitHub: https://github.com/NoraH2004/adv-absa

Email: nora.hofer@uibk.ac.at

Adversarial Examples Against a BERT ABSA Model

Fooling Bert With L33T, Misspellign, and Punctuation,

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