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

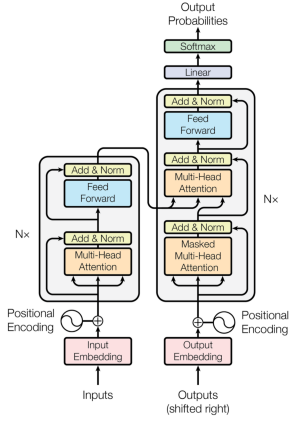


Figure: Transformer Model Architecture (Vaswani et al., 2017)

Motivation

Adversarial Examples Against a BERT ABSA Model

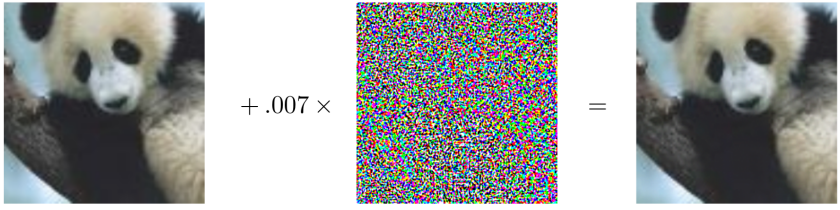


Figure: Adversarial Examples in Computer Vision (Goodfellow et al, ICLR 2015)

Motivation

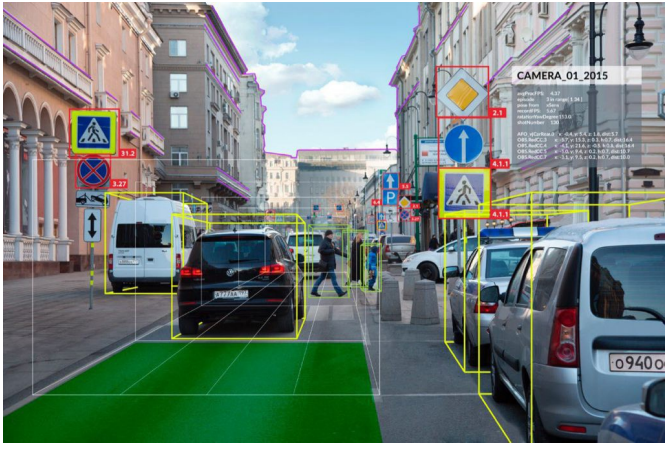


Figure: Object detection in autonomous driving (Source: becominghuman.ai)

Motivation

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

Motivation

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

Motivation

Figure: 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: 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

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

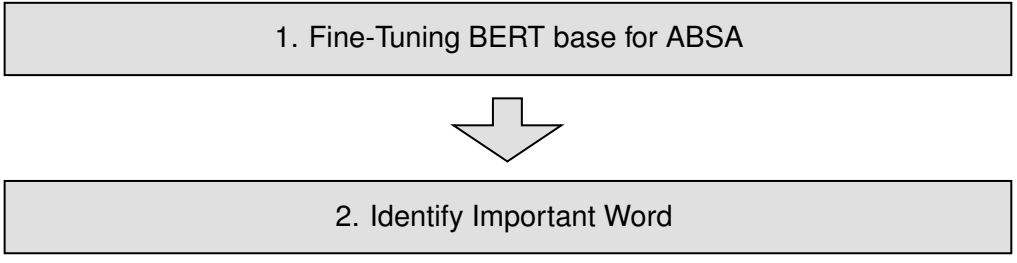
1. Fine-Tuning BERT base for ABSA

Aspect-based Sentiment Analysis

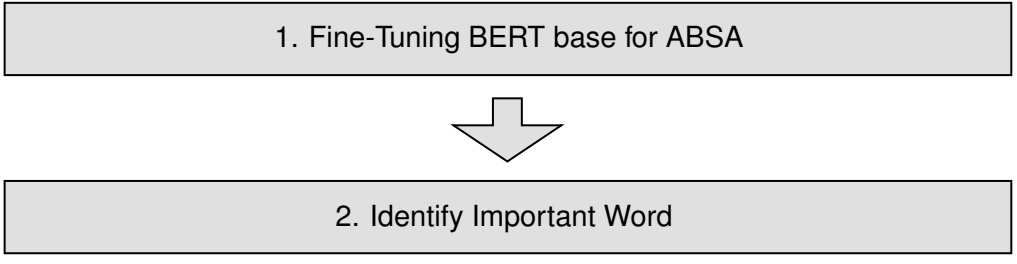
The computer is excellent for gaming but I think it is way too expensive !!

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

Adversarial Attacks against BERT for ABSA

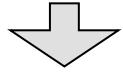


Adversarial Attacks against BERT for ABSA



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

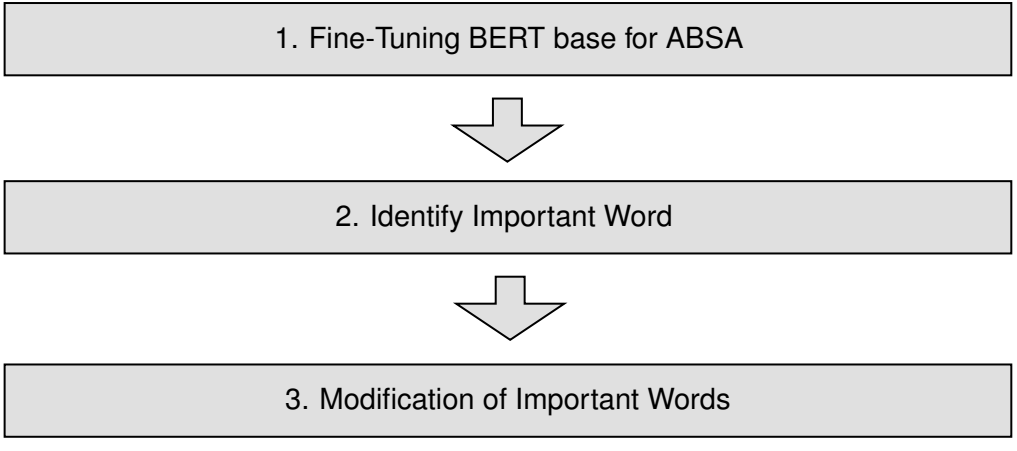
computer *- The is* *excellent for gaming* *but I think it is* *way too expensive!!* Gaming - POS; Price - NEG

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

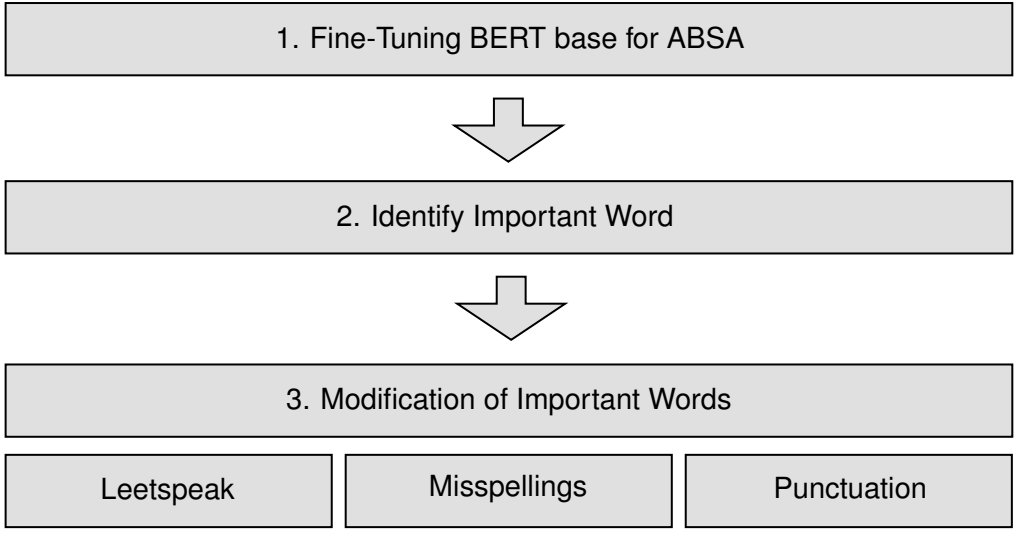
excellent *- The computer is for gaming but I think it is* *way too expensive!!* Price - NEG

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

Adversarial Attacks against BERT for ABSA



Adversarial Attacks against BERT for ABSA



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

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

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

Qualitative 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.

Qualitative 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.

Qualitative 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.

Qualitative 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
Overall Success Rate	**47.76%**	**31.01%**	**14.95%**
Distinct Success Rate	**88.07%**	**70.19%**	**26.83%**

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
  - Falsly placed 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, Misspelligh, and Punctuation,

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