CS779 Project

Adversarial Techniques In NLP

Dept. of CSE, IIT Kanpur

Mentors

Aishwarya
Ajita Shree

Guide

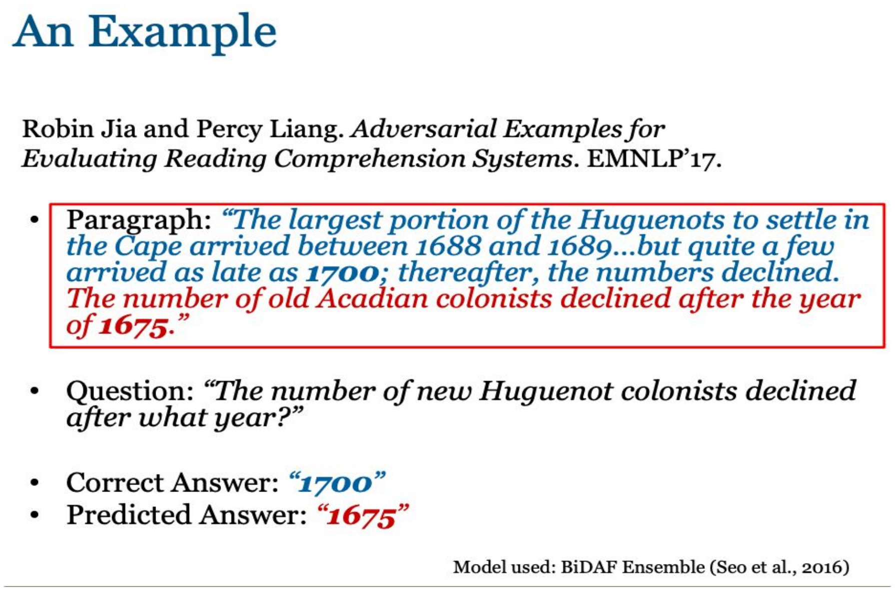
Dr. Ashutosh Modi
Assistant Professor
Dept. of CSE

Group 3:

Abhishek Krishna- 20111002
Deeksha Arora- 20111017
Preeti Singh- 20111044
Sambrant Maurya- 20111054
Shruti Sharma- 20111061

Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey

- *WEI EMMA ZHANG, QUAN Z. SHENG, and AHOUD ALHAZMI, CHENLIANG LI*



An Example

Robin Jia and Percy Liang. *Adversarial Examples for Evaluating Reading Comprehension Systems*. EMNLP'17.

- Paragraph: *"The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as **1700**; thereafter, the numbers declined. The number of old Acadian colonists declined after the year of **1675**."*
- Question: *"The number of new Huguenot colonists declined after what year?"*
- Correct Answer: ***"1700"***
- Predicted Answer: ***"1675"***

Model used: BiDAF Ensemble (Seo et al., 2016)

Adversarial examples are the strategically modified samples, with infinitesimally small perturbations that may fool the model to give false predictions.

Example point: x' = x + € (noise),
s.t. f(x) = y,
but, f(x') = y' ≠ y, € is small perturbation picked from a perturbation set S & added to **original input example x.**

Terminology

- **Adversarial Attack (Evasion Attack)**
  - ❖ Method for generating adversarial examples that deviate correct prediction to incorrect/pre-specified one so that model's performance degrades.
  - ❖ It is of 2 types :
    - ➢ Targeted -> f(x') = y'
    - ➢ Untargeted -> f(x') != y

- **Adversarial Training**
  - ❖ Process to introduce adv egs in model to improve its generalization & robustness towards worst-case examples.
  - ❖ Ways of adversarial training:
    - ➢ Re-train a model using adversarial examples.
    - ➢ Incorporate input perturbations as part of the model training process.

- ***But how exactly to incorporate perturbations?***
  - ❖ Visual Similarity
  - ❖ Semantic Similarity

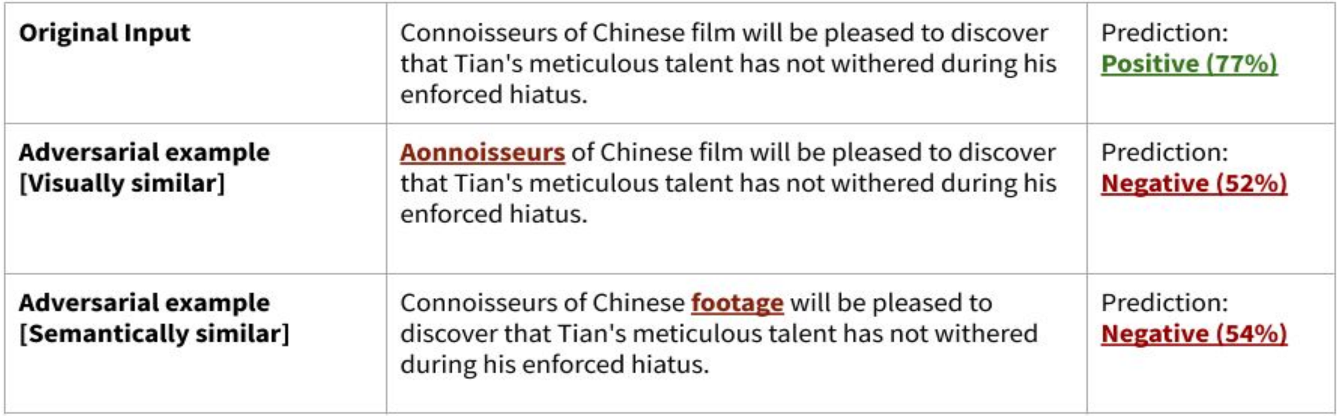
Ref: https://medium.com/analytics-vidhya/everything-you-need-to-know-about-adversarial-training-in-nlp-b249301b6229

Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey

- *WEI EMMA ZHANG, QUAN Z. SHENG, and AHOUD ALHAZMI, CHENLIANG LI*

- Preparing for Attack?

- ❖ Black-box *vs* White-box
- ❖ Un-targeted *vs* Targeted
- ❖ Granularity - use of word, character, sentence or subword level embeddings. ~ *Attack Position*
- ❖ Attack (evaluate robustness of DNN) vs Defense (robustify DNN)

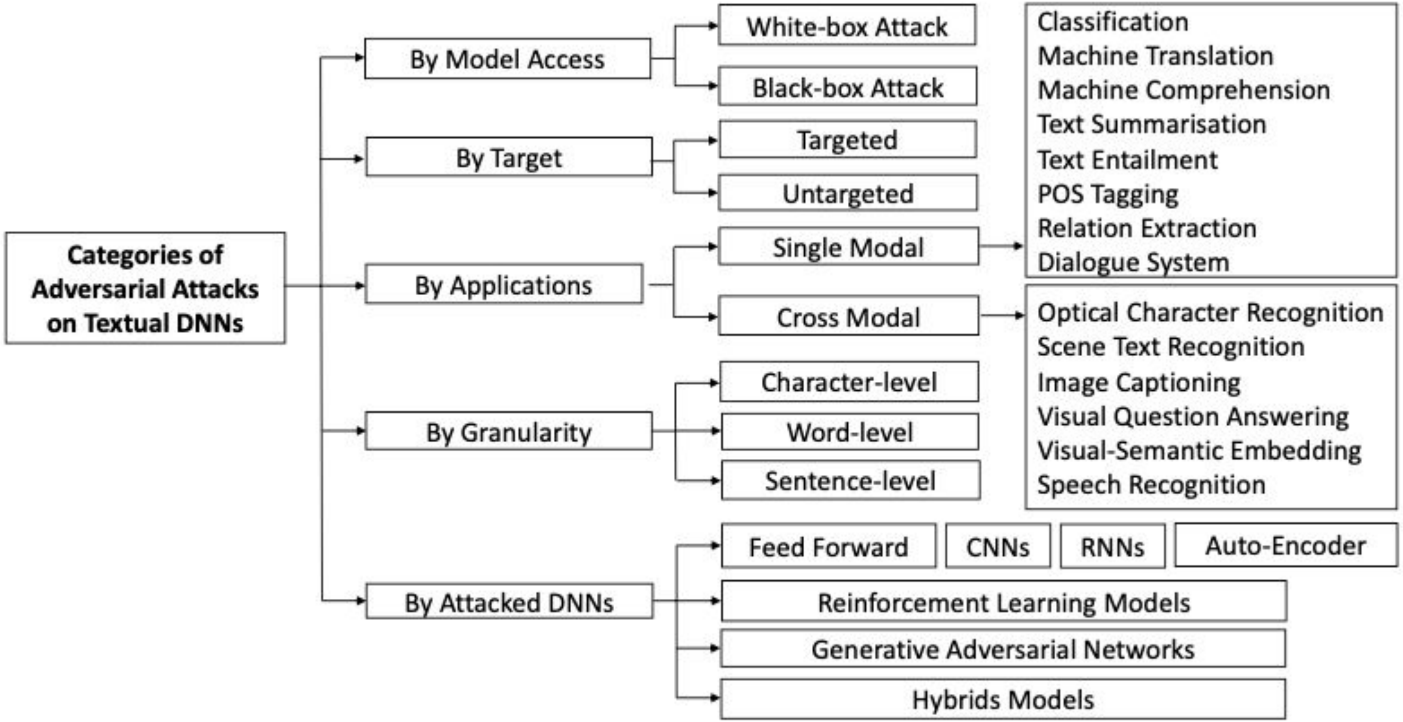


Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Positive (77%)**
Adversarial example [Visually similar]	**Aonnoisseurs** of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (52%)**
Adversarial example [Semantically similar]	Connoisseurs of Chinese **footage** will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (54%)**

Ref - https://towardsdatascience.com/what-are-adversarial-examples-in-nlp-f928c574478e

Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey

- WEI EMMA ZHANG, QUAN Z. SHENG, and AHOUD ALHAZMI, CHENLIANG LI



Ref - Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey
https://dl.acm.org/doi/fullHtml/10.1145/3374217

BERT-ATTACK: Adversarial Attack Against BERT Using BERT

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, Xipeng Qiu. EMNLP 2020

- Fine tuned BERT models are powerful on downstream tasks.
  - Adversarial attacks are challenging *(Jin et al., 2019)*
  - Generated examples should be fluent and semantically consistent.

What if we use BERT against BERT?

- Using masked language models for generating perturbations:
  - Find perturbations that maximise the risk of making wrong predictions.*
  - BERT - a pre trained masked language model
  - Perturbations consider the context around
    - Fluent and reasonable
  - Can be used to generate better substitutions

- **BERT Attack:**
  - Uses BERT as a language model
  - Attacker uses BERT to attack another BERT model

* Goodfellow et al., 2014

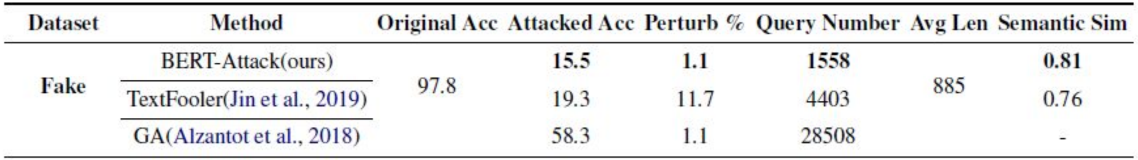
Paper: https://www.aclweb.org/anthology/2020.emnlp-main.500.pdf

BERT-ATTACK: Adversarial Attack Against BERT Using BERT

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, Xipeng Qiu. EMNLP 2020

- Straightforward algorithm:
  - Find the vulnerable words
  - Apply BERT!

- **Experimental results:**
  - Tested on datasets - Fake, IMDB, Yelp, AG news, MNLI and SNLI



Dataset	Method	Original Acc	Attacked Acc	Perturb %	Query Number	Avg Len	Semantic Sim
Fake	BERT-Attack(ours)	97.8	**15.5**	**1.1**	**1558**	885	**0.81**
	TextFooler(Jin et al., 2019)		19.3	11.7	4403		0.76
	GA(Alzantot et al., 2018)		58.3	1.1	28508		-

  - Fools SOTA BERT models successfully
  - Lower perturb percentage and query number
  - Semantic preservation is high
  - 3 times faster than *Textfooler**
  - Can be used for attacking other models as well
    - LSTM based models

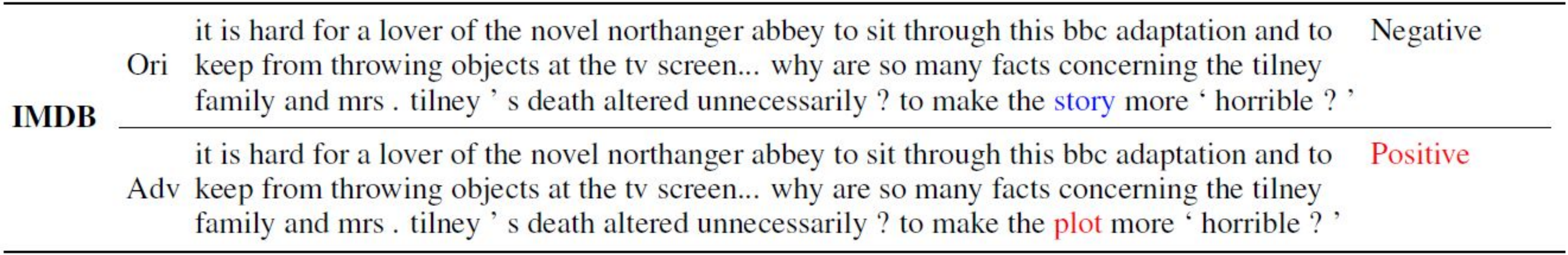
Image:
https://www.aclweb.org/anthology/2020.emnlp-main.500.pdf

*Textfooler(Jin et al., 2019)

BERT-ATTACK: Adversarial Attack Against BERT Using BERT

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, Xipeng Qiu. EMNLP 2020

- An example:



IMDB	Ori	it is hard for a lover of the novel northanger abbey to sit through this bbc adaptation and to keep from throwing objects at the tv screen... why are so many facts concerning the tilney family and mrs . tilney ’ s death altered unnecessarily ? to make the story more ‘ horrible ? ’	Negative
	Adv	it is hard for a lover of the novel northanger abbey to sit through this bbc adaptation and to keep from throwing objects at the tv screen... why are so many facts concerning the tilney family and mrs . tilney ’ s death altered unnecessarily ? to make the plot more ‘ horrible ? ’	Positive

- Future work:
  - BERT is rapidly becoming mainstream
    - Google search
    - Chatbots
    - VideoBERT, PatentBERT, DocBERT etc
  - Attacks against BERT based chatbots can be tried

Framework: https://github.com/LinyangLee/BERT-Attack

Image: https://www.aclweb.org/anthology/2020.emnlp-main.500.pdf

MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models

Le, Suhang Wang, Dongwon Lee. ICDM 2020.

- A malicious comment generation framework to fool fake news detection models.

- SOTA fake news detector models:
  - Use features of an article to predict its credibility
  - Exploit user engagement

- Existing attacks on SOTA fake news detectors
  - Careful manipulation of features of the article
  - Hiding questionable content of an article or altering its source

- Limitations of these attacks
  - Difficult to exercise post publish attacks
  - Adversarial texts generated by the attackers are detectable by naked eye
  - User engagement remains unexplored

Paper: https://arxiv.org/pdf/2009.01048.pdf

Framework: https://github.com/lethaiq/MALCOM

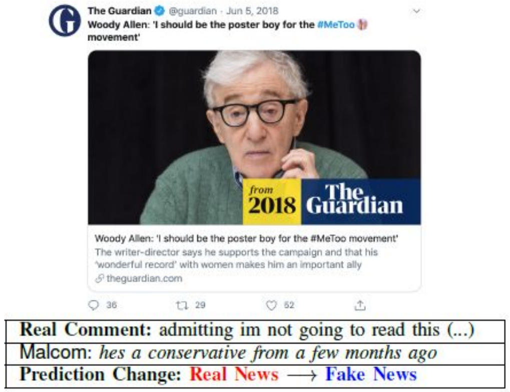
MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models

Le, Suhang Wang, Dongwon Lee. ICDM 2020.

- What’s a better approach?
  - Many SOTA models use users’ comments as a feature
  - Target user comments!

- **Datasets used:**
  - **GOSSIPCOP:** dataset of real and fake news collected from GossipCop website.
  - **PHEME:** dataset of rumors and non-rumors.

- **An example:**



The Guardian @guardian · Jun 5, 2018

Woody Allen: 'I should be the poster boy for the #MeToo movement'

Woody Allen: 'I should be the poster boy for the #MeToo movement'

The writer-director says he supports the campaign and that his 'wonderful record' with women makes him an important ally

theguardian.com

36 29 52

Real Comment: admitting im not going to read this (...)
Malcom: *hes a conservative from a few months ago*
Prediction Change: Real News ⟶ Fake News

Ref: https://arxiv.org/pdf/2009.01048.pdf

Image: https://arxiv.org/pdf/2009.01048.pdf

MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models

Le, Suhang Wang, Dongwon Lee. ICDM 2020.

- **MALCOM Framework:**
  - **Conditional Comment Generator (G):** malicious comment generator- a conditional sequential text generator model
  - **Style Module:** used to improve quality of generated adversarial comments.
    - Topic relevant
    - Coherent
  - **Attack Module:** fine-tunes G to improve the success attack rate.

- **Performance:**
  - Successfully misleads latest NN models to output targeted real/fake labels
  - Fools black box models with a success rate of 90%
  - Robust even in the presence of a defence system
  - Capable of not only promoting fake news but also demoting real news

Ref: https://arxiv.org/pdf/2009.01048.pdf

HotFlip: White-Box Adversarial Examples for Text Classification

Javid Ebrahimi, Anyi Rao, Daniel Lowd, Dejing Dou. ACL 2018.

- White-box adversarial attack that tricks character-level and word-level neural models.
- Character-level : Atomic flip using gradient based approach
- Goal function: Untargeted classification
- Search algorithm: Beam search
- Word-level : derivatives with respect to one-hot vectors + semantics-preserving constraints
  - Constraints: cosine similarity, same part-of-speech, disallow replacing of stop-words.

*Results***:**

- 34% of examples are misclassified after one change using the derivative-based approach.
- Given the strict set of constraints, the authors were able to create only 41 examples for word-level attack.

Ref: https://www.aclweb.org/anthology/P18-2006/

HotFlip: White-Box Adversarial Examples for Text Classification

Javid Ebrahimi, Anyi Rao, Daniel Lowd, Dejing Dou. ACL 2018.

- ❖ *Generate adversarial examples with character “flips”.*

  Given one hot representation of the input, a character flip in the j-th character of the i-th word (a → b) is represented as:

$$\vec{v}_{ijb} = (\vec{0},..;(\vec{0},..(0,..\overset{a}{\overset{\downarrow}{-1}},0,..,\overset{b}{\overset{\downarrow}{1}},0)_j,..\vec{0})_i;\vec{0},..)$$

“aid”=”bid”

- ❖ *Maximize first order approximation of change in loss to find best character swap.*
- ❖ *After trying all possible single flip using beam search, keep top flips that have highest loss and continue for r steps.*

Future work: Flips can be extended to insertion and deletion.Also adding targeted attacks to HotFlip

Ref: https://www.aclweb.org/anthology/P18-2006.pdf

TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, Yanjun Qi

- TextAttack:
  - Python framework
  - Benchmarking and comparing NLP attacks on different models and datasets
  - Framework: Goal Function, Set of constraints, Transformation, Search Method.

- Goal Function: Untargeted/Targeted Classification

- Constraints: Stopwords Modification, Min. Word length, Grammatically, Semantics.

- Transformations: Word swap with character transform/gradient based, word deletion etc.

- Search Method: Greedy search with word importance ranking, Genetic Algorithm.

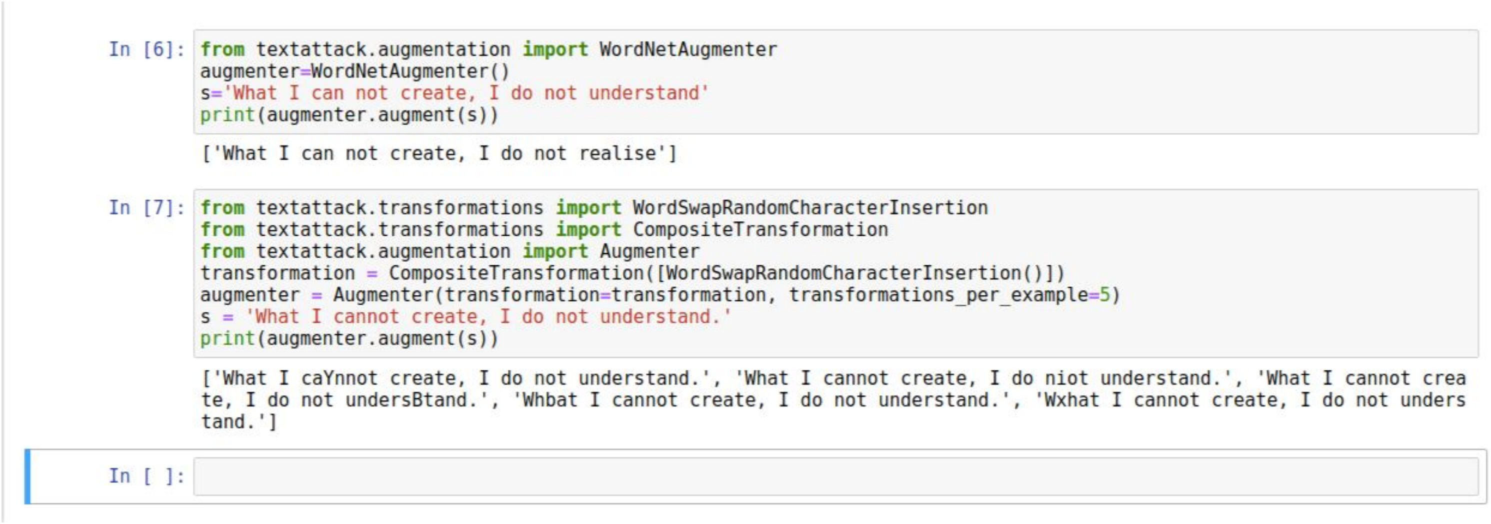
Paper Link: https://www.aclweb.org/anthology/2020.emnlp-demos.16.pdf, Framework: https://github.com/QData/TextAttack

Augmenting Text Example:

textattack.WordNetAugmenter: replacing words with WordNet synonyms

textattack.EmbeddingAugmenter: replacing words with neighbors s.t. Cosine Similarity is at least 0.8

textattack.CharSwapAugmenter: augments text by substituting, deleting, inserting, and swapping adjacent characters

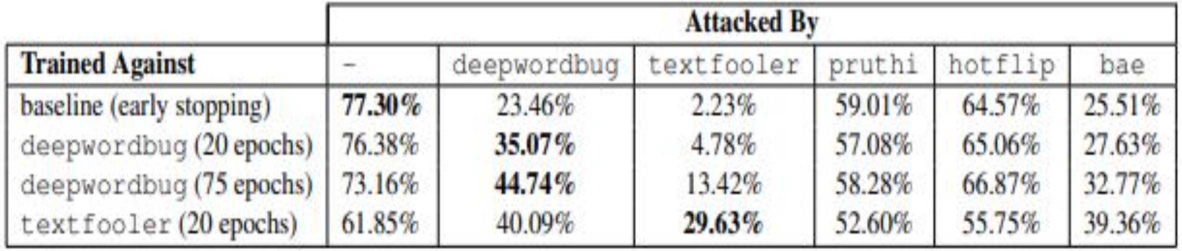


```
In [6]: from textattack.augmentation import WordNetAugmenter
        augmenter=WordNetAugmenter()
        s='What I can not create, I do not understand'
        print(augmenter.augment(s))

        ['What I can not create, I do not realise']

In [7]: from textattack.transformations import WordSwapRandomCharacterInsertion
        from textattack.transformations import CompositeTransformation
        from textattack.augmentation import Augmenter
        transformation = CompositeTransformation([WordSwapRandomCharacterInsertion()])
        augmenter = Augmenter(transformation=transformation, transformations_per_example=5)
        s = 'What I cannot create, I do not understand.'
        print(augmenter.augment(s))

        ['What I caYnnot create, I do not understand.', 'What I cannot create, I do niot understand.', 'What I cannot crea
        te, I do not undersBtand.', 'Whbat I cannot create, I do not understand.', 'Wxhat I cannot create, I do not unders
        tand.']

In [ ]:
```

Ref: https://pypi.org/project/textattack/0.0.3.1/

Pros:

- Able to benchmark and compare different models and datasets.
- Attacks, Data Augmentation, Adversarial Training all are there.
- Model Agnostic.
- Caching/Memoization.
- HuggingFace support



	Attacked By					
Trained Against	-	deepwordbug	textfooler	pruthi	hotflip	bae
baseline (early stopping)	**77.30%**	23.46%	2.23%	59.01%	64.57%	25.51%
deepwordbug (20 epochs)	76.38%	**35.07%**	4.78%	57.08%	65.06%	27.63%
deepwordbug (75 epochs)	73.16%	**44.74%**	13.42%	58.28%	66.87%	32.77%
textfooler (20 epochs)	61.85%	40.09%	**29.63%**	52.60%	55.75%	39.36%

- Dataset: Stanford Sentiment Treebank(SST2)
- Model trained on Standard LSTM using 3K samples.
- Accuracy under attack on eval set is reported.

Future Direction:

- Combining different transformations, search methods and evaluating performance

Ref: https://arxiv.org/pdf/2010.01724.pdf

Summary

Paper	Strategy	Granularity	Attacked Model
Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey	Comparing all the work done on adversarial techniques in NLP	Character, word, subword, sentence, PE embeddings, medical features	CNN, LSTM, OpenNMT, Ensembles
BERT-ATTACK	Use BERT to generate fluent and semantically consistent word replacements	Word	BERT, LSTM
MALCOM	Generates malicious comments to attack the fake news detection models	Sentence	NN
HotFlip: White-Box Adversarial Examples for Text Classification	Use beam search to get the best flip which will maximize the loss.	Character	CharCNN-LSTM
TextAttack: A Framework	Benchmarking and comparing NLP attacks on different models and dataset	Character, word	LSTM, CNN, Bert

Report’s Rough Sketch

- Introduction
    - Problem Statement & Terminology
- Literature Review/Related Work
    - Classification of attacks based on white box/black box
    - Classification of attacks based on targeted/untargeted attacks
    - Classification based on multi-modal/unimodal attacks
    - Attacks on some specific domains (Malware Detection, Sentence/Text Classification, Machine Comprehension, Medical Records etc.)
    - Attacks on Character-level/Word-level/Sentence level embeddings
    - Attacks on DNNs (feed-forward, CNN, RNN, GAN etc.)
    - Attack vs Defence
- Proposed methodology and possible improvements to enrich robustness in NLP models.
- Conclusion and Future Direction
- References

Questions

- NLP Tasks - Text Summarisation, Sentiment Analysis, Machine Translation, Text Classification.
  - To study attacks on all or choose a specific task?
- Analyse existing frameworks like Python’s TextAttack, DeepRobust or implement and test them on new datasets?
- The final expectation from project - a survey report! (?)

Contribution:

- Shruti Sharma (20111061)
  - Introductory Slides
  - Paper- Adv. Attacks on DNN in NLP
- Sambrant Maurya (20111054)
  - Paper- Bert Attack
  - Report’s Rough Sketch
- Preeti Singh (20111044)
  - Paper- HotFlip Attack
  - Report’s Rough Sketch
- Deeksha Arora (20111017)
  - Paper- Malcom Attack
  - Report’s Rough Sketch
- Abhishek Krishna (20111002)
  - Paper- TextAttack
  - Report’s Rough Sketch

Thank You!