Alma Mater Studiorum – University of Bologna
Department of Computer Science and Engineering
Master's Degree in Artificial Intelligence

Master Thesis in Natural Language Processing

SYNBA: A CONTEXTUALIZED SYNONYM-BASED ADVERSARIAL ATTACK FOR TEXT CLASSIFICATION

Supervisor: Prof. Paolo Torroni

Co-supervisors: Dr. Federico Ruggeri
Dr. Giulia De Poli

Candidate: Giuseppe Murro

Academic Year 2021-2022 - Third session

Introduction

Adversarial Machine Learning

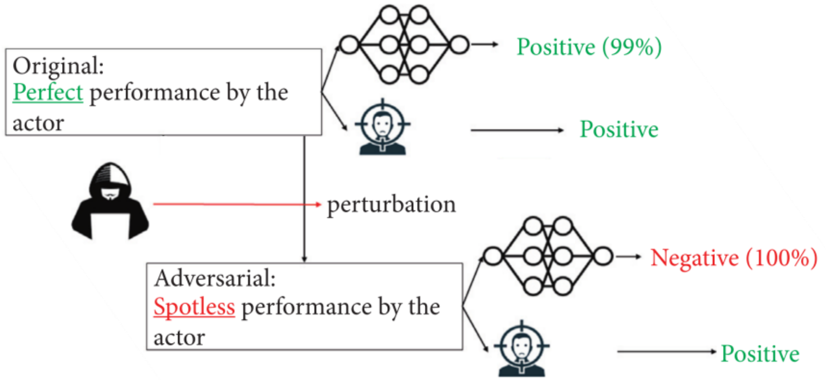
- Intersection of Machine Learning and Cybersecurity
- Aims to trick ML models by providing deceptive input
- Two kinds of Artificial Intelligence Attacks
  - **Data poisoning**
  - **Evasion** or **Adversarial Attack**

Successful textual adversarial example

It is a carefully designed instance with small **perturbations** that cause a machine learning model to make a false prediction, while theinstance is still correctly classified by a human observer

Paradigm shift: from CV to NLP

- Adversarial examples originated in the Computer Vision community
- It is hard to guarantee that the **quality** of the generated textual adversarial examples is **high**
- The **robustness** of ML models can be improved by including high-quality adversarial examples in the training set



Adversarial attacks on Text Classification

Attacks from literature

- They can be classified according to the **semantic granularity** of the perturbation
- Researchers proposed special attacks in the text domain to maintain **semantic consistency** and **syntactic correctness**
- In practice, they frequently **violate** linguistic **constraints**

GRANULARITY	ATTACK METHODS	ADVERSARIAL EXAMPLES GENERATED
CHARACTER-LEVEL	**DeepWordBug** Introduce typos for some words	*Original (POS)*: This film has a special place in my heart *Perturbed (NEG)*: This film has a special p1ace in my herat
WORD-LEVEL	**TextFooler** Find synonyms exploting cosine similarity between word embeddings	*Original (POS)*: generates an enormous feeling of empathy for its characters *Perturbed (NEG)*: leeds an enormous foreboding of empathy for its fonts
	BERT-Attack (BAE) Use BERT as MLM for word replacement	*Original (NEG)*: bears is even worse than i imagined a movie ever could be. *Perturbed (POS)*: bears is even greater than i imagined a movie ever could be.
SENTENCE-LEVEL	**SCPNs** Use an encoder-decoder model to produce a paraphrase of the original sentence	*Original (NEG)*: I'd have to say the director is the big problem here *Perturbed (POS)*: By the way, you know, the director is the big problem

Proposed solution

SynBA

- **TextFooler** and **BERT-Attack** suffer respectively from a lack of context and semantic similarity
- We tried to combine their strengths in a new recipe called **SynBA** (Synonym-Based adversarial Attack)

Transformation

- For each selected word, replacement candidates are obtained from a **weighted function**
- The **rank** of candidates for word replacement is computed by summing up three **normalized scores**
- The weights were defined after a **hyperparameter tuning** phase

$$\text{SynBA-Score} = \lambda_1 * \text{MLM-Score} + \lambda_2 * \text{Thesaurus-Score} + \lambda_3 * \text{WordEmb-Score}$$

The confidence of candidates obtained by **MLM** (BERT) in a descending order

WordNet is used to retrieve **synonyms** and **antonyms** of the original word

The **cosine similarity** is computed between the reference words and the top closest embeddings

Proposed solution

Constraints

- Constraints are used to **avoid** the generation of adversarial examples that are **too different** from the original input text

    - **Part-of-speech** - constraints perturbations to only swap words with the same part of speech
    - **Max modification rate** - limits the number of words that can be perturbed in the input text to a maximum of 20% of total words
    - **Word embedding distance** - throws away perturbations by words with a cosine similarity lower than 0.6
    - **Semantic Textual Similarity** - checks whether the similarity between the original and the perturbed text is higher than a threshold t = 0.7 using a **Sentence-BERT** pre-trained model

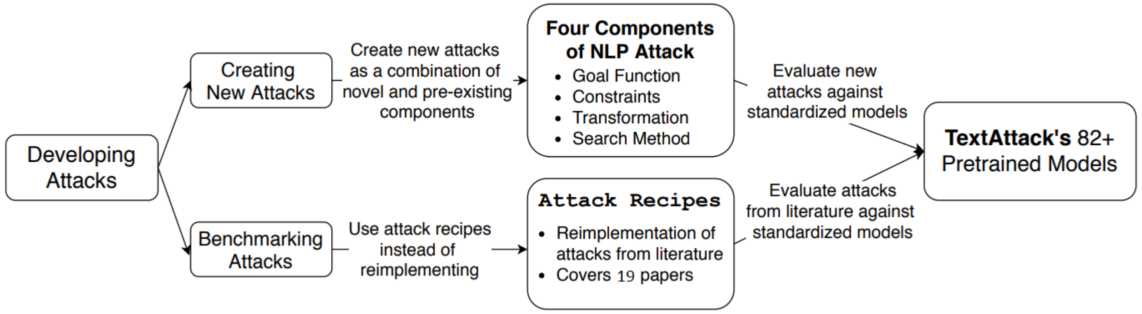
Empirical evaluation

TextAttack 🐙

- Python framework that provides implementations for 19 adversarial attacks methods from the literature
- Directly integrated with 🤗 HuggingFace's transformers

Components

- **goal function**, determines whether the attack is successful in terms of the model outputs
- **constraints**, determine if a perturbation is valid
- **transformation**, given an input, generates a set of potential perturbations
- **search method**, selects promising perturbations from a set of transformations



Experimental results

Datasets

- Sampled 1000 examples from two **movie review** datasets
- **IMDB** – avg. 229 words (± 162) per example
- **Rotten Tomatoes** – avg. 19 words (±9) per example

Target models

- **BERT-base** uncased model fine-tuned according to the dataset used as input

Quality assessment on IMDB

Metric	TextFooler	BERT-Attack	SynBA
Attack success rate (↑)	**98.39**	65.24	92.7
Original/perturbed perplexity (↓)	41.48/ 63.0	**41.78/ 48.4**	41.5/ 50.74
Sentence-BERT similarity (↑)	0.928	**0.964**	0.944
Contradiction rate (↓)	0.053	0.164	**0.049**

Quality assessment on Rotten Tomatoes

Metric	TextFooler	BERT-Attack	SynBA
Attack success rate (↑)	**89.44**	61.92	68.56
Original/perturbed perplexity (↓)	72.58/ 154.52	**76.96/ 99.91**	72.05/ 112.08
Sentence-BERT similarity (↑)	0.805	0.776	**0.901**
Contradiction rate (↓)	0.196	0.536	**0.123**

Human evaluation

Human prediction consistency assessment

- Three **human annotators** were asked to evaluate 100 successful adversarial examples for each attack method
- The task is to **decide** if the perturbed sample is **consistent** with the original one

Annotation interface

```
[3]  annotations = annotate(
       attack_samples,
       options=['CONSISTENT', 'INCONSISTENT', 'UNCLEAR'],
       display_fn=lambda string: display(HTML(string))
     )
```

0 examples annotated, 100 examples left

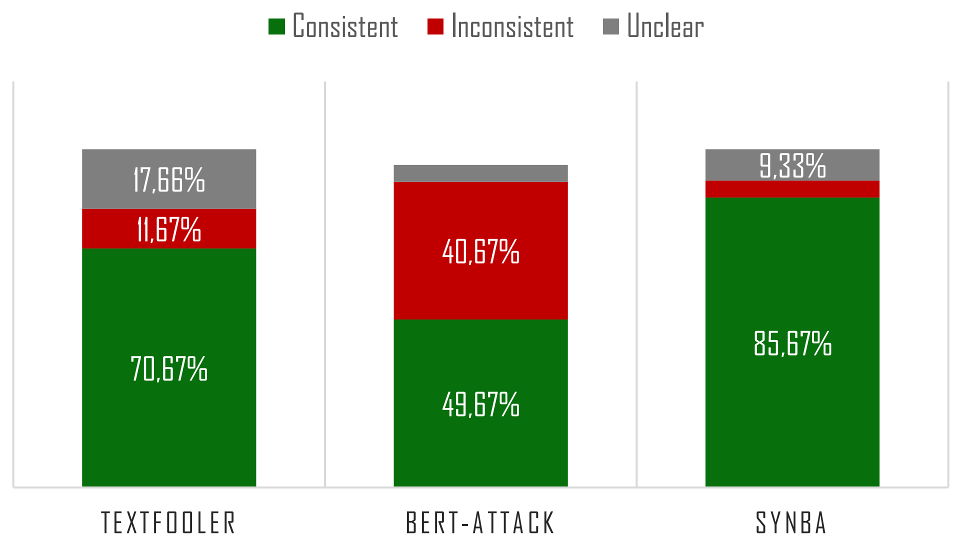
CONSISTENT	INCONSISTENT	UNCLEAR	skip

POS (99.31%) --> NEG (71.76%)

Original: the philosophical musings of the dialogue jar against the tawdry soap opera antics of the film's action in a way that is surprisingly enjoyable .

Perturbed: the philosophical musings of the dialogue jar against the tawdry soap opera antics of the film's action in a way that is oddly likeable .

Survey results



Some adversarial examples

Method	Label	Adversary
Original	POS (99.95%)	*peppered with witty dialogue and inventive moments.*
TextFooler	NEG (97.79%)	riddled with witty dialogue and inventive min.
BERT-Attack	FAILED	charming with easy ways and dry wit.
SynBA	NEG (99.88%)	riddled with witty dialogue and inventive moments.
Original	POS (99.95%)	*the most ingenious film comedy since being john malkovich.*
TextFooler	NEG (95.55%)	the most malignant film comedy since being john malkovich.
BERT-Attack	NEG (98.28%)	the most difficult film comedy since being john malkovich.
SynBA	NEG (92.26%)	the most artful film comedy since being john malkovich.
Original	NEG (99.94%)	*an unsophisticated sci-fi drama that takes itself all too seriously.*
TextFooler	POS (99.88%)	an impressionable sci-fi drama that takes itself all too attentively.
BERT-Attack	POS (98.36%)	an awesome sci-fi drama that takes itself all too soon.
SynBA	FAILED	an unsophisticated sci-fi tragedy that took itself all too heavily.

Conclusions

Limitations

- If the **MLM-Score** is much higher than the others in the final SynBA score, the best candidate is likely to be an **antonym** or **inconsistent** with the original counterpart
- Assessment performed only on **movie reviews** datasets
- **Contradiction rate** metric seems to be promising in the context of **sentiment analysis**, but it could be less informative for other tasks

Future developments

- Extend SynBA to **other tasks** like machine translation, question answering, and text summarization
- Add a **new constraint** which penalizes the use of words that lead to a contradiction
- Exploit enhanced **language models** to generate more semantically related perturbations

Thank you
all for your attention!