Team members:

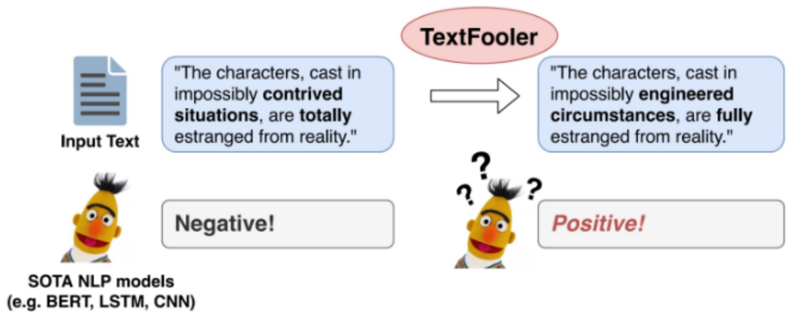
- Gia Trong Nguyen
- Nikita Sergeev

Current progress:

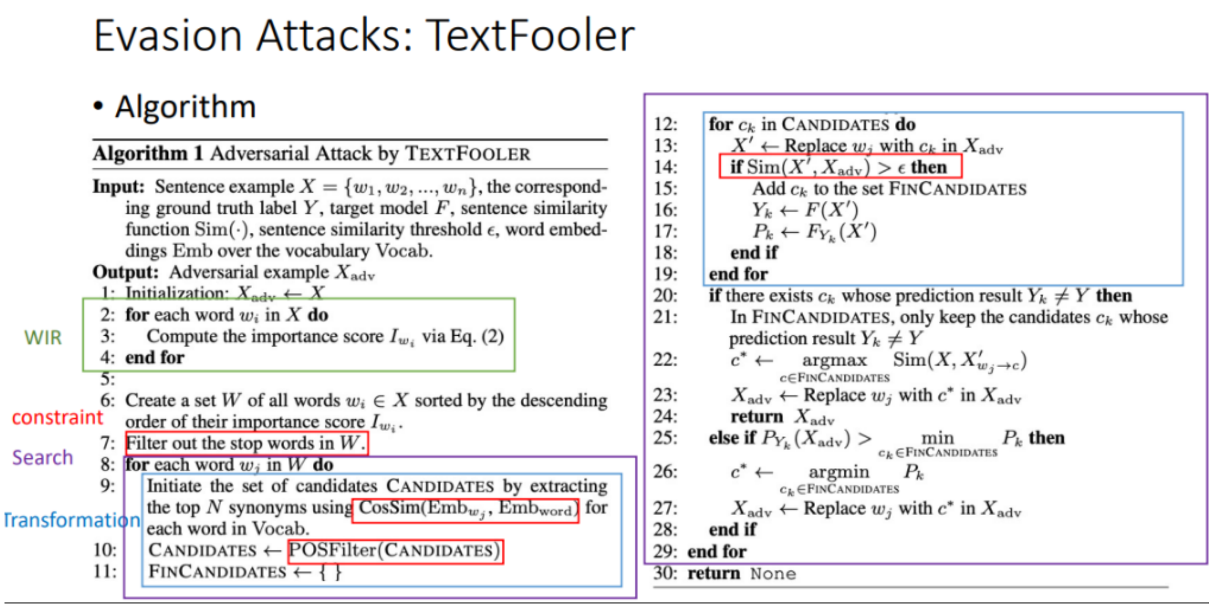
Implementing attack algorithm from scratch

- Algorithm was taken from this article. In the previous week we used an implemented version of this algorithm by *textattack* python package, but it's not good to use the already implemented code, because it gives zero understanding. So, we decided to implement this algorithm from scratch for better understanding of nlp adversarial attack.
- Some information about this algorithm:

Goal	Constraints	Transformation	Search Method
Untargeted Classification	1. Word embedding distance 2. USE sentence similarity 3. POS consistency	Word substitution by counter-fitted GloVe embedding space	Greedy search with word importance ranking



- Algorithm look like this:



- Let’s talk a little bit about constraints. As you can see, one of the steps of implementing is to implement constraints with cosine similarity. We need to define a corpus of synonym words. Our choice is the dataset from this [article](). Then we compute for every word top 50 synonyms. It was a hard part of our work, because we did not have a lot of resources. POS filter we do using *nltk* package. Stop words we also take from *nltk* package. To compute similarity of sentences we use *[Universal Sentence Encoder]()*.
- Implementation of algorithm you can find in **TextFooler FROM SCRATCH** block in our notebook.
- **We attack only toxic labels.**

Get the result:

- To understand if our attack works or not we try to attack 5% of train and valid datasets. Then make the pandas dataframe to visualize the results. Here is an example:

	Original Label	Attacked Label	Original	Attacked	Original target probability	Attacked target probability	Similarity of Sentences	Is Attack succeed?
25677	1	1	while i would dearly love to have my way with her in a wild and uninhibited fashion , jolin needs to realise she is chinese and thus not fucking african american . god damn that woman is confused . and watching her with luo zhixiang in that new video trying to dance like some hip hop slut is just embarrassing . stupid cunt that girl is . bet she fucks like a tiger though . diskindofsexist racist mntokhere ? ? ? ? ? ? ? wmf dapits ((((202 . 86 . 135 . 18	while i would dearly loved to have my way with her in a wild and playfulness means , jolin needs to recognizing she is chinese and thus not shit chau american . theology fucked that woman is confused . and observes her with luo zhixiang in that innovative video struggling to dance like some hips breaks slut is just embarrassing . nonsense lesbo that female is . bet she degenerates like a tiger though . diskindofsexist racism mntokhere ? ? ? ? ? ? ? wmf dapits ((((202 . 86 . 135 . 18	0.977109	0.887330	0.500560	No
71182	1	0	look at your own ridiculous behavior first . by the way , i'm a different user than the one you seem to be writing to about , even though i agree with this user more than i agree with you .	look at your own absurd behavior first . by the way , i'm a different user than the one you seem to be writing to about , even though i agree with this user more than i agree with you .	0.693431	0.228694	0.813941	Yes
135376	1	1	she 's probaly a gurgurshite .	she 's probaly a gurgurshite .	0.716491	0.716491	0.999655	No
125802	0	0	we , bulgarians , have nothing in common with any mongols or turks . this is an evil campaign started by our enemies about a hundred years ago to divide the bulgarians . only in the russian federation there are at least 20 million people of direct bulgarian origin , and the russian authorities simply do n't want them united because this is a threat to their empire , no wonder this article is within the scope of wikiproject russia , they support this anti bulgarian turkic crappy propaganda that has nothing to do with reality and is extremely abusive against us , bulgarians . we know very well who we are and where do we come from , we need no stupid shitopedia to brainwash us with this pathetic neo turkic crap .	we , bulgarians , have nothing in common with any mongols or turks . this is an evil campaign started by our enemies about a hundred years ago to divide the bulgarians . only in the russian federation there are at least 20 million people of direct bulgarian origin , and the russian authorities simply do n't want them united because this is a threat to their empire , no wonder this article is within the scope of wikiproject russia , they support this anti bulgarian turkic crappy propaganda that has nothing to do with reality and is extremely abusive against us , bulgarians . we know very well who we are and where do we come from , we	0.888947	0.888947	1.000000	Skipped
86723	1	1	yo you are a very big mother fucking idiot loser .	yo you are a greatly big mother damned idiot loses .	0.981596	0.955354	0.505714	No

Meaning of columns:

* *Original Label* - Label, which model predict

* *Attacked Label* - Label, after attacking

* *Original* - Original text

* *Attacked* - Attacked text

* *Original target probability* - Probability of Label, which model predict

* *Attacked target probability* - Probability of original Label, after attacking

* *Similarity of Sentences* - Cosine similarity, which show us semantic similarity of sentences

* *Is Attack succeed?* - Attack succeeded or not

- Some statistic:
  5% of Train dataset and 5% of Val dataset

```
Yes        279
No         170
Skipped     51
Name: Is Attack succeed?, dtype: int64
```

```
No          22
Yes         21
Skipped      7
Name: Is Attack succeed?, dtype: int64
```

  As you can see, our attack worked.
  * Yes - Attack is success
  * No - Attack is failed
  * Skipped - Model predict toxic label as non-toxic

Train robust model (adversarial training)

- To train a robust model we add success attack samples to our original dataset and then we retrain the model.

- So, after the previous step we compare the *performance* of two models (**Compute accuracy only for toxic label, exclude non-toxic samples**).

Comparing performance of original and adversarial trained models on valid dataset with adversarial examples.

```
[ ]
from IPython.display import clear_output

true, pred = evaluate(model, prepared_toxic_val_iterator, verbose=False)
clear_output()
print(f"Accuracy score of the Original model: {accuracy_score(true, pred)}")
```

```
Accuracy score of the Original model: 0.8349328214971209
```

```
[ ]
true, pred = evaluate(adversarial_model, prepared_toxic_val_iterator, verbose=False)
clear_output()
print(f"Accuracy score of the Adversarial trained model: {accuracy_score(true, pred)}")
```

```
Accuracy score of the Adversarial trained model: 0.9385796545105566
```

Conclusion: Here, as you can see, model, which not trained on adversarial examples give bad performance comparing with model trained on adversarial attack.

Comparing performance of original and adversarial trained models on valid original dataset.

```
[ ]
true, pred = evaluate(model, original_toxic_val_iterator, verbose=False)
clear_output()
print(f"Accuracy score of the Original model: {accuracy_score(true, pred)}")
```

```
Accuracy score of the Original model: 0.87
```

```
[ ]
true, pred = evaluate(adversarial_model, original_toxic_val_iterator, verbose=False)
clear_output()
print(f"Accuracy score of the Original model: {accuracy_score(true, pred)}")
```

```
Accuracy score of the Original model: 0.938
```

- To evaluate *robustness* of two models we use this formula:

$$\text{attack success rate} = \frac{\text{\# of successful attacks}}{\text{\# of total attacks}}$$

```
Original model attack success rate: 0.621380846325167
Adverasarial trained model attack success rate: 0.4863157894736842
```

As we can see, adversarial trained model become more robust than the original one.

Each member contribution:

Gia Trong - Implementing main algorithm of adversarial attack, Train models, Evaluate the models, Write Report
Nikita - Compute top 50 synonyms of the corpus, Prepare data for training and evaluating processes.

Future work:

- Fixing some little bugs. Bring beauty.
- Try to use other methods of attack from *textattack* package
- Prepare presentation for defending project