# Progress Report -- September 2022

In this directed study, CS 607 Topics in Computer Science, I have 24 to 25 research papers to read about the adversarial examples/attacks against Natural Language Processing (NLP) tasks, such as text classifications in general and sentiment analysis in particular. To study the evolution of these adversarial examples/attacks and to preserve their timeline, I divided these research papers into four groups based on their year of publication: 2018 (5 papers), 2019 (6 papers), 2020 (10 papers), and 2021+2022 (4 papers), and read them accordingly.

This past month of September 2022, I have read all the research papers published in 2018 (five papers); two papers discussed white-box attacks, and the others discussed black-box attacks. Here is a list of these papers:

1. Alzantot et al. (2018). *Generating Natural Language Adversarial Examples.* EMNLP 2018.
2. Gao et al. (2018). *Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers.* IEEE Symposium on Security and Privacy 2018.
3. Ebrahimi et al. (2018). *HotFlip: White-Box Adversarial Examples for Text Classification.* ACL 2018.
4. Kuleshov et al. (2018). *Adversarial Examples for Natural Language Classification Problems.* Rejected at ICLR 2018, but good paper to read.
5. Feng et al. (2018). *Pathologies of Neural Models Make Interpretation Difficult.* EMNLP 2018.

Throughout my reading of these five research papers, I kept a brief taxonomy of the adversarial examples/attacks in terms of the attack type, the targeted space, and the used method, hoping to turn this taxonomy into a review paper one day. The table below shows the brief taxonomy of the already-read papers for the year 2018.

| # | Paper | Authors | Attack | Space | Method |
|---|---|---|---|---|---|
| **1** | Generating Natural Language Adversarial Examples | Alzantot et al. (2018) | Black-box Attack | Embedding Space | Word-level Replacements |
| **2** | Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers | Gao et al. (2018) | Black-box Attack | Input Space | Character-level Replacements |
| **3** | HotFlip: White-Box Adversarial Examples for Text Classification | Ebrahimi et al. (2018) | White-Box Attack | Input Space | [+]Word-level Replacements [*]Character-level Replacements |
| **4** | Adversarial Examples for Natural Language Classification Problems | Kuleshov et al. (2018) | White-Box Attack | Embedding Space | Word-level Replacements |
| **5** | Pathologies of Neural Models Make Interpretation Difficult | Feng et al. (2018) | N/A[**] | Input Reduction | Word-level Removal |

[**]Feng et al. (2018) did not discuss the adversarial examples/attacks directly because the paper's main goal was to study the pathologies that make neural models challenging to interpret. Yet, the authors drew connections to adversarial examples/attacks and confidence calibration and explained why the observed pathologies are a consequence of the overconfidence of neural models.

Lastly, in the next month, October 2022, I will read all papers published in 2019 and 3 to 5 papers published in 2020. I also will keep building the brief taxonomy of the adversarial examples/attacks.

**--- End of Report ---**