# Exploring robustness of Language Models

Midterm check
CSE 598: ML Security and Fairness (2022 Fall)

Nakul Vaidya(1222318305)
Ayush Kalani(1219505612)

# MOTIVATION OF THE PROJECT

Existing large language models such as GPT3, BERT, and T5 hold particular promise as social science generalist tools to mimic general human behavior. We want to test the robustness of these models against data sets that are designed to fool the system, and which can be human judged correctly. While existing works test adversarial examples on image and language data sets, we want to test the robustness and generalizability of these models using TextFooler baseline benchmark.

# Related works

Recent there has been work on Generating Natural Language Adversarial Examples. These examples have been previously studied on older language models with the paper on Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment.

# PROBLEM STATEMENT

With the motivations from previous research we will be targeting GPT-3 with in this project as there has been very limited work that attacks GPT-3 model on text classification and entailment tasks.

# Motivation

- GPT-3 is a black box to everyone where no model architecture or parameters are accessible.
- This study will give us its comparison with the BERT model and help us in answering some of the advancements used in GPT-3.
- This project will require an in depth understanding of creating adversarial examples and learning the architecture of several large language models.
- This project will give us an opportunity to play with GPT-3 and discover new findings about it.

# Generating Adversarial Examples

Word Importance Ranking

- Find words that have more impact on accuracy by deleting words and calculating difference in accuracy. Calculate importance accordingly and rank words.

Word Transformer

- Using Synonyms extraction and semantic checking find replacement words for top important words.

# Experimental designs

1. In this project we will perform adversarial attacks on BERT and GPT-3 and benchmark results.
2. Use TextFooler framework to compare the attack between the two language models.
3. We are first going to finetune both the models on YELP and MultiNLI training set.
4. We are then testing the model on both the original training set and adversarial examples generated from the training set.
5. We will then compare the results obtained from both the language models .

# adversarial Example for MNLI dataset

**orig premise**: Cultures , and adults within them responsible for socialization , select different tasks for children 's learning .

**orig hypothesis**: ['Many', 'adults', 'have', 'learned', 'from', 'their', 'own', 'children', '.']

**adversarial hypothesis**: Many mature have learn from their own children .

# Adversarial Example for yelp dataset

**orig sent (1):** this is a great little place to grab sushi for lunch the staff are friendly and the prices are fair a great alternative to the typical lunch fair

**adv sent (0):** this is a gargantuan little place to grab sushi for lunch the staff are chatty and the fare are fair a great alternative to the typical lunch fair

# Initial results

- For this phase we have fine-tuned bert-base-uncased model.
- For sequence classification fine-tuned using yelp_polarity training dataset and tested on original testing and TextAttack Dataset.
- For textual entailment fine-tuned using MNLI training dataset and tested on original testing and TextAttack Dataset.

- https://github.com/ayushkalani/robust-language-models

| Dataset | Original Accuracy | After-Attack Accuracy |
|---|---|---|
| **YELP** | 0.96994 | 0.06681 |
| **MNLI** | 0.85135 | 0.09631 |

# References


- https://arxiv.org/pdf/1907.11932.pdf
- https://huggingface.co/textattack
- https://github.com/jind11/TextFooler
- https://beta.openai.com/docs/guides/fine-tuning
- https://cims.nyu.edu/~sbowman/multinli/