The problem:

- Nowadays, online social media spreads out many news but some of them send out false information called fake news which are currently severe problems. In this assignment, we will mainly focus on **fake news detection on online social media**.
- How is fake news generated?
    - By human
    - By programmed bots (mainly and mostly)
- If the fake news is created by programmed bots, then the problem is really hard to solve since the bots are becoming more complex making detection much harder.
- The existing algorithms can detect whether the news is fake or not, but not perfectly.

Detect whether the news was generated from GPT-2 or not.

1. Two AI algorithms we would apply:
    a. Classifier 1
    b. Classifier 2

(Note: use classifiers like Neural network, SVM, bagging, boosting, Naive Bayes, GLTR, RoBERTa...)

2. Dataset: https://github.com/openai/gpt-2-output-dataset/tree/master/detector
    a. Webtext (combine train+valid+test, then get 50000 randomly)
    b. A dataset generated from gpt2
        i. `xl-1542M-k40.${split}.jsonl`
3. Approach of the experiments
    a. Data preparation
        i. Download data and combine the two data sets (Huong, by Tuesday 11/24)
        ii. Data exploration (Minh, Fri 11/27)
        iii. Data preprocessing (Minh, Sat 11/28)
        iv. Representation for features: doc2vec (word embedding), tf-idf,.... (Huong, by Sun 11/29)
            - https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-10-neural-network-with-a6441269aa3c
    b. Applied classifiers (will also use Stratified k-Fold Cross Validation along with the classifiers to improve the performance):
        i. Classifier 1 (CNN) (Huong by Tue, Dec 1)
        ii. Classifier 2 (Bagging) (Minh, by Dec 1)
    c. For each classifier, we will collect criteria scores to evaluate the performance. Criterias to evaluate the performance, for example:

```
(from sklearn.metrics import accuracy_score, f1_score,
recall_score, precision_score)
```

        i. Accuracy/ AUC
        ii. Precision
            1. Precision = True Positive/(True Positive+False Positive)

        iii. Recall
            1. Recall = True Positive/(True Positive+False Negative)
        iv. F1 score (is the weighted average of precision and recall).
            1. F1 = 2*(Precision*Recall)/(Precision+Recall)
4. Result of the experiments
5. Conclusion and discussion

Idea 2:
Dataset:

- Webtext
- A dataset generated from gpt2
    - `xl-1542M-k40.${split}.jsonl`

GPT-2 model is trained on the WebText training set.
How good is GPT-2 at producing an article of similar content?
Approach:

- Checking grammar

Reference Link:

1. Fakes news https://en.wikipedia.org/wiki/Fake_news
2. Cross validation https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85
3. Detecting Fake News With and Without Code https://towardsdatascience.com/detecting-fake-news-with-and-without-code-dd330ed449d9
4. Full Pipeline Project: Python AI for detecting fake news https://towardsdatascience.com/full-pipeline-project-python-ai-for-detecting-fake-news-with-nlp-bbb1eec4936d
5. An Exhaustive Guide to Detecting and Fighting Neural Fake News using NLP https://www.analyticsvidhya.com/blog/2019/12/detect-fight-neural-fake-news-nlp/
6. OpenAI's GPT-2: A Simple Guide to Build the World's Most Advanced Text Generator in Python https://www.analyticsvidhya.com/blog/2019/07/openai-gpt2-text-generator-python/
7. RoBERTa https://huggingface.co/transformers/model_doc/roberta.html
8. Combine dataframe from different files https://realpython.com/python-keras-text-classification/
9. Cleaning text https://medium.com/analytics-vidhya/twitter-sentiment-analysis-b9a12dbb2043
10. Detect language https://stackoverflow.com/questions/39142778/python-how-to-determine-the-language

Related researches:

1. Fake news detection within online social media using supervised artificial intelligence algorithms https://www.sciencedirect.com/science/article/abs/pii/S0378437119317546
2. Detecting Fake News in Social Media Networks https://www.sciencedirect.com/science/article/pii/S1877050918318210
3. H https://arxiv.org/ftp/arxiv/papers/1908/1908.09203.pdf

Note for detecting actual news or faked news which are generated from GPT-2:

- Shorter lengths are more difficult to classify (real or fake).
- "Fine-tuning GPT-2 on more narrow datasets tends to increase the perceived humanness of GPT-2-generated text. Fine-tuning is a key variable to take into account in the context of both human and ML-based detection." (https://arxiv.org/ftp/arxiv/papers/1908/1908.09203.pdf)