Homework: Large Scale Generation of Falsified Scientific Literature and Detection

Due: August 7th, 2022, 11:59:59 p.m. PT (No Extension)

---

1. Overview

Counteracting neural disinformation with Grover

Exploring the surprising effectiveness of a fake news generator for fake news detection

Rowan Zellers Follow

Jun 18, 2019 · 8 min read

By Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi

Artificial Intelligence has enormous potential to benefit society. However, the same technology can cause harm, particularly if used by malicious adversaries. One important threat is that of "Neural Fake News": machine-written disinformation at scale. Our

Figure 1: Neural network generated falsified information.

In this second assignment, we will leverage your augmented Bik dataset and features and use it to generate believable, fake scientific literature. To do this, we will leverage one technique for falsified content generation at scale - **Generative Pre-trained Transformer 2** (GPT-2).

Your goal is to parse out and extract the texts from the 200 scientific publications identified in the Bik dataset using Apache Tika, and use the texts to train a new GPT-2 model. After training, GPT-2 can generate new "fake" scientific, believable texts. You will then convert texts to PDFs.

2. Objective

The objective of the assignment is to take the next steps in applying content extraction and feature generation by seeing if you can generate a training set of falsified media. The training set will be used to automatically generate falsified papers. In the first assignment, you focused on the unintended consequences of Big Data and the associated features to see if you could explain why the papers in the Bik dataset were manipulated. In this assignment, you will focus on using the contents to train a model for scientific paper falsification generation. A scientific paper usually has texts, images, and tables. To simplify the paper generation process, you only need to generate texts using GPT-2 for a paper.

You will generate in this assignment a new TSV dataset, which has the new rows for generated falsified papers. To do this, you will generate the features in the original Bik dataset for your new papers. My suggestion is to use services, such as Fake Name Generator to generate author names. For the institutional information, you can sample from the distributions of Institutions in your TSV dataset from homework 1. You can get creative and even generate these features for your new 500 papers based on similarity of text, and other distance metrics like we talked about during the Clustering and Deduplication lectures.

3. Tasks:

1. Write a Python or other script to download all 200 papers in the Bik dataset. You should be able to access the papers institutionally through USC's connection. Store this dataset of PDFs you will use it later in the next steps

2. Run Tika-Python on the downloaded PDFs to extract out the text from the PDFs

3. Write a Python script using sites, like FakeNameGenerator, to generate the author names. For affiliations, write a python script to sample from your 200 Bik papers for your new falsified papers.
   Note: For the name generation, you can use some of the free online resources to generate the fake name-surname combinations (e.g., https://fungenerators.com/api/fakeidentity/, https://parser.name/api/generate-random-name/) or you can make your own fake name generator using some of the names databases. For the title, institution name and other features, sampling from the existing values is sufficient. How to randomly sampling institution names and other features, please refer to https://pynative.com/python-random-choice/

4. Train (Fine Tune) a GPT-2 (124M) model
   using the texts from step 2. Notes:
   - You can find the GPT-2 implementation online or implement your own GPT-2. One implementation I suggest is https://github.com/minimaxir/gpt-2-simple
   - If your computer does not have a modern CPU or GPU, you can use

Kaggle or Google Colab

5. Generate 500 fake papers from the trained GPT-2 in step 4.

6. Write a Python script to automatically generate PDFs of false papers. Each PDF includes paper title from GPT-2, author's information from step 3, and texts from GPT-2. Please refer to the downloaded paper from step 1 for the paper format. Save the papers' PDFs in a folder called "falsified_media".

7. Generate the new rows in your TSV for your new papers and their associated features from all prior steps. Each new row has features for the newly generated fake papers, such as author names, affiliations, title.

Extra Credits:
You can obtain extra credits from two parts.

The first part is about the training (fine-tuning) data for GPT-2. You will get 20% extra credits if you train GPT-2 using more than 200 papers from step 1. The extra papers are from the reference lists of 200 papers. Therefore, you need to download not only 200 papers in the Bik dataset but also the referenced papers listed in 200 papers.

The second part is about generating PDFs. You will get 20% extra credits if you can automatically use LaTex to convert texts to PDFs. Hints: Check out the code.

4. Assignment Setup

4.1.Group Formation

You should keep the same group from your assignment one. There is no need to send any emails for this step, unless there are changes in the groups.

5. Report

Write a short report (no more than 4 pages) describing your observations, i.e. what you noticed about the dataset as you completed the tasks. Please answer the following questions:

1. What did the GPT-2 generated texts look like?
2. Were they believable?
3. Would your associated ancillary features from assignment 1 have been able to discern what was false or not?

Thinking more broadly, please also answer the following (there are no right or wrong answers here, you have the freedom to express your opinion as you find appropriate):

4. How much do you think media falsification is solvable using ancillary metadata features, or using actual content based techniques? Is one better than the other?

5. What other types of datasets could have been used to generate the falsified papers? Pick at least 2 datasets from distinct MIME types.
6. What other sorts of "backstopping" would be required to generate a believable paper trail for the scientific literature?

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 11:59:59 pm PT*** on August 7th, 2022(No Extension), via D2L > My Tools > Assignments. A team can submit multiple times, but only the last submission counts. Anyone from a team can submit. However, we suggest designating one person to submit. In the submission, please include following things:

- All source code is expected to be commented, to compile, and to run. You should have at least the identified Python scripts that you used to generate the falsified media, download the scientific papers and generate associated features.
- Your updated dataset TSV.
- "falsified_media" folder storing the fake papers
- Also prepare a readme.txt describing the files in the submission.
- If you used external libraries other than Tika Python and GPT-2, you should include those necessary files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.
- Save your report as a PDF file (TEAM_NAME_EXTRACT.pdf) and include it in your submission.

Please compress all of the above into a single zip archive and name it according to the following filename convention:

TEAM_NAME_DSCI550_HW2_EXTRACT.zip

Use only standard zip format. Do not use other formats such as zipx, rar, ace, etc.

If your data is too big and exceeds the D2L file limit of 2GB: 1) upload your data to Google drive, 2) include the links to the data in a README file, 3) compress the report, README file and the code and upload it to D2L.

Important Note:

- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.
- Again, please note, a team can submit multiple times, but only the last submission counts. **To avoid confusion: designate someone to submit.**