Across the Sea

601.466 Information Retrieval Final Project

https://github.com/bohanhou14/across_the_sea/
fine-tune model code: https://colab.research.google.com/drive/1kRY0NX-fyV_ww6hI07y3qqjzxz5_ouQP?usp=sharing
HuggingFace Model: https://huggingface.co/bohanhou14/abortion_news_model

Abe Hou
bhou4@jhu.edu

Peter Penev
ppenev1@jhu.edu

Summary: We built a system for helping readers get exposed with alternative opinions of the articles they read. The system composed of a classifier and a crawler. When readers read opinionated articles, the classifier would first analyze its ideology {liberal, neutral, conservative}. Then, the retriever would find and return to the readers articles on the same topic but encompass different ideologies. Using an abortions news dataset we created, we fine-tuned a language model (POLITICS) that was pretrained for ideology and stance detection and built our classifier. For the retriever, we built a web crawler that would search for abortion news article of different ideologies and recommend to the readers. We came up with this project witnessing the ideological divide created by social media and political rhetoric. Instead of using algorithms to find those who think alike, we hope to swim across the sea and learn from others' perspectives.

Cover Photo Credit: *Attack on Titans* Season Three Last Episode

(2) Please provide a brief guide describing in sufficient detail how a user could run your code and/or use your provided web interface, especially if using a non-standard platform or interface.

For a user to run the code, first the user must find a news article that is on the topic of abortion. Then they must run the crawl.py file with the input of python crawl.py *url*.

Example:

Suppose we are reading an article on north-carolina abortion ban reported by nbcnews.

```
python crawl.py https://www.nbcnews.com/politics/politics-
news/north-carolina-republicans-unveil-agreement-12-week-
abortion-ban-rcna82727
```

The program would return:

```
The stance of the article on abortion you are reading is:
neutral

with extremeness: 0.7820116281509399

alternative article with conservative ideology found with an
extreme score of 0.580888032913208

text: Medical doctors continue to testify to the atrocities
committed during partial-birth abortions in three federal court
cases currently being heard in New York, Nebraska and
California. An official with the U.S. Conference of Catholic
Bishops… (abbreviated for better clarity)

url: https://www.catholicnewsagency.com/news/1091/partial-birth-
abortion-trials-show-abortionists-detached-from-reality

----------

alternative article with liberal ideology found with an extreme
score of 0.987561047077179

text: On Thursday, the Center for Reproductive Rights, a pro-
abortion lobby and legal support group, held a briefing on
Capitol Hill to explain to congressional staff its continuing
plan to bring abortion to countries around the world -
particularly to Latin America. The meeting was titled "Advancing
Reproductive Rights in Latin America…
```

url: https://www.catholicnewsagency.com/news/7082/abortion-group-tells-congress-about-plans-to-advance-abortion-in-latin-america

Explanation:

The program first classifies the article which the starting url points to. And then crawl the web starting from the websites “https://www.catholicnewsagency.com/tags/35/abortion” and https://www.nbcnews.com/politics/abortion-news, which are two abortion news websites we selected that are open to crawling. The crawler will find the targeted news articles and print the relevant information. Some important files to keep track of is the extracted.txt file and the visited.txt file. The extracted.txt file shows the extracted files and the score of the document, its url, and the content of the news article. The visited.txt file shows a list of all the urls that the crawler visited.

Note that our program only supports classification and retrieval related to abortion news, because we fine-tuned our model to abortion news.

(3) Give an itemized list of your project’s particular achievements, strengths and selling points, especially things that may not be obvious upon inspection of your code or output, as well your personal assessment of their complexity, scope, and success.

We curated and labeld an abortion news datasets by ourselves and used it to fine-tune a model based on the POLITICS model. It works surprisingly accurate especially on news articles from catholicnewsagency and nbc news. To find which classifier we should be using, we actually experimented several, including the vanilla POLITCS model, OpenAI Davinci, Princeton CSE (a model for sentence embedding). We also fine-tuned an OpenAI Davinci model based on the dataset, but it actually worked less well than our fine-tuned POLITICS model. That’s how we ended up choosing our classifier.

The project has the complexity of a small serious research paper, given that we read many papers for choosing the right models, experimented several of them, and built a news dataset by ourselves.

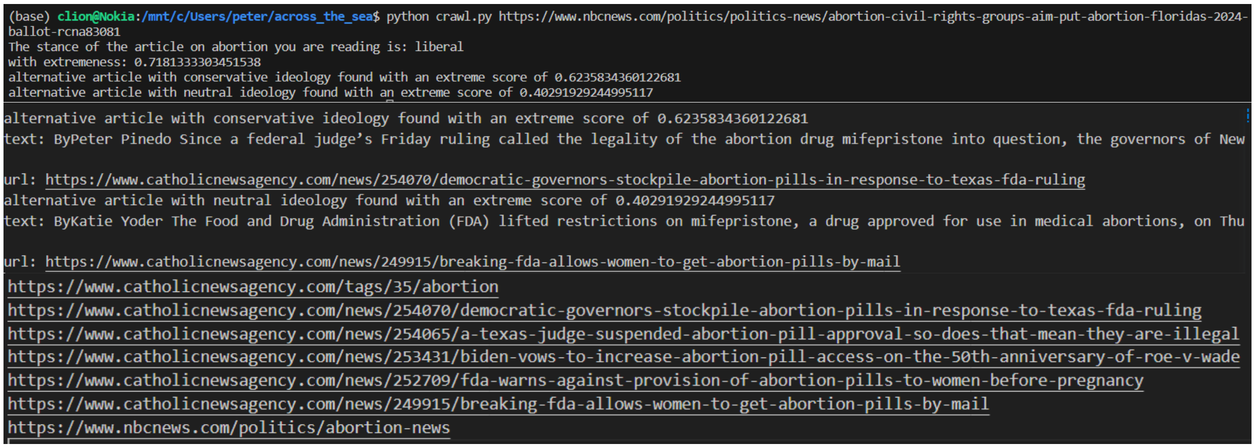
We are satisfied with our project outcome. Three years ago, Abe actually attempted this same idea, but he failed to build a working model because 1) he tried to do too many things – he wants to build a model that supports all kinds of opinion classifications, not limited to abortion news – but this was very hard in 2020; 2) not many models and frameworks were out there at the time, and he tried with LSI which did not work very well; 3) he didn’t try to build a dataset and fine-tune any models; 4) he didn’t have Peter as his partner.

(4) Give a brief itemization of your project’s limitations and list suggested possibilities for improvement or worthwhile extension with additional time.

The main limitation of the current project is that we are currently restricted to only the topic of abortion. Some possibilities for improvement with extra time would be to increase the scope of the topics to either all news or most major news issues. Another possibility is also having the news articles be ranked on a scale from very liberal to very conservative and having the algorithm attempt to find an article of the opposing viewpoint with an almost equal amount of bias instead of labeling some articles as either conservative or liberal. Also, our model performance would degrade on news not from catholicsnewsagency and nbc news. If possible, we would like to build a larger abortion news dataset and improve fine-tuning.

(5 and 6) To both document your projects achievements, as well as the methods of use and expected output(s), please also include a range of samples/screenshots of your project output on a range of project functionalities and input conditions or search queries. In the spirit of all assignments in the class, empirical evaluation appropriate for the nature of the project should also be included.

Liberal base file:



```
(base) clion@Nokia:/mnt/c/Users/peter/across_the_sea$ python crawl.py https://www.nbcnews.com/politics/politics-news/abortion-civil-rights-groups-aim-put-abortion-floridas-2024-
ballot-rcna83081
The stance of the article on abortion you are reading is: liberal
with extremeness: 0.7181333303451538
alternative article with conservative ideology found with an extreme score of 0.6235834360122681
alternative article with neutral ideology found with an extreme score of 0.40291929244995117
```

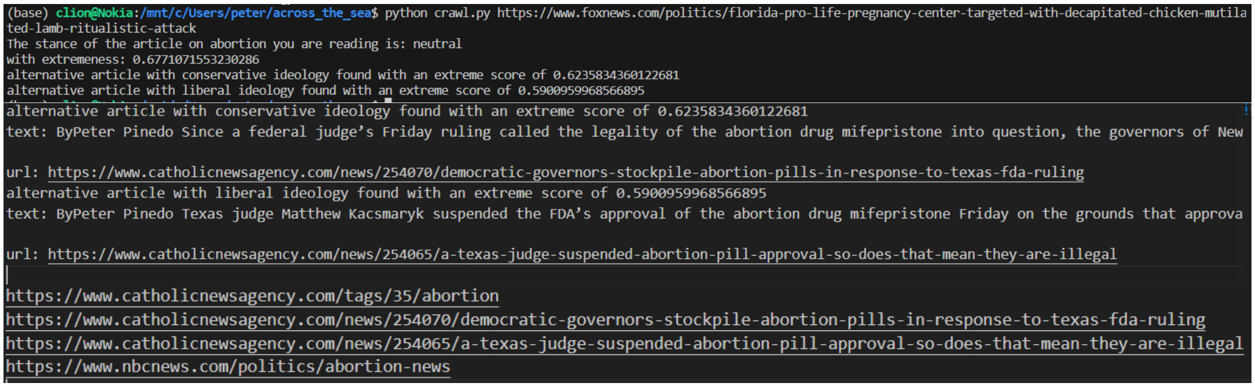
```
alternative article with conservative ideology found with an extreme score of 0.6235834360122681
text: ByPeter Pinedo Since a federal judge’s Friday ruling called the legality of the abortion drug mifepristone into question, the governors of New

url: https://www.catholicnewsagency.com/news/254070/democratic-governors-stockpile-abortion-pills-in-response-to-texas-fda-ruling
alternative article with neutral ideology found with an extreme score of 0.40291929244995117
text: ByKatie Yoder The Food and Drug Administration (FDA) lifted restrictions on mifepristone, a drug approved for use in medical abortions, on Thu

url: https://www.catholicnewsagency.com/news/249915/breaking-fda-allows-women-to-get-abortion-pills-by-mail
```

```
https://www.catholicnewsagency.com/tags/35/abortion
https://www.catholicnewsagency.com/news/254070/democratic-governors-stockpile-abortion-pills-in-response-to-texas-fda-ruling
https://www.catholicnewsagency.com/news/254065/a-texas-judge-suspended-abortion-pill-approval-so-does-that-mean-they-are-illegal
https://www.catholicnewsagency.com/news/253431/biden-vows-to-increase-abortion-pill-access-on-the-50th-anniversary-of-roe-v-wade
https://www.catholicnewsagency.com/news/252709/fda-warns-against-provision-of-abortion-pills-to-women-before-pregnancy
https://www.catholicnewsagency.com/news/249915/breaking-fda-allows-women-to-get-abortion-pills-by-mail
https://www.nbcnews.com/politics/abortion-news
```

Neutral base file:



```
(base) clion@Nokia:/mnt/c/Users/peter/across_the_sea$ python crawl.py https://www.foxnews.com/politics/florida-pro-life-pregnancy-center-targeted-with-decapitated-chicken-mutila
ted-lamb-ritualistic-attack
The stance of the article on abortion you are reading is: neutral
with extremeness: 0.6771071553230286
alternative article with conservative ideology found with an extreme score of 0.6235834360122681
alternative article with liberal ideology found with an extreme score of 0.5900959968566895
```

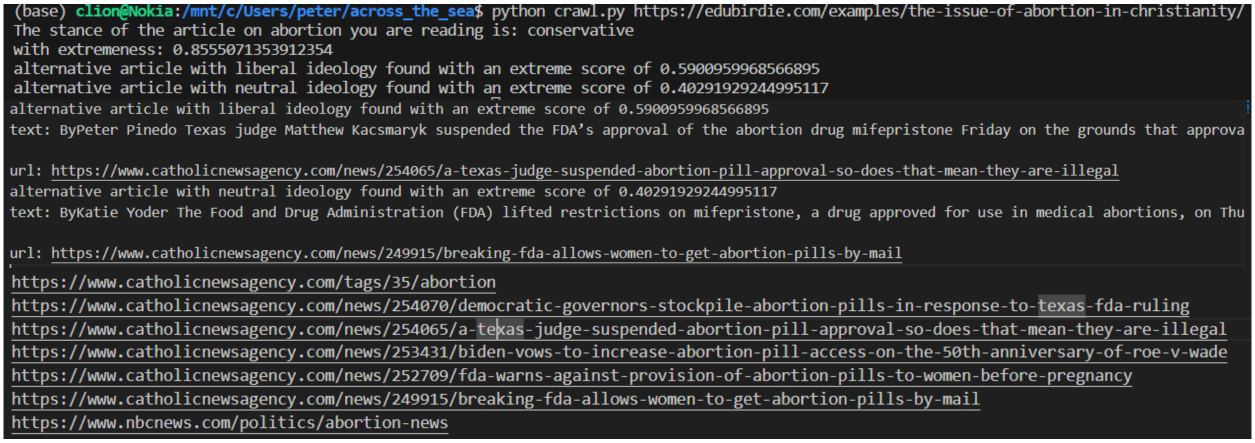
```
alternative article with conservative ideology found with an extreme score of 0.6235834360122681
text: ByPeter Pinedo Since a federal judge’s Friday ruling called the legality of the abortion drug mifepristone into question, the governors of New

url: https://www.catholicnewsagency.com/news/254070/democratic-governors-stockpile-abortion-pills-in-response-to-texas-fda-ruling
alternative article with liberal ideology found with an extreme score of 0.5900959968566895
text: ByPeter Pinedo Texas judge Matthew Kacsmaryk suspended the FDA’s approval of the abortion drug mifepristone Friday on the grounds that approva

url: https://www.catholicnewsagency.com/news/254065/a-texas-judge-suspended-abortion-pill-approval-so-does-that-mean-they-are-illegal
```

```
https://www.catholicnewsagency.com/tags/35/abortion
https://www.catholicnewsagency.com/news/254070/democratic-governors-stockpile-abortion-pills-in-response-to-texas-fda-ruling
https://www.catholicnewsagency.com/news/254065/a-texas-judge-suspended-abortion-pill-approval-so-does-that-mean-they-are-illegal
https://www.nbcnews.com/politics/abortion-news
```

Conservative base file:



```
(base) clion@Nokia:/mnt/c/Users/peter/across_the_sea$ python crawl.py https://edubirdie.com/examples/the-issue-of-abortion-in-christianity/
 The stance of the article on abortion you are reading is: conservative
 with extremeness: 0.8555071353912354
 alternative article with liberal ideology found with an extreme score of 0.5900959968566895
 alternative article with neutral ideology found with an extreme score of 0.40291929244995117
alternative article with liberal ideology found with an extreme score of 0.5900959968566895
text: ByPeter Pinedo Texas judge Matthew Kacsmaryk suspended the FDA’s approval of the abortion drug mifepristone Friday on the grounds that approva

url: https://www.catholicnewsagency.com/news/254065/a-texas-judge-suspended-abortion-pill-approval-so-does-that-mean-they-are-illegal
alternative article with neutral ideology found with an extreme score of 0.40291929244995117
text: ByKatie Yoder The Food and Drug Administration (FDA) lifted restrictions on mifepristone, a drug approved for use in medical abortions, on Thu

url: https://www.catholicnewsagency.com/news/249915/breaking-fda-allows-women-to-get-abortion-pills-by-mail
https://www.catholicnewsagency.com/tags/35/abortion
https://www.catholicnewsagency.com/news/254070/democratic-governors-stockpile-abortion-pills-in-response-to-texas-fda-ruling
https://www.catholicnewsagency.com/news/254065/a-texas-judge-suspended-abortion-pill-approval-so-does-that-mean-they-are-illegal
https://www.catholicnewsagency.com/news/253431/biden-vows-to-increase-abortion-pill-access-on-the-50th-anniversary-of-roe-v-wade
https://www.catholicnewsagency.com/news/252709/fda-warns-against-provision-of-abortion-pills-to-women-before-pregnancy
https://www.catholicnewsagency.com/news/249915/breaking-fda-allows-women-to-get-abortion-pills-by-mail
https://www.nbcnews.com/politics/abortion-news
```

(7) If your code submission includes components that were (a) written by people not on your team or (b) written by yourself for another course, for prior research and/or employment, please very clearly explain/document which components were done elsewhere. Such outside borrowing is permitted if clearly documented, and you are permitted to build upon prior accomplishment, just like you may utilize software libraries from elsewhere, but your project will be evaluated on the original work done for this course.

We utilized a repository of abortion information from https://arxiv.org/pdf/2302.01439.pdf and we also used a pretrained POLITICS model that was significantly modified to help achieve the project goal. However, no code was duplicated or copied form any source excluding the template code for hw4 for this class for the crawler.