Project Proposal
Jason Lunder
CPSC 325

Introduction:
The objective of this project is to develop a model using Generative Adversarial Networks (GANs) that can accurately identify whether a research paper was published by a reputable or non-reputable publisher. The model will be trained on a dataset of research papers published by reputable and non-reputable publishers as per Beal's list. The GANs will be used to generate synthetic papers that are similar in style to those published by the non-reputable publishers to train the discriminator. The discriminator will then be used to classify whether a given paper was published by a reputable or non-reputable publisher. There are many journals that have been identified as predatory and otherwise non-reputable. These journals often do not properly review the papers submitted to them. This can result in nonsensical papers being published. Since I haven't tracked down a dataset of these nonsensical papers, many of which are AI generated, I'll focus on attempting to identify where a paper came from, as in, whether the paper was published in a journal listed on Beal's list or not.

Stakeholders:
The stakeholders interested in this project are researchers, academics, and publishers who want to ensure the quality of the research papers they read, publish or review. Potentially, a model that can accurately classify whether a paper looks like it came from a reputable source could help researchers identify when their papers sound less professional. Additionally, it could aid in identifying what kinds of papers have the potential to spread misinformative facts. It is, however, entirely possible that there will be little to no pattern between papers published in the predatory and non-reputable journals identified by Beal's list and the way that the papers are written. This project will also have relevance to AI researchers who are interested in using GANs for classification problems.

Tech Stack:
The tech stack for this project will include Python, PyTorch, GANs, and Google Cloud Services. PyTorch will be used to build and train the GAN model. The GANs will be used to generate synthetic research papers that are similar in style to those published by non-reputable publishers to train the discriminator. I will also use S3 if necessary to store data, and use BeautifulSoup for webscraping.

Data Sources:
The primary data source for this project will be research papers published by reputable and non-reputable publishers as per Beal's list. The papers themselves will come from the arXiv publicly available dataset of papers on Kaggle. These papers are machine readable and tagged with usefult information, including the journal that the paper was published in. I will then scrape the journals listed in Beal's list of standalone journals to have a base of journals to choose from, as well as attempt to supplement this with journals associated with publishers in Beal's list of publishers. The second task is much harder to do webscraping wise. Once I have a list of predatory journals, I will be able to cross reference this with my arXiv list of papers, to determine how many examples I have in the dataset. I will supplement and prune the list of papers as necessary to have a sufficient proportion of papers published in predatory (non-reputable) and non-predatory (reputable) journals. The dataset will be split into training, validation, and test sets. The training set will be used to train the GAN model, and the validation set will be used to evaluate the performance of the model. The test set will be used to test the accuracy of the model.

Major Analyses and Anticipated Findings:
The major analysis for this project will be to evaluate the performance of the GAN model in identifying research papers published by non-reputable publishers accurately and thereby evaluate if there is a significant and useful pattern in the way that papers published to reputable and non-reputable journals are written. The anticipated findings for this is that though there may be some pattern that allows the discriminator to identify whether a paper was published in a non-reputable journal, but that it may be difficult to pick up on, if it exists at all. This is because the language of the paper comes down to the authors of the paper. In the case of reputable journals that properly engage in the peer review process, one would expect the papers to be more well written, whereas non-reputable journals may be more susceptible to allowing poorly written or gibberish papers to be published. However, the journals considered to be reputable are not immune to the peer review process being executed poorly, as many have been found to contain many papers that are gibberish or nonsense. So, the data may not be clean enough for the model to identify a useful difference between papers from journals on Beal's list and those that are not. However, the generator model may still learn to generate convincing sounding papers regardless of the discriminators ability to accurately pick up on whether the paper came from a journal on Beal's list or not, which is a useful result in and of itself.

Deployment:
The deployment of this project will, ideally, involve hosting the GAN model on a cloud platform, such as Google Cloud or another. Ideally, users will access the model through an API. Short of these goals, the model and associated API to interact with it will be made useable through a docker image. The cost of deployment will depend on the scale of the project and the number of users. However, academic credits can be obtained from cloud providers to offset the cost.

Budget:
The budget for this project will depend on the scope and scale of the project. The cost of deployment will be the primary expense, and academic credits will be utilized to offset this cost. Other expenses may include the cost of acquiring additional research paper datasets, if necessary.

Risks and Mitigation Plans:
The primary risk of this project is the accuracy of the GAN model, which can be mitigated through rigorous testing and evaluation. As mentioned earlier, there is a risk that the dataset will not be clean enough for the model to learn what I want it to learn. This can be mitigated by, potentially, using off the shelf models trained to recognize whether text is part of a language or not to prune models consisting of non-English text. This could have the added effect of removing papers that are English but contain nonsense or gibberish. This could be problematic as these papers are an integral part of what I am trying to test/train for, but this could be attempted if necessary. It is also important to recognize that the correlation required for a model to learn to discriminate between papers published in journals on Beal's list and not may not exist or may not be significant enough to learn from, and this would be a valid result of the research, and, if there is a high proportion of nonsense papers that weren't peer reviewed in journals on Beal's list, and my model is incapable of learning to discriminate between papers published in journals of either type, it may indicate that the "reputable" journals are also inundated with nonsense papers that made it past their peer review processes.