Open for Business or Open for Attack? MAUDE and the Information Warfare Risks of Open Government

Abstract

Open government data is a concept where governments make their data freely available to citizens in order to help them better participate in democratic processes, and to help keep their representation honest. While open government does not apply to data and documents that pose a risk to national security, the data which is available may still pose a risk with regards to information warfare. The movement towards open government provides bad actors hoping to use information warfare with an interface they can use for negative purposes, and that it is easy to leverage the volume of text with currently available text generation techniques to augment an information campaign. These campaigns can be used to waste government resources trying to investigate adversary-created phenomena, and these phenomena, especially at large volumes can be used as evidence in creation of false narratives, or can trigger real journalistic inquiries into companies causing loss of reputation and profits. While the scope of what information is available in open government initiatives is limited to data that cannot be used to threaten national security, we believe that bad actors can exploit the data currently available by using current text-generation techniques to create threats to governmental functions. The impact of these techniques has not been fully explored in this domain.

To limit the scope of this project, we will focus on one relatively innocuous open database, the MAUDE database, and explore the possible opportunities for its (mis)use in information campaigns. The MAUDE (Manufacturer And User facility Device Experience) database is a US FDA database for reporting adverse events to the FDA. Manufacturers use adverse when a medical device manufacturer would like to start selling their product in another market, as well as to ensure that their devices, as well as similar devices made by their competitors, are safe. I anticipate that machine-generated false adverse event reports to the MAUDE database will be indistinguishable from human-written adverse events to 90% of subject matter experts.

This research will help to highlight a gap between the current-defined practices of what data governments should share in open-government portals, and what governments currently share from a security perspective.

Section 1: Introduction and Overview

Open government data is a concept where governments make their data freely available to citizens in order to help them better participate in democratic processes, and to enable citizens to help ensure that their representation is properly representing them (OECD, n.d.). This data can cover a wide variety of different topics. At a local level, open datasets might include city employee salaries, or how many people use a specific bus or library, on a national level, this might include census data, or data that helps to track the safety of different products. Ideally, when these open government platforms are well managed, and well utilized, they help citizens to stay informed, and to vote in their best interests.

Open government does not, however, mean that a government is sharing all of its documents and data with its citizens, especially in situations where the availability of a document may pose a risk to national security (Ubaldi, 2013). While this seems like it might be clear as to what is or is not a risk to national security, there are gray areas, like the potential for someone to leverage open government data in an information warfare campaign. In this project, the problem I am exploring is the possibility for bad actors to use text-generation techniques with the vast amount of freely available MAUDE text data to create sustained information warfare campaigns. This research is a stepping-stone to help develop strategies for building resilience to information warfare campaigns leveraging open government data.

To limit the research scope, I would like to explore these opportunities in the context of just one relatively innocuous open database, the MAUDE database, and explore the possible opportunities for its (mis)use in information campaigns. The MAUDE (Manufacturer And User facility Device Experience) database is a US FDA database for reporting adverse events to the FDA. Healthcare researchers use regulatory databases like MAUDE to assess the risks associated with specific devices and treatments (Frauger et al., 2011). Large numbers of adverse events being reported rapidly, even without a coherent description could, in aggregate, give the appearance of product failings. These can hurt a device manufacturer's chances of selling a product in another market, or if the evidence is compelling enough, lead to a product recall, or even a type of treatment being deemed by the FDA as not being safe.

To generate the samples, we will experiment with two different techniques: first, we will use a GPT-based approach that uses neural networks to predict the next word in a sequence (Radford et al., 2018); second, we will use a Malware-Induced Misperception (MIM) based approach to plagiarize an adverse event and replace a couple of pre-determined key-phrases with content-relevant substitutions (Sharevski and Jachim, 2020). I hypothesize that using either GPT-based and MIM-based text-generation techniques 90% of surveyed subject matter experts will not be able to reliably identify which MAUDE adverse events were written by humans and which were generated using a machine-generated approach. We will test this hypothesis by surveying subject-matter experts familiar with healthcare writing to see whether they can consistently determine which adverse events were written by humans, and which ones were written using an automated generation technique.

The movement towards open government provides bad actors hoping to use information warfare with an interface they can use for negative purposes, and that it is easy to leverage the volume of text with currently available text generation techniques to augment an information warfare campaign. To make it easier for consumers to report all sorts of incidents to the appropriate Attorney General, the US Government has a portal called Oversight.gov that can be used by consumers to report any malfeasance to the government (Council of the Inspectors General on Integrity and Efficiency, n.d.). Beyond MAUDE, Oversight.gov makes it possible for

a bad actor to use the same principles as the attack used in this project, open government data-sources and consumer reporting portals can be abused across a variety of sectors, including making false reports of fraud by energy companies, worker safety violations, illegal tobacco claims, accidental drug overdoses, housing discrimination, inappropriate behavior by mail carriers, etc.. Any of these attacks could lead to stretching US government oversight groups thin, or casting suspicion on people or organizations when it is not needed.

In this proposal, I am requesting financial support in order to explore this problem using a couple of different approaches. First, I would like to generate a false reports using a few different approaches, second, I would like to measure whether humans can consistently distinguish between human-written and machine-written reports, and finally, I would like to complete a document analysis to explore the possibility for automating this approach to create a long-term threat.

Section 2: Related Work

My work builds on work on text-generation, as well as information warfare. There have been multiple papers that have presented techniques for machine-generated text that is hard for humans to distinguish, including using the two primary techniques that will be used to generate the false adverse event reports, GPT and MIM, as well as research that puts information warfare campaigns into strategic context.

Machine-generated text has been widely studied and applied in the data science community, with many papers offering unique approaches to the problem. One 2019 paper that I found particularly interesting, by Liao et al., used a GPT based approach to generate Classical Chinese poetry, which was able to generate high-quality Classical Chinese poetry using a neural network. The network was initially trained on a large corpus of Chinese news articles, then trained on Classical Chinese poems. One thing that's interesting about this approach is that classical Chinese is a unique written language that is distinct from the vernacular Chinese that modern Chinese is based on, so they were able to use a large modern corpus, and fine-tune it with a collection using a totally different type of text for finally generating the samples. This approach is pretty similar, because the adverse events are typically written in a formal language, and even though lots of the language used in the reports is common, it's used in a precise way that often differs from typical English usage.

The MIM technique I will test in this paper builds on my research with DePaul's Divergent Design Lab, where we successfully used a botnet to make edits to Wikipedia articles. One of our applications was to erase the existence of Uyghurs using a Markov chain to make context relevant word swaps (Sharevski & Jachim, 2020). In this paper, a list of words was passed to the botnet, which were identified in text, and replaced using a Markov chain to identify words based on the words preceding the deletion. The Markov chain was trained on similar articles, like other articles on Chinese History for the Uyghur example, so one of the word substitutions replaced the word "Uyghur" with the word "Han," which is another Chinese ethnic group. This technique was used in the context of systematically censoring topics from public discourse, but with minor modifications it could be applied to replace words specific to one type of operation or device with words related to another.

The potential for machine-generated text to be mis-used has been studied, with the development of a few different approaches. One 2019 paper by Gehrmann et al., presented a

tool called GLTR that could help humans to identify machine-generated text through the use of a text visualization that highlighted words that were less-likely to appear in a sentence together, because machine-generated text relies on statistical relationships between words to figure out which words will go together. While the 2019 paper presenting GLTR demonstrated its success against GPT-based models, the MIM samples generated by WikipediaBot were mostly written by humans, and so using GLTR, the text was indistinguishable from text written by a human.

Another body of work that my project builds on approaches to information warfare and disinformation. In 2009, Eloff and Granova presented a line of reasoning is that the best information warfare defense is a strong offense, which enables an understanding of potential adversarial approaches, so that defenses can be built. They conceived of information warfare defenses consisting of a mix of legal and technical defenses to a wide range of different types of information attacks, including PsyOps (which is what they would consider my research), as well as technical attacks based around husting access to information like DDOS attacks. Another approach that they mention is that in some situations, a pre-emptive information warfare strike can help to reduce an adversary from being able to do the same to you.

Section 3: Research Design and Methodology

I am exploring the possibility for open government platforms to be abused for disinformation. To support this exploration, I will use a few different approaches, spanning across a couple different disciplines. First, I will collect text data from the MAUDE, and use it to build text generators, which I will use to generate text samples. Next, I will use human subjects to evaluate the machine generated text samples. Finally, I will perform a qualitative content analysis on the documentation surrounding MAUDE to evaluate the potential for this attack to be automated, and to review the scope of the damages that could be caused using such an attack.

Because the text generator will be trained on text created from previous adverse events, and there are enough adverse events to generate realistic seeming adverse event reports. Additionally, these reports include technical vocabulary, that is used precisely, and consistently, and is not common in normal parlance. This means that the resulting sentences are relatively likely to be consistent enough to create high-quality sentences, and the uniqueness of the terms will make it so that state of the art forensics tools based on word frequencies and presence of specific tools will not be adequate in identifying these sentences without extensive modification.

To create the false reports, which will be used to test the possibility for SMEs to distinguish between machine-generated and human-written adverse events, I will experiment with a couple of different methods for generating text samples. The first technique that I will use is a neural network-based approach to generate sequences based on a GPT. I will initially try training the model just using adverse event reports, but may switch to pre-training the model on a larger corpus written English before fine-tuning the model on the adverse event samples. This will likely follow a similar structure to the Classical Chinese generator project completed by Liao et al in 2019, when they used a massive corpus of general Chinese language texts before switching to using Classical Chinese poems.

An even simpler approach to text generation is the use of a Markov chain, which can be used to generate text samples based on the probability of a word being used after another word within a training dataset. These models are very simple, but in my previous experience, I have found that the word sequences are not always coherent. In the context of adverse event reports, however, the repetitiveness of the input data might mean that the model is able to pick up on

many of the relevant patterns, and it might be feasible to create passable adverse event reports. While humans will be more likely to identify the sentences as being machine-written, the presence of many uncommon words which the Markov chain is more likely to pick up than the neural network means that state of the art machine generated text identifiers are less likely to consider the text to be machine written, but I anticipate that a human will be able to easily tell that the text is not written by a human.

The last text samples that I will generate will use an ambient tactical deception-based approach, which will keep text from existing reports, and replace key words with substitutions identified using either the ANN-based approach or the Markov chain approach. Because the majority of the text will be human-written, except for key words which will be machine-selected. I do not expect that a human or a tool for identifying machine-written text will be able to correctly identify the text as being machine-written, because so little of the text is written by a human.

I anticipate that the automatically generated text samples can persuade humans that they are legitimate reports. To test this, I will survey 100 people on Amazon Mechanical Turk, and review their responses. Amazon Mechanical Turk allows careful selection of who will participate in each survey, if it is possible, I would like to limit the scope to people who work in healthcare administration, who will be familiar with technical writing about health, and will be the most likely to be able to identify inconsistencies in the generated samples (e.g. a rubber dam clamp that led to a patient having an air emphysema is not very likely). Additionally, I will load the text into GLTR with an explanation to see if GLTR combined with domain knowledge by the survey respondents will be able to successfully identify the responses.

The final element to this being a legitimate threat is for bad actors to be able to use this attack is for the adversary to be able to post the responses to MAUDE. Due to the critical nature of this system, it is not ethical or legal for me to make a false submission to the FDA. To see what defenses are in place, I will perform an extensive document review to see what the submission interface is like for members of the public. This will include reviewing the HTML in the submission interface to see if the FDA has built any defenses against scripting tools like Selenium or other approaches that could fully automate this attack. This analysis will help to put the remainder of my project into context of what is possible mechanically in terms of actually weaponizing an attack like this, as well as the ramifications.

Success in confirming the attack will come primarily form the survey responses, where individuals will try to determine whether a sample of text was human written or not. If the majority of respondents cannot differentiate, then that will highlight a major gap in our ability to identify machine-written technical reports.

Section 4: Plan of Work and Outcomes[1]

Assuming my ability to obtain funding, I anticipate completing the project over the course of three and a half months, to include all steps from downloading data, to typing up and editing a final paper.

[1] All date estimates assume a start date of July 1st

To provide a more lucid picture of the steps I hope to complete, I have listed each of the steps that I will complete, along with the date range in which I hope to complete them (assuming I secure funding 2021-07-01).

NO.	PROJECT PHASE	TASK	COMPLETE ON	PREDECESSOR
1-1	Project Prep	Data downloaded, structured into PostgreSQL database.	7/3/2021	
1-2		FDA notified of intentions.	7/5/2021	
1-3		Initial literature review.	7/14/2021	
1-4		IRB submission document prepared.	7/14/2021	1-3
2-1	Attack Legwork	Document analysis on MAUDE documentation, including outlining processes for submitting documents to MAUDE, requirements, etc.	8/15/2021	
2-2		Code written and executed for text generation.	8/1/2021	1-1
2-3		Samples of generated adverse events created	8/15/2021	2-2
3-1	Survey	Survey created based on the generated text samples.	8/20/2021	1-4, 2-3
3-2		Pre-flight survey	8/23/2021	3-1
3-3		Survey adjustments	8/26/2021	3-2
3-4		Survey posted to Amazon Mechanical Turk	8/26/2021	3-3
3-5		Results collected and analyzed	9/15/2021	3-4
4-1	Results Analysis and Document Writing	Notes taken throughout process compiled into outline, rewritten in more formal style.	9/22/2021	1-*, 2-*, 3-*
4-2		Analysis written.	9/29/2021	4-1
4-3		Editing.	10/6/2021	4-2
4-4		Submission.	10/16/2021	4-3

My primary deliverable will be a completed paper. The paper will outline the risks associated with open government, and will hopefully initiate discussions that can help ensure the security and availability of these resources.

Assuming a project start date of 2021-07-01, and all intermediary steps being completed by the dates listed in the time above, I anticipate that the project will be complete by 2021-06-10.

Section 5. Conclusions and Future Work

If funded, I believe that this project can help to promote adversarial thinking with respect to open government databases. One situation where open government does not work is when the resources provided in the open government threaten national security (Ubaldi, 2013), and while the threats I identify do not necessarily directly threaten national security, the possibility for similar approaches to be used in other situations could lead to threats to national security, which would increase the likelihood of citizens no longer being granted to these government resources. In addition to finding ways to misuse open government data, I would ultimately like to develop approaches for protecting these resources from being misused.

Section 6: References

Council of the Inspectors General on Integrity and Efficiency. (n.d.). *Oversight.gov* [Government Website]. Oversight.Gov. Retrieved June 12, 2021, from https://www.oversight.gov/

Eloff, J., & Granova, A. (2009). Chapter 39—Information Warfare. In J. R. Vacca (Ed.), *Computer and Information Security Handbook* (pp. 677–690). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-374354-1.00039-X

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *ArXiv:1906.04043 [Cs]*. http://arxiv.org/abs/1906.04043

Liao, Y., Wang, Y., Liu, Q., & Jiang, X. (2019). GPT-based Generation for Classical Chinese Poetry. ArXiv:1907.00151 [Cs]. http://arxiv.org/abs/1907.00151

MacFarlane, D., Hurlstone, M. J., & Ecker, U. K. H. (2020). Protecting consumers from fraudulent health claims: A taxonomy of psychological drivers, interventions, barriers, and treatments. *Social Science & Medicine*, *259*, 112790. https://doi.org/10.1016/j.socscimed.2020.112790

OECD. (n.d.). Open Government Data—OECD [NGO]. OECD.Org. Retrieved April 25, 2021, from https://www.oecd.org/gov/digital-government/open-government-data.htm

Pauly, V., Frauger, E., Pradel, V., Rouby, F., Berbis, J., Natali, F., Reggio, P., Coudert, H., Micallef, J., & Thirion, X. (2011). Which indicators can public health authorities use to monitor prescription drug abuse and evaluate the impact of regulatory measures? Controlling High Dosage Buprenorphine abuse. *Drug and Alcohol Dependence*, *113*(1), 29–36. https://doi.org/10.1016/j.drugalcdep.2010.06.016

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. OpenAI (PrePrint). https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Sharevski, F., & Jachim, P. (2020). WikipediaBot: Automated Adversarial Manipulation of Wikipedia Articles. *ArXiv:2006.13990 [Cs]*. http://arxiv.org/abs/2006.13990

Ubaldi, B. (2013). *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*. https://doi.org/10.1787/5k46bj4f03s7-en

US FDA. (n.d.). *MAUDE - Manufacturer and User Facility Device Experience* [Government Website]. Retrieved April 24, 2021, from https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm

Appendix: Budget and Budget Narrative

To estimate the costs associated with each line item, I will use PERT, which is a technique for estimating quantities based on a weighted average of a best estimate, the best case scenario, and the worst case scenario.

Survey Responses, $1,243.38

Item	Best Case	Best Estimate	Worst Case
Reward	$5.00	$7.00	$10.00
Amazon's Cut	$1.00	$1.40	$2.00
US bachelor's degree	$0.50	$0.50	$0.50
Work full-time	$0.35	$0.35	$0.35
Healthcare	$0.40	$0.40	$0.40
Administrator	$0.35	$0.35	$0.35
Cost per Participant	$7.60	$10.00	$13.60
Pilot Participants	100	100	100
Participants	15	15	15
Total (Before Tax)	$874.00	$1,150.00	$1,564.00
Tax	$52.44	$69.00	$93.84
Total (After Tax)	$926.44	$1,219.00	$1,657.84

To pay for the surveys to be completed, I have to create a reward for each participant, which I will set at $5.00, then Amazon gets a cut, which will cost 20% of the participant reward. Additionally, Amazon charges for each filter for who will take the survey. To make the paper as robust as possible, I would like for the people reviewing the survey to be subject matter experts in healthcare who are familiar with reading technical health-related reports. So, ideally, I would like the survey respondents to be people who have a US bachelor's degree (+$0.50), work full-time (+$0.35) in healthcare (+$0.40) administration (+$0.35).

Due to the limited pool of people who fit into all of those filters and are responding to surveys on Amazon Mechanical Turk, there is a chance that I need to increase the cost of the reward, perhaps to $7.00 0r $10.00.

With 100 participants, and a 15-person pilot study, I estimate that the total cost will be $1,243.38.

Compute Power, $151.58

Item	Best Case	Best Estimate	Worst Case
Hours	10	40	90
Amazon EC2 (f1.4xlarge)	3.3	3.3	3.3
Total (Before Tax)	$33.00	$132.00	$297.00
Tax	$1.98	$7.92	$17.82
Total (After Tax)	$34.98	$139.92	$314.82

Ideally, by using additional compute power, the training will be done after about 10 hours, but it is possible for the training to take dramatically longer, or for the training to have to be totally repeated due to a technical error, or if the model's quality is lacking.

In total, the weighted estimate for these different possibilities is $151.58.