Extended Abstract: Leveraging Large Language Models to Identify Internet Censorship through Network Data

Tianyu Gao
Graduate Center, City University of New York
New York, NY, USA
tgao@gradcenter.cuny.edu

Ping Ji
Graduate Center, City University of New York
New York, NY, USA
pji@gc.cuny.edu

ABSTRACT

With the intensification of internet censorship measures implemented across the globe, detecting and analyzing censorship events have become crucial for preserving Internet freedom. The network reachability data collected by existing censorship monitoring platforms such as OONI [5], Censored Planet [8], and ICLab [7] offers a comprehensive and longitudinal view of global censorship efforts but presents challenges in data analysis with growing data volume and complexity. In this study, we aim to navigate a novel approach to detect internet censorship by applying Large Language Models (LLMs) to analyze network reachability data, with the goal of mitigating challenges posed by the massive volumes of data collected by the platforms. We explore the potential of LLMs, such as GPT-3 [2] and BERT [3], on processing and interpreting the extensive, complex network data to identify patterns indicative of censorship. By integrating innovative LLM methodologies with data collected from existing censorship monitoring platforms, we hope to enhance the accuracy and scalability of censorship detection, so that more robust defense strategies can be developed.

KEYWORDS

Internet Censorship Detection, Large Language Models, Anomaly Detection, Pattern Recognition

1 INTRODUCTION

This work aims to innovate in the domain of internet censorship detection by applying Large Language Models (LLMs) to analyze network data. Leveraging the proven success of LLMs in complex Natural Language Processing (NLP) tasks, we propose a novel approach to identify censorship patterns within the network measurement data collected by internet censorship monitoring platforms such as Censored Planet [8], ICLab [7], and OONI [5].

The challenge of detecting internet censorship has become increasingly complex with the evolution of the internet, as censors adapt to technological advancements by employing more sophisticated techniques to control and restrict access to information. Currently, the predominant methodologies for detecting internet censorship rely largely on rule-based systems. These systems identify censorship by comparing measurement data with censorship fingerprints, which are usually regular expressions designed by human experts [4]. While computationally efficient, these rule-based approaches face significant limitations. The predefined rules for identifying censorship are less adaptable to the evolving techniques used by censors, leading to potential oversights of network anomalies and false negatives. As the volume of network data continues to expand, rule-based systems struggle to scale effectively. Their performance can degrade under the weight of processing vast datasets, in contrast to learning-based models, which thrive on large data volumes. Additionally, rule-based mechanisms lack the nuanced understanding of context that learning-based approaches can offer, which is crucial when attempting to detect sophisticated censorship techniques that do not trigger predefined rules.

Large Language Models (LLMs) have transformed the field of Natural Language Processing (NLP) by demonstrating exceptional capabilities in understanding and generating natural language. This has led to significant advancements in tasks such as machine translation, sentiment analysis, and question-answering systems. The success of LLMs is largely attributed to their ability to leverage vast amounts of unlabeled data to learn complex language patterns and semantics. Our research builds on this foundation, adapting LLMs to analyze network reachability data to uncover patterns of internet censorship with precision and depth. The rationale for this approach is based on the similarities between the scale, structure, and semantic richness of NLP corpora and network data. [6]

The existing internet censorship monitoring platforms handle massive volumes of data daily, much of which remains unlabeled. While a prominent language model like BERT [3] was trained on 16 GB of the Books Corpus and English Wikipedia, the measurement data of internet censorship available for analysis often surpasses this in daily volume alone. Network reachability data is not only ample but also semantically rich, mirroring the complexity and nuances found in natural language. Platforms like ICLab collect detailed network data, which, similar to text in a language model, contains detailed and contextually meaningful information, ranging from comprehensive metadata to protocol analysis. The structure and content of network data, from packet field category to DNS responses, carry significant semantic meaning. These can be grouped into meaningful clusters analogous to how words and sentences are treated in the field of NLP. By drawing parallels between the scales and semantic-rich aspects of both domains, this work aims to advance the field of internet censorship detection using learning-based approaches.

2 RELATED WORKS

Historically, research in the field of internet censorship detection has primarily focused on rule-based systems, which often struggle with the dynamic nature of censorship techniques and the vast

Free and Open Communications on the Internet 2024 (2), 10–12

amounts of data involved. However, with the rapid developments in artificial intelligence, the shift toward more adaptive and scalable learning-based models has proven to be more effective.

For example, [10] introduced CenDTect, a system that utilizes decision trees and a novel clustering method to analyze data collected from Censored Planet [8], along with the OONI [5] dataset, as on-the-ground confirmation. CenDTect effectively identifies patterns of blocking policies and provides interpretable insights into censorship dynamics at both local and country levels.

Furthering the application of advanced NLP techniques to network data, [4] proposed a system using a sequence-to-sequence model to analyze the Censored Planet[8] dataset as sequential data. This system also employs an image classification approach by treating network reachability data as grayscale images and utilizing Convolutional Neural Networks (CNNs) to distinguish between censored and uncensored content.

Moreover, [1] utilized a combination of supervised and unsupervised machine learning models to detect DNS-based internet censorship from datasets collected by Satellite [9] and OONI [5]. Their approach involved training supervised models on expert-derived labeled data to refine anomaly detection heuristics, achieving high true positive rates, while unsupervised models were used to establish a baseline of "normal" behavior to identify censorship anomalies. This method has been effective in detecting both known and new instances of DNS censorship.

The exploration of Large Language Model in this context is still in its early stages. Leveraging their success in NLP tasks, LLMs such as BERT[3] and GPT-3[2] are believed to be competent for understanding the complex and often subtle signs of censorship embedded within large-scale network data. Their ability to process and interpret vast, unlabeled datasets can potentially uncover patterns of censorship that are not detectable by traditional models. As this research progresses, we will delve deeper into how these models are being adapted for our purposes, compare their efficacy with these innovative approaches, and discuss their potential to reshape the landscape of censorship detection technology, just like they have in the field of NLP.

3 HYPOTHESES

This research is built on the hypotheses that LLMs can detect the patterns indicative of internet censorship within data collected by monitoring platform.

Data Pattern Recognition LLMs will identify and categorize network anomalies as potential censorship activities. To validate this hypothesis, we will train LLMs on a curated dataset comprising both normal internet traffic and documented instances of censorship. LLMs are expected to adapt to recognizing patterns in these datasets. In the context of internet censorship detection, this means analyzing network data, such as packet headers, timing information, and payload characteristics, to distinguish between regular traffic patterns and anomalies that may indicate censorship. We hypothesize that LLMs trained on extensive network logs and known instances of censorship can learn to identify subtle signs of information blocking or throttling. Features like traffic volume spikes, unusual request-response delays, and the presence of reset packets are used as indicators of censorship.

Temporal and Geographical Analysis The models will reveal temporal and geographical patterns of censorship, providing insights of the measures used by different regimes. We employ LLMs to analyze the timing and origin of network traffic, correlating these with known political or social events. By training the models on longitudinal data from various regions and times, they can learn to recognize the temporal and geographical patterns associated with censorship activities.

Interpretability Techniques like attention mechanisms within LLMs can help in tracing the decision-making process of the model. By examining which features of the data the model focuses on when making predictions, researchers can gain insights into the reasons behind its conclusions.

4 PROPOSED FUTURE WORKS

In this work, we plan to employ Large Language Models like BERT (Bidirectional Encoder Representations from Transformers) [3] and GPT-3 (Generative Pretrained Transformer-3) [2] to identify internet censorship through network data analysis.

Datasets Our preliminary choice of dataset is the ICLab [7] dataset because its detailed data collection across all levels of the network stack. The rich semantic content of ICLab's network data would enhance the learning capabilities of Large Language Models. Although ICLab may not have the same scales and coverage as OONI [5] or Censored Planet [8], its detailed, context-rich data offers a playground for LLMs to explore. To further enhance the accuracy of our findings and reduce potential false negatives, we plan to use data from OONI and Censored Planet to cross-reference results obtained from ICLab. This will help validate our model predictions, ensuring the precision of our censorship detection approach.

Data Preprocessing We plan to transform raw network traffic data into a structured format suitable for LLMs analysis. This step involves extracting useful features from the data. These features serve as indicators of normal and potentially anomalous network behaviors. Following extraction, the data is tokenized just as it is in natural language processing (NLP), allowing the LLMs to process and analyze the network anomalies as if they were textual data.

Choice of Models Our preliminary choice of BERT and GPT-3 is based on their strengths in understanding and generating context. Additionally, these models are well-supported by extensive research. BERT's bidirectional processing capability is ideal for identifying patterns of censorship within sequential network data. GPT's extensive training on a diverse array of internet texts enables it to understand a broader context and predict potential censorship activities. Further research will be conducted in this area.

Model Training The training process will incorporate both supervised and unsupervised learning methods. In supervised learning, we train the models on datasets with known instances of censorship, enabling them to distinguish between censored and uncensored network activities. Unsupervised learning, in contrast, enables models to autonomously discover clusters or patterns in the data that indicate censorship, without the labels.

REFERENCES

[1] Jacob Brown, Xi Jiang, Van Tran, Arjun Nitin Bhagoji, Nguyen Phong Hoang, Nick Feamster, Prateek Mittal, and Vinod Yegneswaran. 2023. Augmenting Rule-based DNS Censorship Detection at Scale with Machine Learning. arXiv:2302.02031 [cs.LG]

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[4] Shawn P. Duncan and Hui Chen. 2023. Detecting network-based internet censorship via latent feature representation learning. *Computers & Security* 128 (May 2023), 103138. https://doi.org/10.1016/j.cose.2023.103138

[5] Arturo Filasto and Jacob Appelbaum. 2012. OONI: Open Observatory of Network Interference. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*.

[6] Franck Le, Mudhakar Srivatsa, Raghu Ganti, and Vyas Sekar. 2022. Rethinking Data-driven Networking with Foundation Models: Challenges and Opportunities. arXiv:2211.06494 [cs.NI]

[7] Arian Akhavan Niaki, Shinyoung Cho, Zachary Weinberg, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, and Phillipa Gill. 2020. ICLab: A Global, Longitudinal Internet Censorship Measurement Platform. In *2020 IEEE Symposium on Security and Privacy (SP)*. 135–151. https://doi.org/10.1109/SP40000.2020.00014

[8] Ram Sundara Raman, Prerana Shenoy, Katharina Kohls, and Roya Ensafi. 2020. Censored Planet: An Internet-wide, Longitudinal Censorship Observatory. In *Proceedings of the ACM SIGCOMM 2020 Conference*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3372297.3417883

[9] Will Scott, Thomas Anderson, Tadayoshi Kohno, and Arvind Krishnamurthy. 2016. Satellite: Joint Analysis of CDNs and Network-Level Interference. In *USENIX Annual Technical Conference*.

[10] Elisa Tsai, Ram Sundara Raman, Atul Prakash, and Roya Ensafi. 2024. Modeling and Detecting Internet Censorship Events. *Network and Distributed System Security (NDSS) Symposium* (2024).