Training Data Voids: Novel Attacks Against NLP Content Moderation

Jordan S. Huffaker
Computer Science & Engineering
University of Michigan, Ann Arbor
jhuffak@umich.edu

Jonathan K. Kummerfeld
Computer Science & Engineering
University of Michigan, Ann Arbor
jkummerf@umich.edu

Walter S. Lasecki
Computer Science & Engineering
Center for Hybrid Intelligence Systems
University of Michigan, Ann Arbor
wlasecki@umich.edu

Mark S. Ackerman
Computer Science & Engineering
School of Information
University of Michigan, Ann Arbor
ackerm@umich.edu

ABSTRACT

Machine learning-based content moderation systems make classification decisions by leveraging patterns learned from training data. However, patterns that are under or unrepresented in a system's training data —which we call *training data voids*— cannot be learned and may be exploited by adversarial users to confuse the system. Specifically, adversarial users may creatively construct harmful content that differ from known training examples, leading to uncertain classification. Here, we call this type of attack against machine learning classifiers a *novelty attack* and distinguish it from a more widely known class of attacks (i.e., *adversarial attacks*). Additionally, we contribute a study design for exploring the extent to which novel harmful content can be constructed and for characterizing the effects of novelty on classification results in several text-based content moderation

CSCW '19, Nov. 09–13, 2019, Austin, TX

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

domains. The findings of this study are important for highlighting a potential vulnerability of machine learning-based content moderation systems and may suggest that such systems will remain limited in the near future. We propose that including human intelligence in content moderation systems may be an effective approach for mitigating potential exploitation.

CCS CONCEPTS

• **Human-centered computing → Computer supported cooperative work**.

KEYWORDS

content moderation, machine learning, natural language processing, data voids, novelty attacks

ACM Reference Format:
Jordan S. Huffaker, Jonathan K. Kummerfeld, Walter S. Lasecki, and Mark S. Ackerman. 2019. Training Data Voids: Novel Attacks Against NLP Content Moderation. In *CSCW '19: CSCW Workshop on Volunteer Work: Mapping the Future of Moderation Research, Nov. 09–13, 2019, Austin, TX*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

INTRODUCTION

Embedding machine learning into content moderation systems could greatly reduce the human labor required to moderate content at scale. However, the extent to which machine learning can accurately moderate user generated content in-the-wild is largely still unknown. Some have proposed that machine moderators may replace human moderators altogether [1]. Others believe that machine learning is still limited [8], that it may remain limited [2], or that its applications should be constrained to supporting human moderators existing practices [4]. In this study design, our goal is to examine a specific vulnerability that exists within machine learning-based content moderation systems so that systems builders can be more informed when building future systems.

Specifically, we explore the unique challenges associated with designing machine learning-based systems to counter the efforts of adversarial users who are highly motivated to find and exploit vulnerabilities. Adversarial users often have a broader agenda that involves shaping the content other users are exposed to when browsing platform [13]. For example, in their study of search engine algorithms Golebiewski and boyd found that adversarial users exploit "data voids"[1] —search terms where available relevant data is limited or non-existent— by introducing problematic search results that fill the void for those search terms [7]. Filling data voids at key moments can result in innocuous information seekers finding exclusively misinformation, extremist pages, and other harmful content when searching the terms.

In content moderation settings, a similar type of attack may be used to exploit *training data voids*, content that lacks enough similar training data to be accurately classified by a machine learning

[1]Golebiewski and boyd provide examples search terms like "black on white crime", "Harrold-Oklaunion", and "Sutherland Springs", which all initially had very little data associated with them to return in search results. Adversarial actors capitalized on these terms by using them in problematic content and posting to Twitter, reddit, and other websites. Afterward, searching these terms yielded the problematic content desired by the actors.

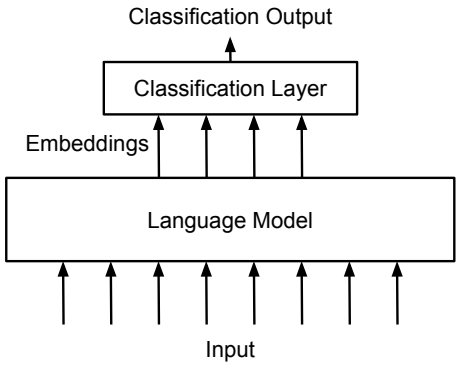

Figure 1: State-of-the-art classifiers for unstructured text typically leverage language models trained on massive collections of text. These language models provide embeddings, representations of input words and phrases, based on their patterns of usage in the language model's training data. The embeddings are used as the input layer to a fine-tuned classifier trained with another dataset that is task specific. We will manipulate the training datasets for both layers in our study.

classifier. Training data voids differ from *blindspots* (see [16] for a formal definition) in that blindspots refer to aspects of a trained model whereas training data voids refer to aspects of training data. State-of-the-art approaches to classifying unstructured text make classification decisions by leveraging language patterns learned from labeled datasets. However, even with a massive dataset, it is nearly impossible to capture all possible ways that language can be used to communicate hate speech, harassment, and other harmful messages. Even known phrases can be permuted in such a way that they differ substantially from examples seen in training data, yet carry the same meaning (e.g., permuting "political puppet" into "political muppet"). Therefore, machine learning-based classifiers may be vulnerable to *novelty attacks*, where adversarial users leverage novel language patterns to hide rule-violating content within a training data void.

Novelty attacks are distinguished from adversarial attacks [14, 19] in that adversarial attacks leverage information about the classification model to cause it to behave a particular way (i.e., control classification output, see [6] for examples) whereas novelty attacks target the data used to train the model, leveraging undiscovered patterns to create confusion. In many content moderation settings, uncertain classifications default to assume the best intentions of the poster, thus ensuring attacker's content end up in front of other users. Additionally, while some prior work has explored vulnerabilities to adversarial attacks, novelty attacks remain unstudied despite the potential risks of deploying machine learning-based systems in sensitive settings such as hate speech and harassment moderation. Like other sociotechnical systems deployed in-the-wild (e.g., crowdsourcing [10, 11], information retrieval [7], and social media [17]), content moderation systems must consider and respond to potential threats. To understand the extent of this vulnerability, we present a study design that aims to uncover how much novelty is needed to deceive machine learning-based content moderation systems. Our study is guided by two research questions:

(1) How can adversarial actors use novelty to attack a machine learning-based classifier?
(2) How does the novelty of test examples affect classification confidence and accuracy?

STUDY DESIGN

In the first part of our study, we plan to perform a qualitative analysis of several labeled text datasets for various moderation tasks (e.g., hate speech detection [3], offensive language detection [18], personal attack detection[2], etc.), searching specifically for novel ways to construct harmful messages. We will use the themes discovered through our analysis to create a novelty scale that can be used to rate specific examples by the difficulty to generate them. A low novelty rating would represent trivial permutations of messages, and a high novelty rating would represent examples that elicit never-before-seen meaning from otherwise known words.

[2]https://www.kaggle.com/jigsaw-team/wikipedia-talk-labels-personal-attacks#attack_annotated_comments.csv

Additionally, we will answer the second research question by synthetically creating training data voids within training data, then by evaluating classifier performance with a static test dataset. We design our procedure to simulate conditions of a real-world classifier by systematically removing training data matching pre-defined "language patterns" (examples that use a particular creative format as identified in the first part of the study), enabling us to control both the amount of relevant training data available for a specific test example and the amount of novelty in the test examples. We plan to bootstrap our classifier with GloVe [15], as it can be retrained in a reasonable amount of time and it comes close to the performance of more massive language models (e.g., BERT [5], Figure 1). Therefore, our characterization consists of measuring classifier performance after varying three variables: 1) number of relevant training data examples used to train the language model, 2) number of relevant training data examples used to train the fine-tuned classifier, and 3) novelty of test examples.

IMPLICATIONS

We expect to find that content moderation systems are vulnerable (at least to some degree) to novelty attacks, a finding that would have implications for the design of future content moderation systems. Importantly, existing approaches for evaluating system performance with static test and training datasets may not be sufficient for gauging how robust a system is to exploitation. A test procedure that varies the novelty of input may be more apt at evaluating in-the-wild performance.

More generally, vulnerability to novelty attacks represent the manifestation of a "socio-technical gap" between the capabilities of content moderation systems and the varied and changing needs of people in their application domains [2]. Including the intelligence of people in these systems may be a path to bridge this gap, as people are better suited to evaluate novel content. We hope that the findings of this study will serve as motivation for the development of hybrid intelligence systems that leverage the intelligence of both humans and machines. In other domains, prior work has shown that these systems can provide the intelligence of people at the scale and speed of automated approaches [9, 12]. We hope that similar systems will be developed for content moderation.

ACKNOWLEDGMENTS

This research was supported in part by the National Aeronautics and Space Administration (NASA) and the Defense Advanced Research Projects Agency (DARPA). Both NASA and DARPA provide funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not NASA, DARPA, or any other government entity.

REFERENCES

[1] 2018. Transcript of Mark Zuckerberg's Senate Hearing. *Washington Post* (Apr 2018). https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/

[2] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[3] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 54–63.

[4] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. (2019).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1625–1634.

[7] Michael Golebiewski and danah boyd. 2018. Data voids: Where missing data can easily be exploited. *New York: Data & Society Research Institute* (2018), 1–6.

[8] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[9] Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. 2011. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 23–32.

[10] Walter S Lasecki, Jaime Teevan, and Ece Kamar. 2014. Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 248–256.

[11] Walter S Lasecki, Jaime Teevan, and Ece Kamar. 2015. The cost of asking crowd workers to behave maliciously. In *Proc. the AAMAS Workshop on Human-Agent Interaction Design and Models*.

[12] Alan Lundgard, Yiwei Yang, Maya L Foster, and Walter S Lasecki. 2018. Bolt: Instantaneous crowdsourcing via just-in-time training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 467.

[13] Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute* (2017).

[14] Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355* (2019).

[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[16] Ramya Ramakrishnan, Ece Kamar, Besmira Nushi, Debadeepta Dey, Julie Shah, and Eric Horvitz. 2019. Overcoming Blind Spots in the Real World: Leveraging Complementary Abilities for Joint Execution. (2019).

[17] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. (2019).

[18] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983* (2019).

[19] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326* (2018).