Security Challenges in Natural Language Processing Models

Qiongkai Xu[1,3] and **Xuanli He**[2]
[1] School of Computing and Information System, the University of Melbourne, Australia
[2] Department of Computer Science, University College London, United Kingdom
[3] School of Computing, FSE, Macquarie University, Australia
qiongkai.xu@mq.edu.au, xuanli.he@ucl.ac.uk

Abstract

Large-scale natural language processing models have been developed and integrated into numerous applications, given the advantage of their remarkable performance. Nonetheless, the security concerns associated with these models prevent the widespread adoption of these black-box machine learning models. In this tutorial, we will dive into three emerging security issues in NLP research, i.e., backdoor attacks, private data leakage, and imitation attacks. These threats will be introduced in accordance with their threatening usage scenarios, attack methodologies, and defense technologies.

1 Tutorial Content

1.1 Introduction

Large-scale natural language processing models have recently garnered substantial attention due to their exceptional performance. This promotes a significant proliferation in the development and deployment of black-box NLP APIs across a wide range of applications. Simultaneously, an expanding body of research has revealed profound security vulnerabilities associated with these black-box APIs, encompassing issues such as dysfunctional failures (Gu et al., 2017; Dai et al., 2019; Huang et al., 2023), concerns related to privacy and data leakage (Coavoux et al., 2018; Carlini et al., 2021), and infringements on intellectual property (Wallace et al., 2020; Xu et al., 2022). Those security challenges can lead to issues like data misuse, financial loss, reputation damage, legal disputes, and more. It is worth noting that these security vulnerabilities are not mere theoretical assumptions. Previous research has demonstrated that both commercial APIs and publicly available models can be easily compromised (Wallace et al., 2020; Carlini et al., 2021; Xu et al., 2022). This tutorial aims to provide a comprehensive overview of the latest research concerning security challenges in NLP models.

1.2 Security Challenges in NLP

This section will delineate three prevalent security challenges encountered in NLP research and applications. These include (1) backdoor attacks, (2) privacy concerns and data leakage, and (3) imitation attacks. For each of these challenges, we will first commence by introducing their threat model in real-world applications. Subsequently, we will delve into the techniques used to execute these attacks, illustrating their impact on vulnerable applications. Finally, we will discuss the countermeasures and defense technologies available to mitigate these attacks.

Adversarial and Backdoor Attacks. Our discussion commences with adversarial attacks in the context of NLP tasks. These attacks involve the manipulation of inputs to compromise the performance of a target model (Alzantot et al., 2018; Ebrahimi et al., 2018; Li et al., 2018). More specifically, by altering specific characters or words, it becomes possible to deceive a text classifier into assigning an incorrect label. This research underscores the inherent vulnerability of trained NLP models. A notable subset of these attacks is the backdoor attack, where the victim model is induced to associate misbehavior with specific triggers (Dai et al., 2019). During the inference stage, poisoned models exhibit normal behavior on clean inputs, but their misbehavior is triggered when malicious patterns are presented. Those malevolent actions can range from deceiving text classifiers (Dai et al., 2019; Kurita et al., 2020) to mistranslating neutral phrases into controversial ones (Xu et al., 2021).

In the literature, there exist two primary strategies for embedding backdoor triggers: (1) data poisoning and (2) weight poisoning. Data poisoning seeks to infiltrate triggers into a victim model by poisoning a small fraction of the training data,

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 7–12

as demonstrated in various studies (Dai et al., 2019; Chen et al., 2021; Qi et al., 2021b; Wang et al., 2021; Xu et al., 2021). Regarding weight poisoning, attackers surreptitiously integrate the triggers into the victim model's weights (Kurita et al., 2020; Li et al., 2021; Yang et al., 2021a) or their embedding dictionary (Huang et al., 2023). It is noteworthy that the majority of backdoor attacks have centered on supervised learning. However, with the growing prominence of instruction tuning (Ouyang et al., 2022; Wei et al., 2022), we will delve into the manipulation of large language models through instruction tuning poisoning in subsequent discussions (Wan et al., 2023; Xu et al., 2023; Shu et al., 2023).

In conjunction with the literature on backdoor attacks, we will cover multiple defensive approaches that aim at mitigating the vulnerabilities caused by these attacks. Depending on the level of access to the training data, these defensive measures can be categorized into two types: (1) *training-stage* defense and (2) *test-stage* defense. The former method aims at identifying poisoned data by analyzing the anomalous characteristics of the training data (Sun et al., 2021; He et al., 2023b). The latter approach leverages external tools (Qi et al., 2021a) or the victim language models themselves (Yang et al., 2021b; Chen et al., 2022; He et al., 2023a) to either remove the triggers or entirely discard the poisoned data samples during the inference.

Privacy and Data Leakage. Another challenge in NLP models is the potential risk of disclosing data, particularly sensitive content, to untrustworthy parties. A recent widely recognized example is the capability of pre-trained language models, e.g., GPT-2, to generate sentences containing sensitive information when provided with carefully designed prompts (Carlini et al., 2021). Another concern revolves around the possibility that certain information from the training data is inferred through the model's parameters or the gradient updates, such as membership inference and text data recovery (Melis et al., 2019; Gupta et al., 2022). These types of attacks pose significant challenges to collaborative learning of language models (Yang et al., 2019).

Privacy and data leakage present a contentious challenge in NLP models. In this discussion, we will introduce technologies aimed at addressing these concerns, including (1) unlearning specific private training data, known as machine unlearning (Bourtoule et al., 2021), (2) methods for identifying the generated outputs that may contain sensitive attributes (Xu et al., 2020) and (3) techniques that obscure the intermediate representation of NLP models, such as the application of differential privacy (Lyu et al., 2020; Shi et al., 2022).

Imitation Attack. The final security challenge within our scope will be the imitation attack on NLP models. With the advancement of NLP models, particularly large pre-trained language models, companies have encapsulated exceptional models into commercial APIs, serving millions of end-users. In order to foster a profitable market, service providers commonly implement pay-as-you-use policies for those APIs. To circumvent service charges, a seminal work (Tramèr et al., 2016) proposed the imitation of the functionality of commercial APIs by relying on predictions from those APIs. Subsequent research has revealed vulnerabilities associated with imitation attacks that extend beyond the violation of intellectual property, e.g., one can employ the imitation model to craft transferable adversarial examples capable of deceiving the victim model as well (Wallace et al., 2020; He et al., 2021). Moreover, the interaction between the victim model and the imitator can lead to significant privacy breaches (He et al., 2022a). Furthermore, Xu et al. (2022) demonstrate that imitation models can outperform the imitated victim models, particularly in the context of domain adaptation and model ensemble.

Several studies have devised a range of defensive strategies to mitigate those security threats. Given that imitation attacks depend on the predictions made by victim models, one straightforward solution involves manipulating these predictions such that the imitation models are trained with partial or potentially deceptive information. We will delve into the details of how this has been achieved in text classification and generation problems, including techniques such as customizing and perturbing predicted label distributions (Xu et al., 2022; He et al., 2022a). Additionally, we will explore recent advancements in watermarking technologies for intellectual property protection (Krishna et al., 2020; He et al., 2022b,c; Zhao et al., 2023)

2 Relevance and Importance to Computational Linguistic Community

Large-scale language models have achieved significant performance in many NLP tasks, with many

applications now reliant on those advanced NLP models. However, any uncontrolled misconduct, the inadvertent disclosure of private training data, or potential leaks of model intellectual property could result in substantial financial and social consequences. The imperative to guide the future development of NLP models is shifting from mere task performance to a growing emphasis on the security and ethical concerns of these models. Machine learning models, especially large-scale deep learning models, remain somewhat inscrutable to human comprehension. This opacity raises the challenges in identifying and addressing potential risks associated with these models without comprehensive explanations and a deep understanding of their inner workings. In order to inspire broader discussion and foster research efforts in the domain of security in NLP, this tutorial is dedicated to presenting the principle security challenges in modern natural language processing models. This will include exploration of their threat models, attack methodologies, and defense technologies.

3 Tutorial Information

Tutorial Outline The tutorial is expected to be 3.5 hours, including a half-hour coffee break.

1. Introduction (15 mins)
2. Backdoor Attack (50 min, by Xuanli He)
   (a) Problem definition and motivation;
   (b) Adversarial and Backdoor Attacks on NLP models;
   (c) Defense techniques against backdoor attacks.
3. Privacy and Data Leakage (50 min, by Qiongkai Xu)
   (a) Problem definition and motivation;
   (b) Privacy Leakage in NLP models;
   (c) Data Leakage in NLP models;
   (d) Defense techniques against privacy and data Leakage.
4. Imitation Attack (50 min, by Qiongkai Xu and/or Xuanli He)
   (a) Problem motivation and definition;
   (b) Imitation attack and subsequent attacks;
   (c) Defense techniques against imitation attack.
5. Conclusion and Future Trends (15 mins)

Topic Breadth. Our expectation is that approximately 30% of the content will be drawn from the work of the instructors, while the remaining 70% will be sourced from contributions made by various other researchers. The materials we intend to cover include papers from both academia and industry.

Ethical Considerations. In this tutorial, we shed light on various vulnerabilities found in contemporary NLP models. Our intention in discussing these vulnerabilities is not to endorse any form of attack. Rather, our objective is to emphasize the importance of responsible AI practices in both academic and industrial contexts. Through this approach, we can harness the progress made in AI while concurrently upholding security, privacy, and ethical considerations.

Open Accessibility. We intend to ensure that all instructional materials are available online.[1] Moreover, we grant permission to include slides and video recordings in the ACL anthology.

4 Prerequisites for the Attendees

This tutorial is designed to cater to the needs of both NLP researchers and students in academia, as well as industrial practitioners with an interest in security & privacy in NLP, model explanation, and related areas. While a basic understanding of Machine Learning is beneficial, it is not an obligatory prerequisite.

5 Reading List

- Backdoor Attack (Gu et al., 2017; Dai et al., 2019; Kurita et al., 2020)
- Privacy and Data Leakage (Melis et al., 2019; Carlini et al., 2021; He et al., 2022a)
- Imitation Attack (Wallace et al., 2020; Xu et al., 2022)
- Defense using differential privacy (Lyu et al., 2020; Shi et al., 2022), machine unlearning (Bourtoule et al., 2021), and watermarking (He et al., 2022b)

6 Presenters

Dr. Qiongkai Xu, Research Fellow on Security in NLP, School of Computing and Infor-

[1]The resources pertaining to this tutorial are available at https://emnlp2023-nlp-security.github.io/.

mation System, the University of Melbourne, Australia.[2]

https://xuqiongkai.github.io
https://scholar.google.com/citations?user=wCer2WUAAAAJ

His recent research interest lies in auditing machine learning models, namely 1) privacy and security issues in ML/NLP models and 2) new evaluation paradigms for ML/NLP models. He has published more than 30 papers, with more than 10 of them on the topic of privacy and security in NLP.

Dr. Xuanli He, Research Fellow, Department of Computer Science, University College London, UK.

https://xlhex.github.io/
https://scholar.google.com/citations?user=TU8t0iAAAAAJ&hl

His recent research lies in an intersection between deep learning and natural language processing, with an emphasis on robustness and security in NLP models. He has published more than 10 top-tier conference papers about security in NLP models.

Acknowledgments

We thank Trevor Cohn, Benjamin I. P. Rubinstein, Lingjuan Lyu, and anonymous reviewers for their insightful suggestions, discussion, and comments on this tutorial and involved related work.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Sishuo Chen, Wenkai Yang, Zhiyuan Zhang, Xiaohan Bi, and Xu Sun. 2022. Expose backdoors on the way: A feature-based efficient defense against textual backdoor attacks.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. 2021. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Annual Computer Security Applications Conference*, pages 554–569.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. 2022. Recovering private text in federated learning of language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xuanli He, Chen Chen, Lingjuan Lyu, and Qiongkai Xu. 2022a. Extracted bert model leaks more information than you think! In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Xuanli He, Lingjuan Lyu, Lichao Sun, and Qiongkai Xu. 2021. Model extraction and adversarial transferability, your BERT is vulnerable! In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2006–2012, Online. Association for Computational Linguistics.

Xuanli He, Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2023a. IMBERT: Making BERT immune to insertion-based backdoor attacks. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 287–301, Toronto, Canada. Association for Computational Linguistics.

Xuanli He, Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, and Chenguang Wang. 2022b. Protecting intellectual property of language generation apis with lexical watermark. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10758–10766.

[2]He is now at Macquarie University as a lecturer.

Xuanli He, Qiongkai Xu, Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2023b. Mitigating backdoor poisoning attacks through the lens of spurious correlation. *arXiv preprint arXiv:2305.11596*.

Xuanli He, Qiongkai Xu, Yi Zeng, Lingjuan Lyu, Fangzhao Wu, Jiwei Li, and Ruoxi Jia. 2022c. CATER: Intellectual property protection on text generation APIs via conditional watermarks. In *Advances in Neural Information Processing Systems*.

Yujin Huang, Terry Yue Zhuo, Qiongkai Xu, Han Hu, Xingliang Yuan, and Chunyang Chen. 2023. Training-free lexical backdoor attacks on language models. In *Proceedings of the ACM Web Conference 2023*, pages 2198–2208.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. 2020. Thieves on sesame street! model extraction of bert-based apis. In *International Conference on Learning Representations*.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. 2021. Backdoor attacks on pre-trained models by layerwise weight poisoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3023–3032.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. Onion: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021b. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 443–453.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2022. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, Seattle, United States. Association for Computational Linguistics.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. *arXiv preprint arXiv:2306.17194*.

Xiaofei Sun, Jiwei Li, Xiaoya Li, Ziyao Wang, Tianwei Zhang, Han Qiu, Fei Wu, and Chun Fan. 2021. A general framework for defending against backdoor attacks via influence graph. *arXiv preprint arXiv:2111.14309*.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618.

Eric Wallace, Mitchell Stern, and Dawn Song. 2020. Imitation attacks and defenses for black-box machine translation systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5531–5546.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*.

Jun Wang, Chang Xu, Francisco Guzmán, Ahmed El-Kishky, Yuqing Tang, Benjamin Rubinstein, and Trevor Cohn. 2021. Putting words into the system's mouth: A targeted attack on neural machine translation using monolingual data poisoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1463–1473.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Chang Xu, Jun Wang, Yuqing Tang, Francisco Guzmán, Benjamin IP Rubinstein, and Trevor Cohn. 2021. A targeted attack on black-box neural machine translation with parallel data poisoning. In *Proceedings of the Web Conference 2021*, pages 3638–3650.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2023. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint arXiv:2305.14710*.

Qiongkai Xu, Xuanli He, Lingjuan Lyu, Lizhen Qu, and Gholamreza Haffari. 2022. Student surpasses teacher: Imitation attack for black-box NLP APIs. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2849–2860, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qiongkai Xu, Lizhen Qu, Zeyu Gao, and Gholamreza Haffari. 2020. Personal information leakage detection in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6567–6580.

Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. 2019. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207.

Wenkai Yang, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun, and Bin He. 2021a. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2048–2058, Online. Association for Computational Linguistics.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021b. Rap: Robustness-aware perturbations for defending against backdoor attacks on nlp models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8365–8381.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. Protecting language generation models via invisible watermarking. *arXiv preprint arXiv:2302.03162*.