Tutorial on Artificial Text Detection

Adaku Uchendu[1], **Vladislav Mikhailov**[2], **Jooyoung Lee**[1], **Saranya Venkatraman**[1], **Tatiana Shavrina**[3] and **Ekaterina Artemova**[2,4]

[1]The Pennsylvania State University [2]HSE University
[3]AI Research Institute [4]Huawei Noah's Ark Lab

Abstract

Recent advances in natural language generation have led to the development of models capable of generating high-quality and human-like texts among many languages and domains. However, it is known that such models can be misused for malicious purposes, including but not limited to generating fake news, spreading propaganda, and facilitating fraud. This tutorial aims at bringing awareness of *artificial text detection*, a fast-growing niche field devoted to mitigating the misuse of these models. It targets NLP researchers and industrial practitioners who work with text generative models and/or on mitigating ethical, social, and privacy harms. Our tutorial provides the attendees with a comprehensive background on this topic and reviews in a holistic manner: (1) issues of generative models that can exacerbate their misuse, (2) terminologies and task definitions, (3) models well-studied for the task, (4) existing datasets and benchmarks, (5) approaches to detecting generated texts, (6) standard crowd-sourcing practices and related critical studies, (7) downstream applications, and (8) established risks of harm. We conclude by outlining unresolved methodological problems and future work directions.

Type of Tutorial: Cutting edge.

1 Description

With recent advances in natural language generation (NLG), there has been a paradigm shift from template-based approaches towards using deep learning (DL) methods for text generation. A family of transformer language models (LMs; Vaswani et al., 2017) are now capable of generating high-quality and human-like texts among many languages and domains (Radford et al., 2018, 2019; Brown et al., 2020; Liu et al., 2020; Raffel et al., 2019; Lin et al., 2021).

However, it is known that these text generative models (TGMs) can be misused for generating fake news (Zellers et al., 2020), product reviews (Adelani et al., 2020), and even extremist and abusive content (McGuffie and Newhouse, 2020). Due to a scope of risks (Weidinger et al., 2021), the niche field of *artificial text detection (ATD)* has emerged, aiming at mitigating such misuse of LMs (Jawahar et al., 2020).

Our tutorial calls attention to this problem and provides the audience with a comprehensive review on inter-connected areas of NLG, responsible model development, and defense from adversaries/malicious users. We hope to bring NLP researchers and practitioners to be aware of standard approaches, unresolved methodological issues, and potential risks and encourage the community to propose novel solutions for developing robust, reliable, and interpretable artificial text detectors.

Relevance to the community A fast-growing area of NLG facilitates responsible development of LMs and methods of mitigating their potential misuse. However, the latter is rarely discussed in the literature and has been only partially addressed in a related tutorial by Nakov and Da San Martino (2020) on fact-checking, fake news, and propaganda. To the best of our knowledge, our tutorial is the first one in the *ACL venues that presents a systematic overview of ATD. The content of this tutorial can be found helpful for practitioners who work with TGMs and/or on mitigating risks from their misuse.

Ethical considerations This tutorial discusses ethical, social, and privacy risks of harm from misuse of TGMs and memorization of LMs. We understand that solutions to ATD can reveal techniques to evade detection. However, older detection models will become obsolete as this task becomes even more popular and newer AI text-generators are built and deployed. Therefore, the risk of discussing detection techniques possesses very minimal implications.

2 Outline

The intended tutorial duration is three hours and a half-hour break.

2.1 Introduction (30 minutes)

This section introduces the tutorial by presenting the recent advances in NLG and motivating the task of ATD. We then outline the scope of the tutorial and discuss some of the main issues of TGMs that fall under ATD:

Toxicity and Hate Speech Real-world applications demand substantial safety control over text generation, including toxicity and hate speech. TGMs are known to generate harmful content even when fed with a non-toxic prompt. (Pavlopoulos et al., 2020; Gehman et al., 2020).

Memorization Carlini et al. (2021) report that scaling LMs increases their memorization of data, meaning that scaled TGMs tend to leak more memorized information during inference. This poses a security risk as accurate personally identifiable information can be leaked from LMs.

Hallucinated Content Generation Zhou et al. (2021) show that TGMs sometimes generate texts that are unfaithful to the prompt. Ji et al. (2022) presents a survey on this problem in NLG.

Misinformation Generation Despite the issues mentioned above, TGMs can be used by malicious users to generate polarizing misinformation at scales, such as propaganda, scam content, and fake news (Jawahar et al., 2020; Ahmed et al., 2021). This potential application affects the news ecosystem and causes negative ethical and social impacts.

2.2 Background (25 minutes)

Terminologies There is no common terminology for ATD. We will define different jargon used in the research. These terminologies include:

- **Artificial**: This means not human-made. Different researchers use synonyms of "artificial" to denote the non-human origin of texts: *synthetic*, *AI-generated*, *machine-generated*, *neural*, etc.

- **Artificial text**: also called *synthetic text*, *AI-generated text*, *machine-generated text*, *machine-made text*, *neural text*, and *computer-generated text*. These terms denote a text generated by an AI technology, usually a neural model. Some of the terms are broad; however, we focus on a sub-type of *artificial text*, which is only text generated by neural LMs, specifically Transformer-based ones.

- **Text generative models**: This refers to a family of LMs that can be used to generate texts, e.g. GPT-2/GPT-3. Several other terms for TGM are *AI text-generator*, *machine text-generator*, and *neural text-generator*.

- **Artificial text detectors**: These are ML- or DL-based models that detect artificial texts. They are also known as *machine text detectors*, *neural text detectors* and *AI text detectors*.

Task Definitions: We describe standard task definitions, including human vs. machine classification, neural authorship attribution, testing robustness, and reverse engineering, such as predicting decoding strategy or TGM's size.

Text Generative Models: We briefly introduce popular and state-of-the-art TGMs varying in architecture and pre-training objective. We also list TGMs and their configurations that are well-studied in ATD.

2.3 Datasets (15 minutes)

This section presents English *human vs. machine* datasets, including benchmark datasets used to achieve *artificial text detection*. We also provide references for related datasets in other languages.

Break (30 minutes)

2.4 Artificial Text Detectors (30 minutes)

Here we discuss unsupervised, threshold-based, and supervised methods for detecting artificial texts. As part of this section, we also outline tools for human-model collaboration and visualization of generated text properties.

2.5 Research on Human Evaluation (20 minutes)

We discuss recent papers about performing crowdsourcing human studies to distinguish human-written texts from AI-generated ones.

2.6 Applications (20 minutes)

This section highlights downstream applications of artificial text detectors, such as warning users about potentially fake content on social media and news platforms, filtering corpora augmented with TGMs, defense from abuse of product reviews platforms/adversaries, spam filtering, and propaganda spread with bots.

2.7 Ethical and Social Risks (20 minutes)

We highlight established ethical and social risks of harm from TGMs and the current status on this topic.

2.8 Summary (10 minutes)

The final section summarizes the topics covered in this tutorial, pointing out unresolved methodological problems and future work directions.

3 Breadth

We estimate 85% of the work covered will not be by the tutorial presenters.

4 Diversity

The tutorial covers data for multiple languages (e.g., detection of machine-translated texts among different language pairs), various domains (e.g., social media, news articles, product reviews), and architecture choices (e.g., TGMs' architecture, pretraining objective, size, decoding strategy).
The background of the tutorial presenters is evenly distributed among academia and industry, and also countries and continents (Russia and USA). The team consists of middle and senior NLP researchers, junior and senior Ph.D. students at Pennsylvania State University (PSU) and HSE University. The presenters have years of academic and industry research experience, with the research work published in multiple NLP venues such as LREC, COLING, EACL, EMNLP, AAAI, INLG, and workshops co-located with NeuriPS, EACL, NAACL, and EMNLP.

5 Presenters

Our team is experienced in developing methods, shared tasks, and benchmarks for ATD, establishing privacy risks from memorization in LMs, and leading R&D teams on training extensive TGMs. Research works on these topics are published in EMNLP venues (Uchendu et al., 2020, 2021; Kushnareva et al., 2021).

Adaku Uchendu is a fourth-year Ph.D. student at PSU, working under the guidance of Dr. Dongwon Lee on detection of artificial texts.
E-mail: azu5030@psu.edu

Vladislav Mikhailov works as an invited lecturer in Big Data & IR School at HSE University and advises BA/BSc students on topics related to NLG.
E-mail: vmkhlv@hse.ru

Jooyoung Lee is a second-year Ph.D. student at PSU, working under the guidance of Dr. Dongwon Lee on privacy risks of LMs resulting from memorization.
E-mail: jfl5838@psu.edu

Saranya Venkatraman is a fourth-year Ph.D. student at PSU, working under the guidance of Dr. Prasenjit Mitra on neural dialogue generation.
E-mail: szv4@psu.edu

Tatiana Shavrina is an PI at AI Research Institute. Her main research interests are evaluation of LMs and NLG.
E-mail: shavrina@airi.net

Ekaterina Artemova holds a PostDoc position at CS Faculty at HSE University and advises NLP teams at Huawei Noah's Ark Lab on advanced research topics. Ekaterina focuses on NLU tasks, ranging from ToD systems to NLG.
E-mail: elartemova@hse.ru

6 Additional Details

Technical Requirements: We require Internet access to the tutorial room and stable connection.

Audience Size: Due to the increasing number of submissions to NLG and related tracks in the latest *ACL venues, we estimate that up to 100 attendees will be interested in our tutorial.

Prerequisite Background: Our target audience is general NLP conference attendances, researchers, and industrial practitioners. We expect that the audience is broadly familiar with:

- standard approaches to neural text generation, TGMs' architectures and decoding methods;
- basic supervised learning paradigm and commonly used models, such as logistic regression and deep neural networks;

- general dataset collection and annotation practices.

Open Access: We agree to allow an open access to the tutorial videos, slides, and other material in the ACL Anthology. The materials will additionally be posted on our tutorial GitHub page.

Reading List

Artificial Text Detection: Jawahar et al. (2020) provide a critical survey on ATD, including social impacts of TGMs and existing challenges in the field.

Human Evaluation Studies: van der Lee et al. (2019) review common annotation protocols and guidelines on human evaluation of AI-generated texts. Karpinska et al. (2021) present a survey of 45 papers on human evaluation of open-ended text generation, emphasize the reproducibility problem, and conduct a comprehensive study comparing the results of crowd-sourced workers and expert annotators.

Ethical Risks: We refer the attendees to Smiley et al. (2017); Weidinger et al. (2021) for overviewing ethical implications and social risks from TGMs.

References

David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In *International Conference on Advanced Information Networking and Applications*, pages 1341–1354. Springer.

Alim Al Ayub Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. 2021. Detecting fake news using machine learning: A systematic literature review. *arXiv preprint arXiv:2102.04458*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kris McGuffie and Alex Newhouse. 2020. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. *arXiv preprint arXiv:2009.06807*.

Preslav Nakov and Giovanni Da San Martino. 2020. Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–19, Online. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Charese Smiley, Frank Schilder, Vassilis Plachouras, and Jochen L. Leidner. 2017. Say the right thing right: Ethics issues in natural language generation systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 103–108, Valencia, Spain. Association for Computational Linguistics.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending Against Neural Fake News. *Neurips*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP Findings)*, Virtual.