Between the Lines: Generating, Detecting and Defending against Textual Deepfakes

Elena-Anca PARASCHIV[1,2], Carmen-Elena CÎRNU[1]
[1]National Institute for Research and Development in Informatics - ICI Bucharest
[2]Doctoral School of Electronics, Telecommunications and Information Technology, National University of Sciences and Technologies Politehnica of Bucharest
elena.paraschiv@ici.ro, carmen.cirnu@ici.ro

Abstract: This paper provides an in-depth exploration of textual deepfakes, emphasizing their significance in the realm of cybersecurity. It delves into the sophisticated techniques employed by artificial intelligence (AI) algorithms to generate highly deceptive textual content, highlighting the potential cybersecurity threats they pose. Through the examination of malicious actors' utilization of textual deepfakes for activities such as misinformation campaigns and phishing attacks, this paper highlights the critical need for robust detection mechanisms and defense strategies. By fostering awareness towards the impact of deepfakes, enhancing deepfake detection capabilities, and advocating the responsible usage of AI technologies, this paper aims to address the cybersecurity challenges posed by textual deepfakes and safeguard the integrity of digital communication channels.
Keywords: deepfakes, cybersecurity, textual deepfakes, AI, GPT.

INTRODUCTION

The dynamic landscape of cyber threats poses a major challenge to individuals, institutions, as well as critical infrastructures and the ongoing increase in frequency and complexity of cyberattacks requires continuous innovation in defensive strategies. Conventional cybersecurity techniques mostly rely on responsive measures and signature-based detection. Nevertheless, these approaches struggle to keep up with the acceleration in the development of new attack vectors and the evolving sophistication of malware.

The recent widespread use of generative models and the advent of deepfakes have revolutionized artificial intelligence (AI) and digital content creation. In this context, generative models, driven by deep learning architectures, hold the significant capability of generating real-world scenarios, including texts, audio media, photos and videos. However, even though these developments have released an unprecedented potential in various domains, they have also brought about a number of serious challenges, particularly in the realm of cybersecurity.

https://doi.org/10.54851/v6i1y202401

Generative models provide novel possibilities by allowing the creation of realistic and varied synthetic data, such as simulated malware instances, network attack structures, and threat scenarios. However, synthetic data can also be applied for various defensive goals. Generative models may assist in the development and training of resilient malware detection software or other security measures through the production of a broad variety of realistic attack scenarios.

In order to better understand the complex world of textual deepfakes, this paper will examine the methods used to counteract their harmful impacts as well as the complexities involved in their development and detection. By means of an extensive analysis of extant literature, research methodologies, and case studies, this paper aims to clarify the mechanisms that underlie textual "deepfakery", to evaluate the effectiveness of existing detection techniques, and suggest innovative strategies for ameliorating their effects.

The present study was carried out in the context of a dynamic digital environment that is marked by quickening technological progress and changing communication paradigms. A strong defence against textual deepfakes is vital as society struggles with the spread of fake news, Internet manipulation, and information warfare. By shedding light on the nuances of this burgeoning phenomenon and laying forth a roadmap for countermeasures, the aim is to protect the authenticity of textual communication and uphold the fundamentals of trust in the digital era. That is why this article presents a short introduction about understanding deepfakes, the models used for generation of textual deepfakes as well as some methods for detecting deepfakes and defending against them along with a use case and examples of using language models in order to generate textual deepfakes.

The remainder of this paper is structured as follows. The second section presents a general understanding of what deepfake is, focusing on textual deepfakes. The third section sets forth the most important AI-based techniques used for generating textual deepfakes. Section four presents the basic approaches and algorithms for detecting textual deepfakes and the fifth section gives a brief presentation of what can be done to defend against deepfakes. An use case and several examples of generating and detecting textual deepfakes are presented in the sixth section and some conclusions are drawn at the end of the paper.

UNDERSTANDING DEEPFAKES

Being a combination of "deep learning" and "fake", deepfakes refer to synthetic media (texts, audio media, images, and videos) altered by AI techniques (Somers, M., 2020). While there was a significant development in the multimedia space (photos, video and audio media), deepfakes have subsequently spread to other content types, such as textual narratives (Figure 1). Considering the complexity of deepfakes, it is crucial to also encompass the challenges posed by their textual equivalents.

Figure 1. *Types of Deepfakes*

Deepfakes covers a wide range of synthetic content (Adee, S., 2020) produced by deep learning algorithms. Once applied to text, they manifest as artificially generated textual sequences designed to resemble human writing approaches and linguistic habits, including falsified articles, forged messages, misleading instances conceived to manipulate the readers.

The development of deepfakes is strongly associated with the advancements made in the context of deep learning algorithms together with their use and applications in media synthesis. Early research centered on generating realistic pictures and videos, leading to the implementation of advanced methods capable of creating convincing multimedia deepfakes. Although they are a relatively new phenomenon, textual deepfakes have gained popularity alongside advancements in natural language processing (NLP).

In the late 2010s, deepfakes started to attract public notice, generating concerns. At first, the phenomenon was only restricted to humorous situations, but it did not take too much time for deepfakes to cause negative effects. Technology now highlights a crucial point in the digital age: the thinning line between what is real and what is fake.

Artificially-generated text can be easily used to offer misleading information. There are several publications that highlight the importance of detecting and defending against textual deefakes. Zellers et al. (2019) proved that a language model (GPT-2) can accurately generate fake news articles. Automatically generated emails are another aspect of concern considering the textual deepfakes (Das & Verma, 2018) as well as producing false online restaurant reviews (Yao et al., 2017) that may contribute to users` starting to lose trust in digital content. Such tools can facilitate widespread dissemination of misinformation which may lead to drastic consequences.

Textual deepfakes are evolving in a similar way to their multimedia counterparts, which is indicative of a larger trend towards the convergence of AI-driven content modification technologies. There are many complex reasons for creating textual deepfakes. One primary motivation is the propagation of misinformation for political, financial or cultural advantages. Textual deepfakes are a useful tool for manipulating public opinion and decision-making processes by creating fake news stories, social media posts, emails, or online reviews. Moreover, they can be used for malicious intentions: impersonation, cyberbullying, or defamation. Understanding the root causes behind textual deepfakes is essential in order to effectively detect and mitigate them and to prevent their propagation.

GENERATION OF TEXTUAL DEEPFAKES

The primary ingredient in creating textual deepfakes consists in AI techniques and NLP models, being responsible for producing deepfakes at a lower cost and a much faster rate. There are different complex techniques and methodologies that are currently used for generating textual deepfakes:

- **Language generation models**

Language generation models represent a significant advancement in the field of NLP, allowing AI to generate text that is comparable with human-authored content. The capacity to comprehend and replicate the features, framework, and semantics of human language is the fundamental component of these models. Among the multiple types of language generation models, there are some that need to be considered:

- **Autoregressive models**, such as Generative Pre-Trained Transformer (GPT), generate text sequentially, with each word being constructed based on the preceding words. This process enables models to preserve coherence and context across the generated text.

GPT represents a deep learning model which is built upon a transformer architecture and it enables the use of generative AI-powered applications (such as ChatGPT). The GPT undergoes pre-training on a large amount of text data and it can be fine-tuned for various tasks, including text classification and generation, automatic translation, or sentiment analysis (Gokul et al., 2023). The architecture of a transformer implemented in GPT is considered a substantial progression in relation to earlier NLP methodologies. It enables a self-attention mechanism responsible for incorporating the entire context of a sentence in the word prediction, thereby enhancing its language

comprehension and generation capabilities. The decoder-only architecture of GPT allows the model to be more effective for different kinds of tasks. Eliminating the encoder allows GPT models to streamline the processing of the input, enabling faster data generation.

- **Conditional and Contextual Models,** such as Bidirectional Encoder Representations from Transformers (BERT), takes into consideration the entire input sequence when generating the output.

BERT is considered a bidirectional pre-training approach (Devlin et al., 2019), which allows it to capture rich semantic information and contextual characteristics. It considers the entire input data when it generates the output, proving a good performance for various natural language understanding tasks. It integrates masked language modeling, meaning that, during pre-training, a sequence of the input tokens is randomly masked, and BERT is trained to predict them, depending on the surrounding context. The encoder-only architecture of BERT enables the layers to bidirectionally process the input text, creating contextualized representations of words / tokens from both directions.

Figure 2 illustrates the difference between BERT and GPT, based on the Transformer architecture.

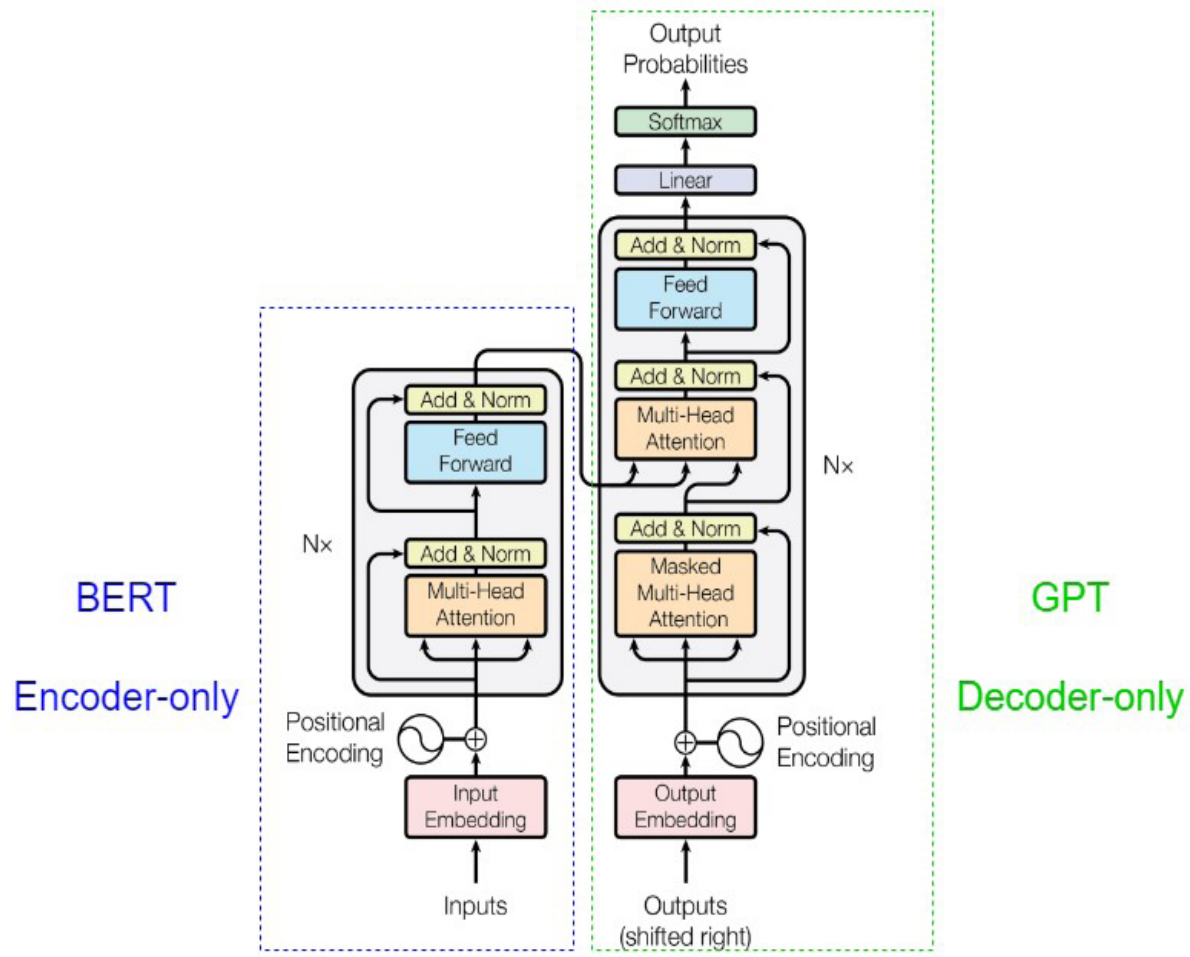

Figure 2. *BERT and GPT based on the Transformer architecture (adapted from Vaswani et al. (2017))*

- **Fine-tuning and Transfer Learning**

Using specific datasets or tasks, pre-trained language models are fine-tuned in order to generate text that is adapted to particular contexts and domains. Transfer learning techniques enable the adaptation of pre-existing knowledge from one task to another, enhancing the performance and flexibility of textual generation models. Transfer learning in language models considers several steps:

- *pre-training* – the model is trained on a significant corpus of text in order to learn the correlations and characteristics of the language;
- *fine-tuning* – the model can be further fine-tuned on a particular task that can involve a smaller dataset, aiming to adapt the linguistic representations for the general use;
- *inference* – the model receives the input and outputs a representation of its prediction for that specific task.

- **Adversarial Training**

Adversarial models enhance the robustness and diversity of the generated text, being trained against adversarial instances or disruptions which enables one to better discern between what is fake and what is real. When asked to generate an adversarial example, ChatGPT responded with the following details illustrated in Figure 3.

You

Generate an adversarial example of visual similarity and semantic similarity for the following text: When Earth becomes uninhabitable in the future, a farmer and ex-NASA pilot, Joseph Cooper, is tasked to pilot a spacecraft, along with a team of researchers, to find a new planet for humans. Please also give a prediction of positive and negative in %

ChatGPT

To create an adversarial example with both visual and semantic similarity, we need to generate a text that is visually and semantically similar to the original text but carries a different meaning. Here's an example:

Original Text:
"When Earth becomes uninhabitable in the future, a farmer and ex-NASA pilot, Joseph Cooper, is tasked to pilot a spacecraft, along with a team of researchers, to find a new planet for humans."

Adversarial Example:
"When Earth becomes uninhabitable in the future, a farmer and ex-NASA pilot, Joseph Cooper, is tasked to pilot a spacecraft, along with a team of researchers, to find a new planet for robots."

This adversarial example maintains the visual and semantic structure of the original text but changes the intended recipient of the new planet from humans to robots.

As for the prediction of positive and negative sentiment, it's subjective and depends on the interpretation of the text. However, considering the potential consequences of humans becoming unable to inhabit Earth and the task of finding a new planet, the sentiment could be:

- Positive: 60%
- Negative: 40%

Figure 3. *Adversarial example*

To generate an adversarial example with both visual and semantic similarity, the original text was analyzed to identify its key components and structure. In this instance, the text revolves around the concept of a future scenario where the Earth becomes uninhabitable, and a protagonist, Joseph Cooper, is tasked with piloting a spacecraft to find a new planet for humans. The goal was to maintain the visual and semantic coherence of the original text while introducing a subtle alteration to convey a different meaning.

The adversarial example was crafted by modifying the intended recipient of the new planet, shifting it from humans to robots. This alteration was strategically chosen to maintain the overall structure and coherence of the text while introducing a semantic shift that subtly alters the intended message. By replacing „humans" with „robots," the adversarial example introduces a new layer of interpretation, implying a future scenario where robots, rather than humans, are the beneficiaries of Joseph Cooper's mission.

Furthermore, sentiment analysis was conducted to assess the potential emotional impact of the adversarial example. Given the speculative nature of the scenario and the implications of humans becoming unable to inhabit Earth, the sentiment was predicted to be slightly negative, reflecting the potential concerns and uncertainties associated with such a scenario. However, it's important to note that sentiment analysis is subjective and can vary based on individual interpretations of the text.

That is why, the increasing prevalence of language models (e.g. GPT) raises significant apprehension regarding security and privacy. A primary concern revolves around the potential misuse of language models for nefarious purposes, including the creation of fabricated news articles or deceptive deep fakes. The capability of GPT to generate text that closely resembles authentic content poses challenges in discerning between genuine and counterfeit information, exacerbating the risks associated with misinformation and fraudulent activities (Gokul et al., 2023).

DETECTING TEXTUAL DEEPFAKES

Detecting textual deepfakes, or synthetic text generated to mimic human-written content, poses a significant challenge due to the increasing sophistication of generative models and the subtle nature of textual manipulation. Understanding the approaches, algorithms, evaluation metrics, and challenges associated with detecting textual deepfakes is crucial for developing effective detection strategies and safeguarding against the spread of misinformation.

Approaches and algorithms for detecting textual *deepfakes:*

- ***Statistical analysis*** involves scrutinizing linguistic features, syntactic structures, and semantic patterns within the text. Anomalies or inconsistencies revealed through statistical analysis may signify the presence of artificially generated text. *Example:* Analyzing the frequency of certain words or phrases that are uncommon in natural language but frequently appear in generated text. For instance, deepfake text might contain an unusually high occurrence of technical jargon or complex terminology that is not typical in human-generated content.
- ***Machine Learning*** models utilize classifiers like support vector machines (SVMs), random forests, or neural networks, trained on datasets containing both real and synthetic text. These models differentiate between genuine and synthetic text based on learned features or embeddings. *Example:* Training a support vector SVM classifier on a labeled dataset containing both genuine and synthetic text. The classifier learns to distinguish between the two types of text based on features such as word frequency, sentence structure, and syntactic patterns. During inference, the model can then classify new text samples as either genuine or fake based on the learned patterns.

- ***Adversarial detection*** employs techniques to uncover adversarial perturbations or artifacts introduced during the generation of textual deepfakes. Adversarial detection methods aim to identify subtle cues or inconsistencies indicative of synthetic text. *Example:* Training a SVM classifier on a labeled dataset containing both genuine and synthetic text. The classifier learns to distinguish between the two types of text based on features such as word frequency, sentence structure, and syntactic patterns. During inference, the model can then classify new text samples as either genuine or fake based on the learned patterns.

Evaluation metrics and benchmarks for gauging detection performance include accuracy, reflecting the proportion of instances accurately classified (i.e. genuine or fake text) by the detection algorithm. Precision signifies the proportion of accurately classified fake instances among all instances classified as fake, while recall represents the proportion of accurately classified fake instances among all genuine fake instances. The F1 Score (Sokolova & Lapalme, 2009) is used for calculating the harmonic mean of precision and recall, offering a balanced assessment of detection performance. The Area Under the ROC Curve (AUC-ROC) (Fawcett, 2006) evaluates the trade-off between sensitivity (true positive rate) and specificity (true negative rate) across different classification thresholds.

However, there are **challenges and limitations in detecting textual deepfakes** which include adversarial generation, where generative models adapt to avoid detection algorithms, resulting in an ongoing challenge between detection and generation techniques. Data imbalance poses challenges due to the limited availability of labeled datasets containing diverse examples of textual deepfakes, making it difficult to train robust detection models. Semantic coherence represents another challenge, as sophisticated generative models can produce text that is contextually relevant and semantically coherent, blurring the distinction between genuine and synthetic content.

DEFENDING AGAINST TEXTUAL DEEPFAKES

As the proliferation of textual deepfakes continues to pose significant challenges to media integrity and societal trust, developing effective defense mechanisms is imperative. Strategies for mitigating the spread and impact of textual deepfakes, along with the adoption of tools and technologies to enhance resilience against such attacks, play a crucial role in safeguarding against the manipulation of online content.

There are several strategies that can be employed to mitigate the spread and impact of textual deepfakes, including:

- ***educational initiatives*** by increasing public awareness about the existence and potential dangers of textual deepfakes through educational campaigns and media literacy programs that can empower individuals to critically evaluate online content which may help in mitigating the impact of misinformation;
- ***content verification and fact-checking*** by leveraging automated content verification tools and fact-checking services to detect and debunk false or misleading information disseminated through textual deepfakes. Collaborative efforts among media organizations, technology companies, and fact-checking agencies can help identify and counteract deceptive content effectively;
- ***algorithmic transparency and accountability*** by promoting transparency and accountability in algorithmic decision-making processes employed by social media platforms and content recommendation systems which can enhance algorithmic transparency and help mitigate the inadvertent amplification of textual deepfakes and other forms of harmful content.

Various tools and technologies can be employed to enhance resilience against textual deepfake attacks, including:

- **Deepfake detection tools:** Deploying machine learning-based detection algorithms and software solutions capable of identifying textual deepfakes and flagging suspicious content in real time. These tools can assist content moderators and platform administrators in combating the spread of deceptive content. Developers of GPT-2 have released an open-source model on Github (Github, 2020) aimed at detecting text generated by GPT-2 in comparison with human-written text. For users lacking coding experience, a convenient Google Chrome plugin called GPTrue or False is available. This plugin allows users to easily discern suspicious text by highlighting it and clicking on the GPTrue or False icon (Starace, 2024). Subsequently, the plugin provides a certainty percentage indicating whether the text is likely human-written or generated by AI.
- **Blockchain and Cryptographic Solutions:** Implementing blockchain technology and cryptographic techniques to verify the authenticity and integrity of textual content. Blockchain-based platforms can provide immutable records of content provenance and facilitate traceability, enhancing trust in online information sources (World Economic Forum, 2021).

To enhance defense mechanisms against textual deepfakes, future research and development efforts should prioritize several key areas. Firstly, there is a need to improve detection technologies, developing more robust and scalable algorithms capable of accurately identifying increasingly sophisticated textual deepfakes. Integration of multimodal approaches, which combine textual and visual cues, can significantly enhance detection performance and reliability. Secondly, strengthening regulatory frameworks is essential. Enacting legislation and regulatory measures to hold accountable entities responsible for the creation, dissemination, and monetization of deceptive content can act as a deterrent against malicious actors. Clear legal guidelines and effective enforcement mechanisms are crucial in encouraging platform providers to implement proactive measures against textual deepfakes. Finally, promoting collaboration and knowledge sharing among researchers, industry stakeholders, and policymakers is vital. Initiatives such as open-source platforms, collaborative projects, and interdisciplinary partnerships can facilitate innovation and collective action in combating online misinformation. By addressing these multifaceted challenges collectively, a safer and more trustworthy online environment could be created.

USE CASE AND EXAMPLES OF GENERATION AND DETECTION OF TEXTUAL DEEPFAKES

Texts generated synthetically could potentially serve as tools to deceive Internet users. For example, they could be utilized in the widespread creation of extremist or violent content with the intention of radicalizing individuals, fabricating fake news for disinformation campaigns, crafting email texts for phishing attacks, or generating fake reviews targeting particular hotels, venues, or restaurants. This collective misuse of synthetic texts has the potential to erode trust in online content for some users while enticing others to partake in antisocial and hazardous actions.

This section presents a use case highlighting the collaboration between cybercriminals and AI-generated text tools to create and distribute deceptive content (Figure 4). Through the integration of AI technologies, cybercriminals exploit various media, including phishing emails, misleading online reviews, fabricated news articles, and social media disinformation campaigns, to manipulate and deceive unsuspecting individuals.

For instance, cybercriminals leverage AI-generated text tools to craft convincing phishing emails designed to trick recipients into divulging sensitive information or clicking on malicious

links. These emails often employ sophisticated social engineering tactics, such as impersonating trusted entities or creating urgent scenarios to elicit a response from recipients.

Similarly, AI-generated text is utilized to fabricate misleading online reviews aimed at artificially boosting the reputation of products or services or tarnishing the reputation of competitors. By generating deceptive reviews that appear authentic, cybercriminals manipulate consumer perceptions and influence purchasing decisions.

In addition, cybercriminals employ AI-generated text to fabricate news articles that propagate false information or sensationalize events for the purpose of spreading disinformation and manipulating public opinion. These fabricated news articles, often disseminated through social media channels, exploit vulnerabilities in information ecosystems and contribute to the proliferation of fake news.

Furthermore, social media platforms serve as a fertile ground for the dissemination of AI-generated text-based disinformation campaigns orchestrated by cybercriminals. By leveraging AI tools to generate and amplify divisive narratives, spread rumors, or incite to discord, cybercriminals exploit the virality and reach of social media platforms to sow confusion and undermine trust in different institutions.

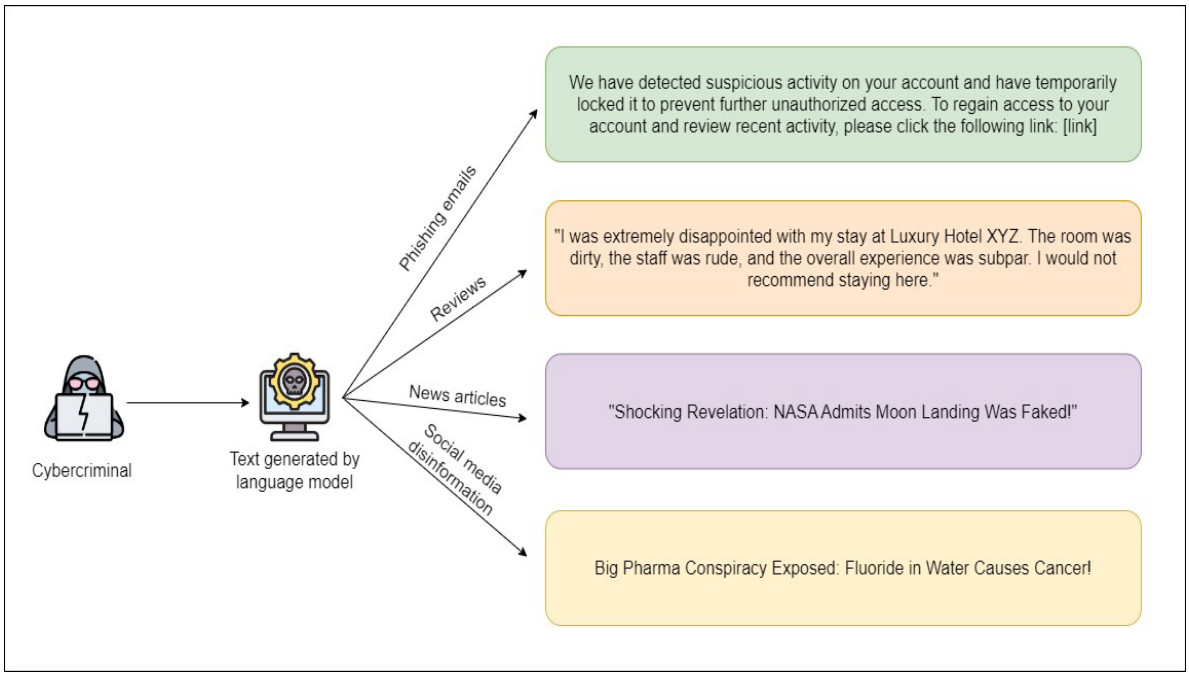

Figure 4. *Generation of different malicious tasks based on a language model*

The detection accuracy of these defense mechanisms was impressive, yet their practical effectiveness, especially in adversarial scenarios, remained uncertain. Evaluation primarily relied on datasets curated by researchers, which lacked exposure to synthetic data encountered in real-world settings. Malicious actors are expected to adapt their tactics to evade detection in practical contexts, a challenge not adequately addressed in current research.

The language model illustrated in Figure 5 serves as the primary source of text generation with two distinct pathways that can be considered in generating a textual deepfake. The normal generation pathway represents the standard process of text generation, where the language model produces text without deliberate modifications. The generated text evolves into deepfake articles. Subsequently, the deepfake articles are subjected to analysis by a deepfake

detection system (Figure 5). In the case of the adaptive generation pathway which involves the autoregressive models or the conditional ones, the malicious actor employs adaptive techniques to generate text with the intention of evading detection. The malicious actor modifies the decoding strategy of the language model to produce modified deepfake text articles or introduces adversarial perturbations during the text generation process, resulting in the creation of modified deepfake text articles. These modified deepfake text articles, incorporating adaptive techniques, are also input to a deepfake detection system for evaluation.

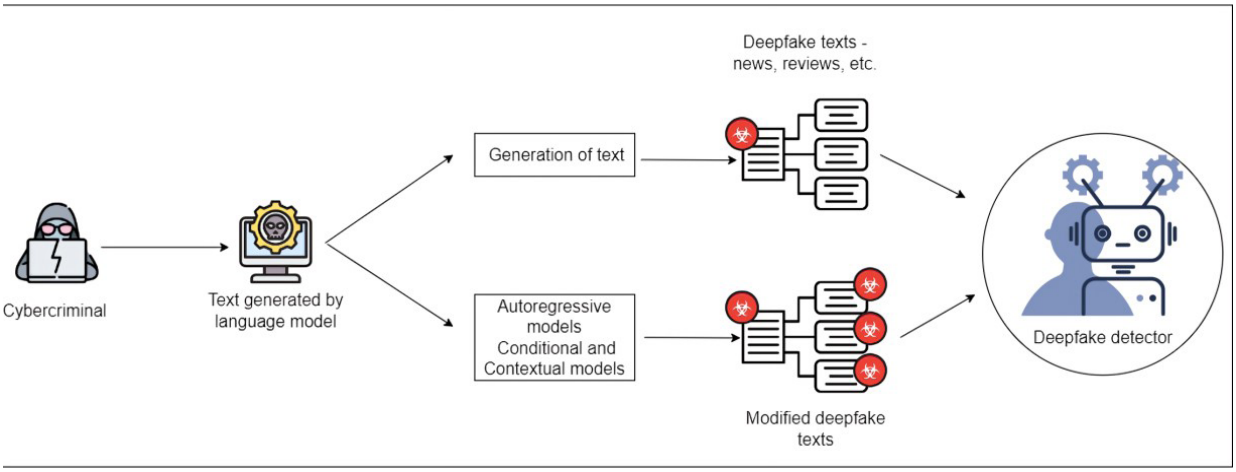

Figure 5. *Generation and detection of deepfake texts*

Once the generated data is fed into the deepfake detector, there are two possible outcomes, such as successfully detecting the deepfakes or failing to do so. Successful detection entails the identification of manipulated content, ensuring the integrity of the data and preventing potential harm or misinformation. Conversely, if the deepfake detector fails to identify the manipulated content, it may result in the dissemination of deceptive information, posing risks to individuals and communities. Therefore, the effectiveness of the deepfake detector plays a critical role in safeguarding against the proliferation of fraudulent content and maintaining trust in digital media.

CONCLUSION

The rise of deepfake technology brings about both significant opportunities and profound challenges across various domains, including media, cybersecurity, and societal trust. This paper explored the capabilities of deepfake technology, its potential applications, and the associated risks. From generating hyper-realistic synthetic media to manipulating public perception through misinformation and disinformation campaigns, the impact of deepfakes on society cannot be understated. Moreover, this paper analysed the ongoing efforts for developing detection mechanisms and countermeasures to mitigate the adverse effects of deepfakes.

While advancements in deepfake detection technology are promising in combatting the spread of manipulated content, the rapidly evolving nature of deepfake generation techniques necessitates continuous innovation and adaptation. Furthermore, addressing the ethical, legal, and societal implications of deepfakes requires a multifaceted approach, involving collaboration among researchers, policymakers, industry stakeholders, and the civil society.

While navigating the complex landscape of deepfakes, it is imperative to remain vigilant, to critically assess the encountered information, and foster digital literacy and resilience among individuals and communities. By promoting awareness towards the impact of deepfakes, enhancing detection capabilities, and fostering responsible use of emerging technologies, the risks posed by deepfakes could be mitigated and the integrity of the digital ecosystem could be safeguarded in a collective manner. Only through concerted efforts and proactive measures it will be possible to navigate the challenges posed by deepfakes and uphold the principles of truth, transparency, and trust in the digital age.

REFERENCE LIST

Adee, S. (2020) *What Are Deepfakes and How Are They Created? Deepfake technologies: What they are, what they do, and how they're made.* https://spectrum.ieee.org/what-is-deepfake. [Accessed 1st March 2024].

Das, A. & Verma, R. (2018) Automated email Generation for Targeted Attacks using Natural Language. *Proceedings of TA-COS 2018: 2nd Workshop on Text Analytics for Cybersecurity and Online Safety, 12 May 2018, Miyazaki, Japan.* European Language Resources Association (ELRA). pp. 23-30.

Github. (2020) *Dataset of GPT-2 outputs for research in detection, biases, and more.* https://github.com/openai/gpt-2-output-dataset [Accessed 2nd March 2024].

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* [cs.CL]. Available at: http://arxiv.org/abs/1810.04805. [Accessed 1st March 2024]

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters.* 27(8), 861-874. doi: 10.1016/j.patrec.2005.10.010.

Gokul, Y., Ramalingam, M., Chemmalar Selvi, G., Supriya, Y., Gautam, S., Praveen, K. R. M., Deepti Raj, G., Rutvij, H. J., Prabadevi, B., Weizheng, W., Athanasios V. V. & Thippa, R. (2023) Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *arXiv* [cs.CL]. Available at: http://arxiv.org/abs/2305.10435. [Accessed 1st March 2024]

Sokolova, M. & Lapalme, G. (2009) A systematic analysis of performance measures for classification tasks. *Information Processing & Management.* 45(4), 427-437. doi: 10.1016/j.ipm.2009.03.002.

Somers, M. (2020) *Deepfakes, explained.* https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained. [Accessed 29nd February 2024].

Starace, G. (2024) *GPTrue or False.* https://www.giuliostarace.com/projects/gptrue-or-false/ [Accessed 8th April 2024].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I., (2017) Attention is all you need. *arXiv* [cs.CL]. Available at: http://arxiv.org/abs/1706.03762. [Accessed 5th March 2024]

World Economic Forum. (2021) *Blockchain can help combat the threat of deepfakes. Here's how.* https://www.weforum.org/agenda/2021/10/how-blockchain-can-help-combat-threat-of-deepfakes/ [Accessed 3rd March 2024]

Yao, Y., Viswanath, B., Cryan, J., Zheng, H. & Zhao, B. Y. (2017) Automated Crowdturfing Attacks and Defenses in Online Review Systems. *arXiv* [cs.CR]. Available at: http://arxiv.org/abs/1708.08151. [Accessed 8th April 2024].

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. & Choi, Y. (2019) Defending Against Neural Fake News. *arXiv* [cs.CL]. Available at: http://arxiv.org/abs/1905.12616. [Accessed 8th April 2024].

This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.