Protecting Your Online Privacy: Insights on Digital Twins and Threat Detection

Sergej Schultenkämper
Bielefeld University of Applied Sciences and Arts
Bielefeld, Germany
sergej.schultenkaemper@hsbi.de

Frederik S. Bäumer
Bielefeld University of Applied Sciences and Arts
Bielefeld, Germany
frederik.baeumer@hsbi.de

***Abstract*—This paper presents considerations for the use of Digital Twins for protecting online privacy and detecting potential threats. While Digital Twins offer a promising approach to modeling individual vulnerability and identifying online threats, there are still significant challenges to be addressed. One of the primary challenges is the need for diverse and comprehensive training data, as Digital Twin instantiation relies heavily on machine learning algorithms. To address this issue, the authors describe two datasets for Digital Twin instantiation based on Computer Vision and Natural Language Processing techniques. In addition, the authors also examine the limitations of current approaches for creating Digital Twins and propose potential areas for future research. The main objective of this work is to provide insights and considerations for the use of Digital Twins in online privacy protection and threat detection.**

Index Terms*—*Digital Twin*; *Privacy*; *Social Networks

I. Introduction

In recent years, the widespread use of the internet and social media has resulted in an explosion of personal data available online. While this has provided a multitude of opportunities for users to connect with others and share information, it has also created significant privacy risks. The practice of doxing, or the public release of personal information, is a particularly concerning example of how online privacy can be compromised [1]–[3]. Doxing can lead to harassment, stalking, and even physical harm [4]. As the amount of personal data available online continues to grow, the risk of doxing and other privacy threats also increases.

To address this issue, researchers have proposed the use of Digital Twins (DTs) as a way to model the vulnerability of individuals to privacy threats on the Web [5]. DTs are computer-based models that simulate or mirror the life of a physical entity or process and are commonly used in the manufacturing and aviation industries to monitor, control, and optimize the life cycle of real-world assets [6]. In the context of privacy threats on the web, a DT would represent a digital representation of a real person instantiated by information available online [5]. Essentially, existing methods and approaches can be used here since modeling knowledge and merging data points of an entity have been practiced on the Web for many years. However, one of the central challenges is to find, extract, and disambiguate the smallest particles of knowledge. Here, modern Artificial Intelligence (AI) methods are used that can evaluate image and text datasets. However, there aren't yet so many available training datasets in the domain of privacy that make it possible to train these models. Furthermore, the individual pieces of information that can be found and assigned to an entity, both individually and in combination with other information, are to be evaluated in terms of their influence on the privacy of users on the Web, in order to ultimately be able to derive from the mass of information when a risk exists and users need to be warned.

The ADRIAN research project explores the use of DTs to model the vulnerability of individuals to privacy threats on the web (see Figure 1). The aim of the project is to warn users about the resulting threats that can arise from combined data and highlight the potential of DTs in mitigating such threats. In this paper, we discuss the concept of DTs and how they can be instantiated using information available on the web. We also present two datasets used by us and discuss the methods used to instantiate the DTs. Our paper reflects the idea of using DTs to model the vulnerability of individuals to privacy threats and the potential for further research in this area.

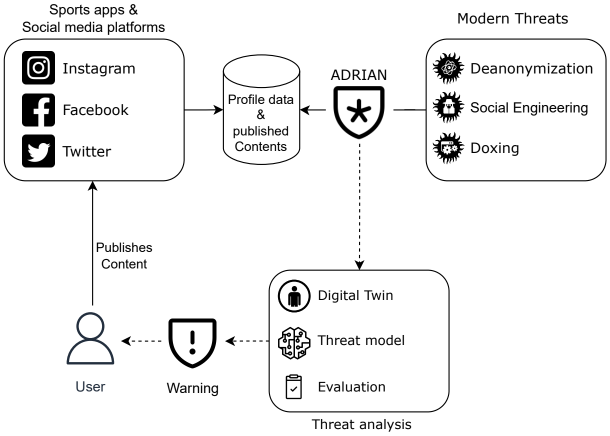


Fig. 1. ADRIAN Privacy Framework Overview

The paper is structured as follows: We discuss related work in the context of privacy research in Section II and describe our DT-based approach and datasets in Section III. Finally, we discuss our ideas in Section IV and draw our conclusions in Section V.

II. Related Work

In the following, we discuss the concept of DTs and show existing approaches for Information Extraction (IE) and modeling, both on text and image data.

A. Concepts for (Human) Digital Twin and Integration Levels

The term DT is an ambiguous concept found in various research and practice areas, such as mechanical engineering, medicine, and computer science [6]. Developments in AI have expanded its usage, with DTs defined as computer-based models that simulate, emulate, mirror, or '*twin*' the life of a physical entity, which may be an object, a process, a human, or a human-related feature [6]. DTs serve as living, intelligent, and evolving models that represent the virtual counterpart of a physical entity or process, used to monitor, control, and optimize their physical counterparts' life cycle.

There are three levels of integration for DTs [6]: (a) *Digital Model*, (b) *Digital Shadow* and (c) *Digital Twin*. A Digital Model is the basic representation of a physical object or system in the virtual world, without any automatic information flow between the virtual and physical worlds. Changes in the physical object must be manually updated in the digital model. A Digital Shadow takes this further and involves a unidirectional automatic information flow from the physical world to the virtual world. Sensors measure information from the physical model and transmit signals to the virtual model. A complete DT exists when the virtual and physical environments communicate bidirectionally, with information flowing automatically between both environments. This allows the DT to accurately reflect the current state and development of its physical counterpart.

With a look at sociotechnical systems, however, the matter becomes different. Sociotechnical systems encompass both human and machine components, making it relevant to explore the notion of a Human DT [7]. Despite its growing significance, there is no consensus on a standard definition or understanding of this concept [8]. The digital data that is available about individuals is often referred to under the term "*Digital Footprint*" or "*Digital Representation*" with the two terms often used interchangeably. These concepts are about data that is left by users on the Internet, often unknowingly and without clear identification or connection to the person. To differentiate the concepts of "*Digital Footprint*", "*Digital Shadow*", and "*Digital Twin*", several aspects can be considered, such as identifiability, active or passive data collection, individualized or aggregated evaluation, real-time or later analysis, decision-making authority, and comprehensive representation [7]. The Human DT aims to store and analyze relevant characteristic properties of an individual for a specific situation. This may include demographic or physiological data, competence or activity profiles, or health status [7].

In the ADRIAN research project, we understand the term as the digital representation of a real person, instantiated by information available on the Web [5]. In this context, the DT can never reflect the entire complexity of a real person, but it reproduces features that, alone or in combination with other attributes, can pose a threat to the real person. In this way, the DT makes it possible to model and measure the vulnerability of a person. The modeling of DTs is based on established and freely available standards of the semantic web, such as Schema.org and Friend of a Friend (FOAF). This makes it possible to easily connect and extend DTs. At the same time, the sheer number of possible sources of information, the quality of the data, and a multitude of contradictory data make modeling challenging. AI-driven methods from various fields, such as Natural Language Processing (NLP) and Computer Vision (CV), can help.

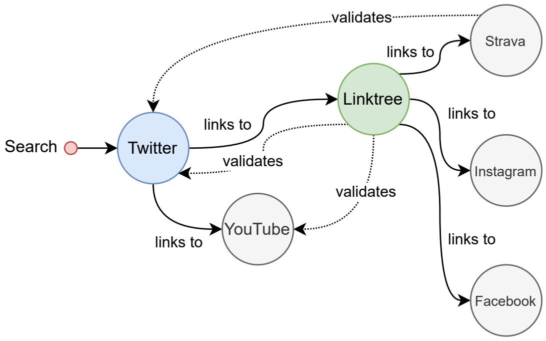


Fig. 2. Matching and Validation of User Profiles [9]

B. Instantiating Digital Twins using Semantic Techniques

Modeling DTs requires information that must be obtained from unstructured sources and heterogeneous datasets in the first step. In the second step, the information must be converted into a logical, semantic, and machine-readable structure. A key challenge is finding information about a person who can basically be on different social networks under different identities. Previous work in this area shows that interlinking between networks (see Figure 2) can be used in many cases to detect and verify matching profiles [9]. In addition, features, such as email addresses and usernames are reused (with variations) and, together with other features, enable profiles to be merged (e.g., profile pictures) [5].

For modeling DTs based on social network data, existing work on ontologies can be used. Different approaches have been devised to tackle the issue of interoperability by establishing shared standards for knowledge and information exchange. Numerous ontologies have been developed for the representation of personal networks, including FOAF and Semantically-Interlinked Online Communities (SIOC) [10], as well as for threats, exemplified by Structured Threat Information eXpression (STIX) [11]. Furthermore, Schema.org serves as an ontology for representing biographical information.

To extract data and map it into these structures (see Figure 3), different techniques are used depending on the type of data. Named Entity Recognition (NER) is essential for understanding and processing unstructured text data [12]. NER facilitates the identification and classification of specific entities like names, places, organizations, and dates found in unstructured text sources, such as social media content, web

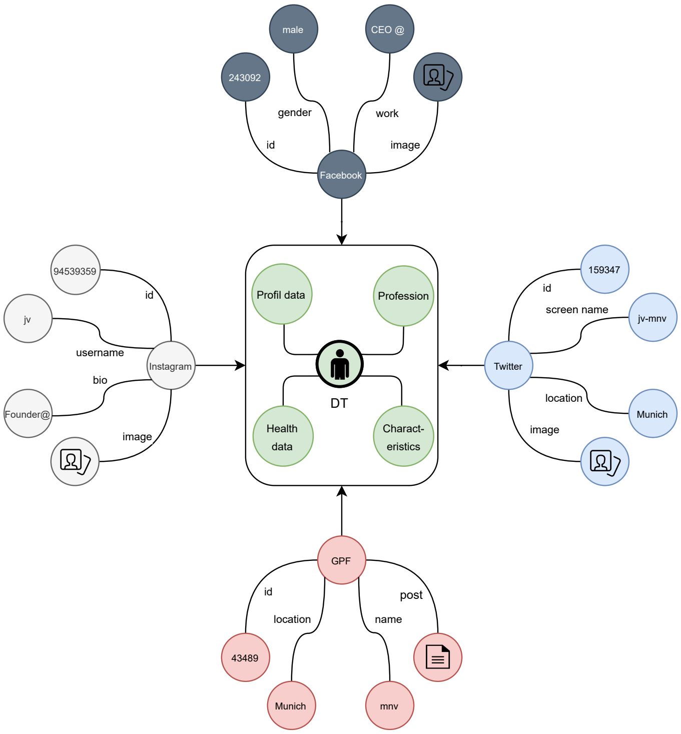


Fig. 3. Instantiated Human DT on Real Data

pages, and chat logs. However, the underlying NER models can be extended to recognize and extract different entities from diverse domains. Furthermore, recent work based on this also considers the connection between entities. One example is LUKE (Language Understanding with Knowledge-based Embeddings), which is a pre-trained contextualized representation model for words and entities based on a bidirectional transformer. It performs exceptionally well on various tasks related to entities, including relationship classification, where the model classifies the relationship between two entities.

Along with texts, images are a predominant type of data to handle. Models for understanding and processing visual information are currently being researched. Vision language models like Bootstrapping Language-Image Pre-training (BLIP) and Bootstrapping Language-Image Pre-training with frozen unimodal models (BLIP-2) have emerged for multimodal deep learning. These models are able to create representations between natural language text and visual input (e.g., images). This representation can be utilized for a wide range of tasks, including image captioning and visual question answering. Vision language models have the potential to enhance the analysis of images in the context of privacy and enrich DTs.

III. Enriching Digital Twins: Datasets

The two approaches to IE mentioned above are based on trained models whose training data do not necessarily consist of data relevant to the use case of this work. For this reason, models are fine-tuned to better fit the requirements of an application domain. In this section, we address two different but related studies that examine the privacy implications of CV and NLP. In Section III-A, we discuss our work on developing a dataset based on the VISPR dataset [13], with a focus on human characteristics that may have privacy implications for visual content. In Section III-B, we explore the use of NER to identify potential privacy risks in German patient forums, particularly with respect to health-related entities from textual information.

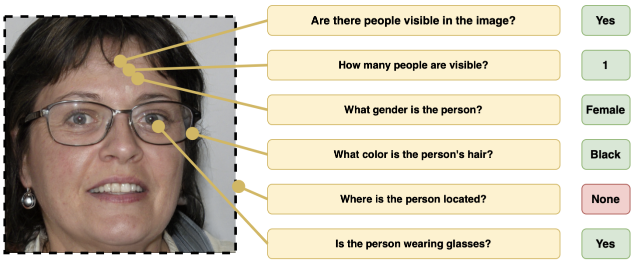

Are there people visible in the image?
Yes
How many people are visible?
1
What gender is the person?
Female
What color is the person's hair?
Black
Where is the person located?
None
Is the person wearing glasses?
Yes

Fig. 4. Attribute Extraction via VQA

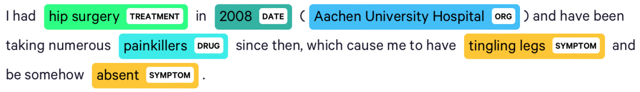
A. VQA Evaluation Dataset

Extracting personal information attributes from images offers great potential to instantiate DTs. Especially on social networks, a lot of portrait photos are shared or photos of vacations (e.g., the Eiffel Tower in the background) and new acquisitions (e.g., cars). We develop methods that can extract and categorize this information. However, suitable evaluation datasets are lacking so far. In our recent work, we have developed a dataset based on the VISPR dataset [13]. The main goal was to extract sensitive information based on various human characteristics from visual content. The original VISPR dataset consists of 22,167 images with a total of 115,742 labels containing 68 personal information attributes. We focused on annotating a subset of these attributes that relate specifically to human characteristics (see Figure 4). Our dataset contains annotations for nine key attributes, as described in Table I.

These attributes include approximate age (*a1_age_approx*), gender (*a4_gender*), eye color (*a5_eye_color*), hair color (*a6_hair_color*), presence of tattoos (*a11_tattoo*), partial nudity (*a12_semi_nudity*), full nudity (*a13_full_nudity*), skin color (*a17_color*), and physical disabilities (*a39_disability_physical*). For each attribute, we specified certain characteristics to ensure the consistency of the dataset. Based on the dataset, we then investigated the effectiveness of a vision-language approach for pre-selecting relevant images and extracting human attributes as data enrichment in the ADRIAN research project. The BLIP model evaluated performed well in detecting the number of people in an image and in detecting certain human attributes, such as age, hair color, nudity, and tattoos. However, it had difficulty determining eye color and distinguishing between different degrees of nudity. Overall, the BLIP model performed well in human characteristic recognition, but showed weaknesses in document extraction.

TABLE I
FURTHER ANNOTATED VISPR CHARACTERISTICS

Attribute Id	Annotations	# Images
a1_age_approx	*child, adult, elderly*	1711
a4_gender	*male, female*	1863
a5_eye_color	*blue, green, gray, brown*	1348
a6_hair_color	*black, blond, brown, gray, red*	1759
a11_tattoo	*yes, no*	45
a12_semi_nudity	*yes, no*	247
a13_full_nudity	*yes, no*	11
a17_color	*black, brown, white*	1914
a39_disability_physical	*yes, no*	41



I had hip surgery in 2008 (Aachen University Hospital) and have been taking numerous painkillers since then, which cause me to have tingling legs and be somehow absent .

Fig. 5. Entity Recognition for Health Data

B. NER Evaluation Dataset

In another recent study, we evaluated different NER models on a dataset we created based on German patient forum texts. The annotated entities, examples of the entities, and the number of labels are shown in Table II. The focus was on the annotation of medical entities relevant to the instantiation of the DT, including "*Anatomy*", "*Diagnosis*", "*Diseases*", "*Drug*","*Symptoms*", and "*Treatment*" (see Figure 5).

The evaluated German BERT (GBERT) [14] and XLM-RoBERTa [15] models showed very good performance in accurately extracting health-related data from the texts with high precision and recall values. GBERT showed a better result for labels with a high number (*Anatomy*, *Diagnosis*, *Diseases*, and *Symptoms*), while XLM-RoBERTa showed better performance on entities with a low number of labels (*Drug* and *Treatment*). Based on these NER models, which can detect privacy-relevant entities in German patient forums, we are able to improve existing DTs. However, our approach primarily targets the recognition of entities in the text as the first stage of our research. Our plan is to extend this work by extracting relationships between these entities in the next phase. By incorporating relationship extraction capabilities, we will be able to identify connections, such as drug-dosage and drug-disease associations. This improvement increases the accuracy of IE and the quality of DTs in the health domain.

TABLE II
NER DATASET FROM GERMAN PATIENT FORUMS

Entity	Examples	# Labels
Anatomy	Eyes, Vessels, Intestine	1294
Diagnosis	ECG, Ultrasound, Gastroscopy	635
Diseases	Flu, Hemorrhoids, Stroke	3022
Drug	Omeprazol, Fluoxetine, Ibuprofen	390
Symptoms	Headache, Fever, Tired	1249
Treatment	Eyeglasses, Massage, Physiotherapy	361

IV. DISCUSSION

The use of DTs as a measure of user vulnerability to privacy threats on the web is an exciting and promising area of research. However, it is important to recognize the limitations of these models, particularly in the context of the complexity of the real-world entities they seek to represent.

One challenge in creating DTs is the need to accurately represent the vast amount of personal data available online. While tools and NLP techniques have improved significantly in recent years (see Section II), they are still limited in their ability to accurately extract and interpret information. Additionally, the use of standards, such as FOAF [16] and Schema.org for modeling DTs can help to improve interoperability, but it does not address the problem of the quality and consistency of the data. Moreover, the use of DTs raises concerns about the ethical and legal implications of the collection and use of personal data. As the amount of information available online continues to grow, the risk of privacy violations also increases. The use of a semantic web further exacerbates this issue, as the interconnectedness of data can lead to the unintentional disclosure of sensitive information [5].

In the context of the ADRIAN research project, it is important to note that while we aim to create DTs that accurately reflect the real-world entities they represent, it is *impossible* to create a complete DT. This is because the vast amount of data available online makes it impossible to monitor every piece of information related to an individual. Moreover, the data sources are very complex, for example, user-generated texts, which, with spelling errors, neologisms, and incompleteness, make IE considerably more difficult. Or images, which can be of very different nature and are also prone to leading to incorrect conclusions (e.g., a poster of the Eiffel Tower in the background as an indication of the location). However, this does not diminish the importance of our work. By combining relevant information from different sources, we can identify potential privacy threats and warn users before they become victims.

In this paper, we present the two datasets that we use to train and evaluate our AI-based methods. The VQA dataset can be used to evaluate further VQA privacy models, which showed very promising results in person characteristic recognition. A major challenge in our evaluations was distinguishing real people from statues, which requires further experimentation. As far as document identification is concerned, the model recognizes passports and credit cards but has difficulty identifying driver's licenses and national ID cards. The second dataset, based on German patient forums, was used to train and evaluate the German GBERT and XLM-RoBERTa models. These models show high accuracy in detecting medical entities. The distribution of entities in patient forums is very unbalanced. GBERT performed better on entities that occur more frequently, while XLM-RoBERTa performed better on entities that occur less frequently.

V. Conclusion

In this paper, we have explored the concept of DTs and their potential use in modeling the vulnerability of individuals to privacy threats on the web. We have demonstrated how DTs can be instantiated using information available on the web and the effectiveness of these models in identifying potential privacy threats. Our study highlights the need for increased awareness of the privacy risks associated with sharing personal information online and the potential for DTs to be used as a measure of user vulnerability.

While our study provides valuable insights into the potential of DTs in the context of online privacy, there is still much work to be done in this area. Further research is needed to explore the effectiveness of DTs in mitigating privacy threats, as well as to investigate the ethical and legal implications of their use. Additionally, the development of standardized methods for instantiating and evaluating DTs would be a valuable contribution to the field [5].

In conclusion, DTs have the potential to revolutionize the way we approach privacy threats on the web. By providing a measure of user vulnerability, these models could help individuals take proactive steps to protect their privacy and prevent doxing and other privacy violations. We look forward to further research in this field.

Acknowledgment

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

References

[1] R. Jones, R. Kumar, B. Pang, and A. Tomkins, "'I know what you did last summer': query logs and user privacy," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*. Lisbon, Portugal: ACM Press, 2007, pp. 909–913.

[2] A. T. McKenna, A. C. Gaudion, and J. L. Evans, "The Role of Satellites and Smart Devices: Data Surprises and Security, Privacy, and Regulatory Challenges," *Penn State Law Review*, vol. 123, no. 3, 2019.

[3] J. Pinchot and D. Cellante, "Privacy Concerns and Data Sharing Habits of Personal Fitness Information Collected via Activity Trackers," *Journal of Information Systems Applied Research*, vol. 14, no. 2, pp. 4–13, 2021.

[4] M. Chen, A. Cheung, and K. Chan, "Doxing: What Adolescents Look for and Their Intentions," *International Journal of Environmental Research and Public Health*, vol. 16, no. 2, Jan. 2019.

[5] F. S. Bäumer, S. Denisov, Y. Su Lee, and M. Geierhos, "Towards Authority-Dependent Risk Identification and Analysis in Online Networks," in *Proceedings of the IST-190 Research Symposium (RSY) on AI, ML and BD for Hybrid Military Operations (AI4HMO)*, A. Halimi and E. Ayday, Eds., oct 2021.

[6] B. R. Barricelli, E. Casiraghi, and D. Fogli, "A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications," *IEEE Access*, vol. 7, pp. 167 653–167 671, 2019.

[7] G. Engels, "Der digitale Fußabdruck, Schatten oder Zwilling von Maschinen und Menschen," *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, vol. 51, no. 3, pp. 363–370, aug 2020.

[8] K. Feher, "Digital identity and the online self: Footprint strategies – An exploratory and comparative research study," *Journal of Information Science*, vol. 47, no. 2, pp. 192–205, oct 2019.

[9] S. Denisov and F. S. Bäumer, "The Only Link You'll Ever Need: How Social Media Reference Landing Pages Speed Up Profile Matching," in *Information and Software Technologies*, A. Lopata, D. Gudonienė, and R. Butkienė, Eds. Cham: Springer International Publishing, 2022, pp. 136–147.

[10] J. Breslin, U. Bojars, A. Passant, S. Fernández, and S. Decker, "SIOC: Content Exchange and Semantic Interoperability Between Social Networks," in *W3C Workshop on the Future of Social Networking*, jan 2009, barcelona, Spain.

[11] OASIS Open, "STIXTM Version 2.1 OASIS Standard," jun 2021, https://docs.oasis-open.org/cti/stix/v2.1/cs01/stix-v2.1-cs01.pdf, retrieved 06/01/23.

[12] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in *Proc. of the Seventh Conf. on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147.

[13] T. Orekondy, B. Schiele, and M. Fritz, "Towards a Visual Privacy Advisor: Understanding and predicting privacy risks in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3686–3695.

[14] B. Chan, S. Schweter, and T. Möller, "German's Next Language Model," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6788–6796.

[15] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proc. of the 58th Annual Meeting of the ACL*. Online: ACL, Jul. 2020, pp. 8440–8451.

[16] N. Pankong, P. Somchai, and M. Buranarach, "A combined semantic social network analysis framework to integrate social media data," jul 2012, pp. 37–42.