SN Computer Science (2024) 5:1109
https://doi.org/10.1007/s42979-024-03413-z

ORIGINAL RESEARCH

Mirroring Privacy Risks with Digital Twins: When Pieces of Personal Data Suddenly Fit Together

Frederik Simon Bäumer[1] · Sergej Schultenkämper[1] · Michaela Geierhos[2] · Yeong Su Lee[2]

Received: 29 May 2024 / Accepted: 14 October 2024

Abstract

With the proliferation of social media, more personal information is being shared online than ever before, raising significant privacy concerns. This paper presents a novel approach to identify and mitigate privacy risks by generating digital twins from social media data. We propose a comprehensive framework that includes data collection, processing, and analysis, with special attention to data standardization, pseudonymization, and the use of synthetic data to ensure privacy compliance. We apply and evaluate state-of-the-art techniques such as Large Language Models, Generative Adversarial Networks, and Vision-Language Models to generate synthetic but realistic social media data that support the construction of accurate and representative digital twins while ensuring strict privacy compliance. Our approach demonstrates the potential for digital twins to help identify and mitigate privacy risks associated with social media use. We discuss the value and feasibility of this concept and suggest that further refinement of the techniques and conditions involved is needed.

Keywords Digital twin · Privacy threat · Synthetic data · Social media profiling

Introduction

The Web is the place where people interact, discuss, and share various types of information. With many opportunities, but also risks, this exchange of information on the Web is creating a vast, freely accessible data source for a variety of data-driven applications. But to what extent can personal information from social media be accessed and aggregated? Ultimately, how likely is it that personal information, when aggregated, will pose a privacy risk to individuals, groups, or locations? Through almost every activity on the Web, users leave active and passive footprints [1]. These include obvious information such as images, text, and video that users knowingly upload. They also include information that is transmitted without user intervention, such as endpoint IP addresses.

Our work focuses on examining how individual pieces of information contribute to privacy threats over time, rather than just the immediate risk. The evolving potential threat is assessed by considering the likelihood that an individual will face such risks as more personal information is collected. It is assumed that a certain amount or combination of information is necessary to pose a significant cyber threat to an individual [2]. To address this, we explore four key research questions: first, determining the optimal starting point in the Social Web for efficiently identifying all relevant pieces of information before assembling them into a *digital twin* (DT); second, identifying the pseudonymization steps required to comply with privacy regulations; third, evaluating the extent to which synthetic data can be used as a substitute or complement for (re)training or fine-tuning AI models; and finally, understanding how to construct, model, instantiate, and enrich a DT from online social networks (OSNs).

We describe our exploratory data engineering approach to processing and aggregating social media data, analyze the

✉ Michaela Geierhos
michaela.geierhos@unibw.de

Frederik Simon Bäumer
frederik.baeumer@hsbi.de

Sergej Schultenkämper
sergej.schultenkaemper@hsbi.de

Yeong Su Lee
yeongsu.lee@unibw.de

[1] Applied AI Working Group, Bielefeld University of Applied Sciences and Arts, Interaktion 1, 33619 Bielefeld, North Rhine-Westphalia, Germany

[2] Research Institute CODE, University of the Bundeswehr Munich, Werner-Heisenberg-Weg 39, 85579 Neubiberg, Bavaria, Germany

Published online: 29 November 2024

patterns in which personal data is most commonly shared, and finally show how to instantiate a DT if desired. We have no interest in collecting personal data. We are even testing synthetic data to train AI models. In other words, we are interested in learning about the types of information that are explicitly or implicitly shared on each social media platform, and how easy or difficult it is to link data across platforms with some degree of confidence. In order to analyze large amounts of data, to make the disclosure of personal data on the Web visible, and to hold a mirror up to users, a number of data engineering challenges in both natural language processing (NLP) and data and knowledge engineering must be overcome.

Background

In the following, we provide some background information on the use of DTs for mirroring data disclosure and the trade-off with synthetic data.

Mirroring Disclosed Data with Digital Twins

The concept of DTs has evolved into a versatile tool used in various fields of research and practice. Originally rooted in mechanical engineering, medicine, and computer science, the term has expanded in scope with advances in artificial intelligence (AI) [2–4]. At their core, DTs can be defined as virtual representations of physical entities such as objects, processes, people, or human-related characteristics [3].

There are three levels of integration for DTs: Digital Model, Digital Shadow, and Digital Twin [3]. A Digital Model requires manual updates to reflect changes in the physical world and serves as a basic representation of a physical object or system in the virtual world. With a Digital Shadow, sensors transmit data to the virtual model, providing an automatic flow of information from the physical world to the virtual world. A complete DT ensures that the digital representation accurately reflects the current state of its physical counterpart by enabling bi-directional communication between the virtual and physical environments [3].

In the context of our work [2], DTs refer to digital representations of real individuals based on information available on the Web (*Human Digital Twin* [4, 5]). DTs focus on characteristics that pose a potential threat to individuals, although they may not capture the full complexity of an individual. DTs make it possible to measure potential risks by modeling an individual's vulnerability. The use of semantic web standards such as Schema.org [6] and FOAF (Friend of a Friend, [7]) allows DTs to be connected and extended. However, there are challenges. These include the large number of data sources, data quality, and conflicting information. Research shows that users unknowingly disclose a significant amount of information on the Web, highlighting the potential risks of aggregating and analyzing this data [8]. In summary, DTs are a powerful tool for modeling and understanding the vulnerabilities of individuals based on data disclosure. By understanding the impact of shared information on the Web, individuals can be better protected from potential threats and privacy risks.

Making Trade-Offs with Synthetic Data

Yet researchers often face barriers to using real data [9]. In some cases, the data is closed source, incomplete, unreliable, biased, or simply unavailable. There are also cases, such as human face datasets [10], where data cannot be shared or exchanged due to privacy concerns or potential security risks. This applies to almost all datasets we work on. But there are also datasets, e.g. photo databases with privacy properties [11, 12], that are open to the public. In most cases this is not the case. This is where synthetic data can play a valuable role in addressing these challenges.

Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm with the aim of solving a (set of) data science problem(s). [9]. In this way, researchers can use synthetic data for various applications without compromising the privacy or security of sensitive information. It has a wide range of potential applications in areas such as privacy, fairness, data augmentation, accelerating development cycles, and democratizing access to data [9].

From a privacy perspective, synthetic data is useful for training machine learning models. In addition to protecting the privacy of individuals, this allows organizations to comply with privacy regulations such as GDPR[1] and HIPAA[2] [9]. Similarly, when dealing with online threats, synthetic data can be used to simulate attacks like spearfishing and test the effectiveness of security measures without exposing real data. As a result, organizations can identify vulnerabilities in their systems and develop strategies to mitigate potential threats. By using synthetic data, organizations can stay ahead of threats and improve their overall security position [13].

There are limitations to the use of synthetic data [9] despite its value in research. One major concern is that synthetic data may not capture the complexity and nuance of real data [10]. Synthetic data can mimic the statistical properties of real data. However, it may not accurately represent the diversity and variability present in real datasets. This can introduce bias or inaccuracy into models trained on synthetic

[1] General data protection regulation: https://gdpr.eu.

[2] Health Insurance portability and accountability act: https://www.hhs.gov/hipaa/.

data, potentially compromising analytical effectiveness. Despite this assumption, research shows that it is possible to synthesize data with minimal domain gap so that trained models can generalize to real, in-the-wild data [10, 14].

Related Work

First, we present the state of the art in DT construction, before discussing pseudonymization approaches and the use of synthetic data in AI research.

Constructing Digital Twins

In a more futuristic view, a DT can be seen as a unique cybernetic representation of an individual that is created at birth and continuously updated throughout life, reflecting both genetic traits and health data [15]. This DT evolves synchronously with its human counterpart, capturing internal and external changes, medical treatments, and personal metrics, while adapting to natural growth patterns. Real-time data is captured and transmitted to cyberspace where the DT is updated, including medical exams, treatments, vaccinations, and other health metrics. This includes the integration of sensor data from wearable technology that reports on various physiological and emotional states. It can even include environmental and lifestyle factors such as diet and exercise habits. To ensure privacy and security, DT systems are interactive and support logins from individual users or authorized individuals using advanced authentication methods. DT provides health assessments and diagnostic feedback to both individuals and healthcare providers using a combination of IoT, big data analytics, and AI techniques such as neural networks. Finally, security within the DT system is rigorous. Multi-factor authentication and secure communication protocols are used to ensure reliable and protected access to DT data. This protects the integrity and confidentiality of sensitive personal health information [15]. However, this vision goes far beyond the definition of DTs here.

Modeling techniques used for DTs are ontologies and knowledge graphs, both of which play a central role in system functionality and efficiency [16]. Ontologies are frameworks that define and categorize data within a domain, ensuring consistent interpretation of data across systems. This consistency is essential when integrating data from disparate sources because all data must be understood in the same way. Ontologies facilitate the semantic interoperability: The ability of different systems to communicate and work together seamlessly using common definitions and structures. They also enable the integration of disparate data sources, such as sensors and databases. Knowledge graphs extend the functionality of ontologies by allowing not only the definition and organization of data, but also the direct linking of data elements. The result is a network of data points that can be traversed and queried, supporting complex data interactions and analysis. Although the technical approaches for instantiating the DT modeled by ontologies or knowledge graphs are of interest here, the pseudonymization of personal data is also crucial for GDPR compliance.

Pseudonymization

Pseudonymizing is a process of replacing personally identifiable information with a pseudonym or encoded value [17]. Mappings between encoded values and original identifiers are stored separately [17, 18]. Several techniques can be used to pseudonymize data, including numbering, random numbers or chunks, hashing, and encryption [19, 20]. Pseudonymization has been applied to various types of text using NLP methods in recent years. Recently, pseudonymization using large language models (LLMs) has been explored. For the CoNLL-2003 dataset, some researchers [21] experimented with different models, including spaCy, Flair, Seq2Seq, GPT3, and ChatGPT, which detect privacy-sensitive entities and replace them with items of the same type. They report that the LLM-based system has the best results for preserving text integrity. Others [22] used GPT-4 for the recognition and replacement of privacy sensitive information in radiology texts and reported the effectiveness of the named entity replacement by GPT-4. However, uploading the data to the server to identify and replace the named entities remains another privacy risk [21–23]. Otherwise, Llama 2 was shown to outperform its counterparts in the specific domain when properly fine-tuned [24].

OSNs are modeled by graphs consisting of a set of entities and the links between them, where the nodes represent the entities and the edges represent the relationship between the entities [25]. In addition, information about users such as age, gender, address, hobbies, education, and work experience is typically included in social network data. This user-related information is called a user profile. Therefore, in order to anonymize the graph structure, graph manipulation such as graph modification and graph generalization is applied on the one hand [26]. On the other hand, the profile data is treated as a table record where the direct identifiers and indirect attributes are typically hidden by the generalization, suppression, permutation, and perturbation [26]. In the context of pseudonymization, personal data should not be identifiable without additional information and must be protected as described above or may be replaced by synthetic data.

Use of Synthetic Data in AI

In recent years, AI algorithms, mainly using advanced techniques such as Generative Adversarial Networks (GANs)

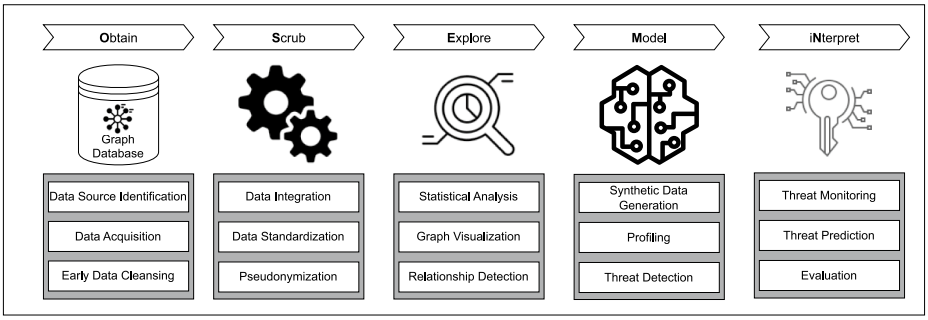

Fig. 1 Framework for authority-dependent risk identification and analysis [2]

and Variational Autoencoders (VAEs), have been applied to various fields. It has been expected to improve privacy, fairness, data augmentation, speed up development processes, and broaden data accessibility [9]. The field has developed rapidly, starting with the introduction of TextGAN, an algorithm for synthetic text generation through adversarial training [27]. In the following years, several specialized models were introduced [28–30].

However, synthetic data isn't always a substitute for real data. It can be altered to protect privacy and may not account for outliers or guarantee privacy without strict controls [9]. While synthetic data offers opportunities to improve the robustness of machine learning, more research is needed to understand its potential and limitations [9]. Recent studies have shown that the integrity of the performance of the model is not compromised if the training data is largely or entirely generated [10]. Furthermore, it is not necessary to train exclusively synthetic data. There have been studies where the training of models is on real data, but the evaluation is on synthetic data. This approach is used to detect privacy risks in images of individuals and groups [31].

Methodology

The research objectives and questions are presented below. This is followed by the theoretical framework that guides us in answering the research questions experimentally. Due to the correlational nature of the study and contextual appropriateness, a mixed methods approach was chosen.

Research Aim and Questions

Our focus is not on the acute threat, but on the contribution of each piece of information to the threat and its evolution over time. The potential threat over time is how likely an individual is to face a threat to their anonymity and privacy as more pieces of personal data are collected. We believe that it takes a certain amount of information, or a certain combination of information, to pose a cyber threat to an individual [2].

RQ1 To efficiently find all relevant pieces of information before assembling them into a DT, what is the appropriate starting point in the Social Web?

RQ2 What pseudonymization steps must be taken to comply with privacy regulations and ethical concerns?

RQ3 To what extent can synthetic data be used as a substitute or complement for (re)training or fine-tuning AI models?

RQ4 How to construct (i.e., model, instantiate, and enrich) a DT from OSNs?

Theoretical Framework

Here we take the approach to actively search, model, measure and highlight threats on the Web. In order to facilitate understanding and reusability, we have based the framework development on OSEMN (Obtain, Scrub, Explore, Model, and iNterpret) [32], which is a standardized model for data science research (Fig. 1).

The first step in the process is to identify relevant data sources and acquire enough of them. For us, this includes text, images and video from OSNs. Initial data cleansing is already included in this step, as the quality of the data can vary widely and user-generated data can usually be assumed to be of low quality. However, this cleansing is far from sufficient. For example, it does not include steps to normalize the data. This is done in the subsequent step. Since the collected data comes from different sources, data preprocessing is essential. Data standardization and pseudonymization

Table 1 Aggregated dataset based on Strava profiles and OSN search

	Twitter	YouTube	Xing	Facebook	Instagram
Users searched	13,902	13,902	13,902	13,902	13,902
Query limits	20	100	10	–	–
Users found	10,549	10,857	7862	–	–
Profiles retrieved	120,675	487,381	43,804	7970	5732

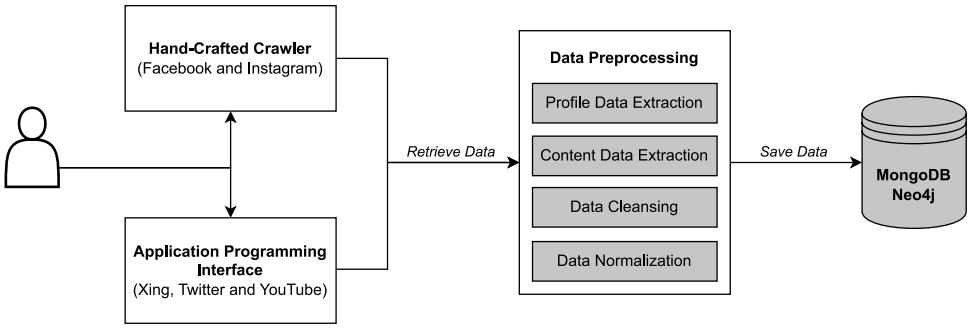

Fig. 2 Data collection and preprocessing pipeline

are two necessary preprocessing methods. Following standards and pseudonymization principles makes it possible to explore across datasets using existing tools and procedures.

In our use case [2], the most important task is to identify relevant data points to explore relationships in the data. We use a graph database to store and analyze data from multiple sources. Thanks to standards compliance, missing information, such as location details, can be integrated from other sources (e.g., Linked Open Data). Furthermore, the generated graphs provide a visualization of the networks and allow a graphical exploration of the datasets. But it is also likely that data collection will become a bottleneck. Therefore, it is planned from the beginning to generate synthetic data. For this purpose, a model is developed that explains the original data as well as possible. From this model, new data are generated that preserve important statistical properties of the original dataset. For our approach, it is necessary to identify and compare AI methods that are suitable for working with heterogeneous data. Using state-of-the-art AI models, we generate additional knowledge and analyze the relationships between user profiles in more detail. This knowledge and analysis can be used, for example, to identify groups of people in specific locations to detect potential threats. Finally, we investigate whether it is possible to identify potential risks at an early stage based on the data we generate and collect.

Research Approach

We conduct research in a practical, problem-solving manner. This is why we use mixed methods and take into account both quantitative and qualitative data. Our goal is to develop effective solutions that can be applied to real-world situations. By integrating different types of data, we aim to answer the above research questions (RQ1 – RQ3) experimentally, while pursuing a case study research design to address RQ4 and gain an in-depth understanding of DT construction based on personal information in OSNs.

Case Study

We implement a longitudinal case design to automatically monitor specific sports apps like Strava and analyze the collected data (Table 1).

Data is collected using either an API (Application Programming Interface), if available, or a hand-crafted crawler (Fig. 2). However, crawlers are limited in how much data they can collect in a given timeframe, so they may not capture all data points. We focus on retrieving users from an initial platform, in this case Strava. For users who registered through Facebook, Strava profiles contain specific information. We perform a direct image comparison using histogram and template matching techniques by searching for the username on Facebook and Instagram using a handcrafted crawler. Consequently, we only list profiles identified on these platforms. On other platforms, we use APIs to search usernames that return large numbers of results faster than our crawler. All discovered profiles are subsequently extracted, stored, and later pre-processed and matched. This allows us to cluster individuals and retrieve further information from the platforms (e.g., Twitter, see Table 2). This allows us to identify potential targets and assess their risk potential. This is done by processing text (e.g., tweets), images (e.g., selfies in front of buildings, maps), and

Table 2 Twitter dataset statistics

Statistics	#
Tweets	24,105,016
Total users	5,013,209
Tweets per user (min)	1
Tweets per user (mean)	3.57
Tweets per user (max)	4.77
Links per tweet (min)	0
Links per tweet (mean)	1.19
Links per tweet (max)	11
Links per user (min)	1
Links per user (mean)	6
Links per user (max)	51,575
Unique Linktree links	540,830

Table 3 Aggregated dataset based on Linktree links

	Twitter	LinkedIn	Facebook	Instagram
Profiles searched	5284	5069	5446	5583
Profiles retrieved	4956	4264	4909	4630

geospatial information (e.g., running routes). In other words, we are dealing with a heterogeneous dataset. Due to its composition, the requirements for processing methods are very different. During data analysis and knowledge extraction, a DT can be constructed. This generates extremely sensitive (meta) data. This information can be correlated with other data to determine the plausibility of a threat to a person or group of people. The technical implementation will combine methods from information retrieval with approaches from forensic linguistics and will use methods from network analysis and clustering to create new evaluation functions for the identification of subjects (people, places, etc.) on the basis of the disclosed information.

Data Collection

In general, we follow the data collection and processing pipeline shown in Fig. 2.

For this study, we used Twitter data with Linktree links from January 1, 2022 to February 23, 2022. We collected additional data for a longer period of time, from January 1, 2022 to April 19, 2023, to further validate our findings and expand our analysis. The result is a more comprehensive dataset, which is presented in Table 2.

We explored ways to collect data more efficiently and more effectively (Sect. Identify an Appropriate Entry Point for Collecting Data.) Therefore, in recent work [33], we evaluated the potential of social media landing pages as an entry point for data collection, resulting in the following dataset (Table 3).

Data scraping for scientific purposes is regulated by the GDPR in Art. 89, which provides exceptions for the processing of personal data for scientific purposes in the public interest. In particular, the collection of data without the participation of the data subject is subject to an exception for purposes of scientific research pursuant to Art. 14 GDPR. Accordingly, the provision of information is not required if it would make the objectives of the data processing for scientific research purposes impossible or seriously impair them. It will not be practically possible to inform all data subjects in the case of mass quantitative analysis of publicly accessible personal data, for example from OSNs (including Twitter) and other public portals on the Internet. Moreover, at least in the case of data that is not exceptionally sensitive, the benefit of informing the data subject would not be proportionate to the effort required to inform him. The information obligation under Art. 14 GDPR is therefore not applicable.

For particularly risky data processing, the GDPR requires a data protection impact assessment to be carried out (Art. 35 GDPR). This is necessary, for example, if extensive special categories of personal data (Art. 9 GDPR) are processed such as health data or data relating to religious or philosophical beliefs. We do not process data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, nor does it process genetic data, biometric data for the purpose of uniquely identifying a natural person, health data, or data concerning the sex life or sexual orientation of a natural person. However, due to ethical concerns, we will pseudonymize the names of individuals and their identifying attributes (Sect. Pseudonymization).

Modeling and Instantiating the Digital Twin

Our goal is to create a comprehensive and sophisticated DT that captures many facets of a person's digital presence. Based on state-of-the-art literature (Sects. Mirroring Disclosed Data with Digital Twins and Constructing Digital Twins), a basic, simplified model of the core elements and their interrelationships within our envisioned DT is illustrated in Fig. 3. It includes entities such as "*Person*", "*OnlineAccount*", "*EducationalOrganization*", "*Organization*", "*SocialMediaPosting*", and other related schemas connected by attributes and relationships such as "*hasRole*", "*hasEduOrg*", "*hasAccount*", and more. This model will be used as a template to develop a more complex and full-featured DT.

It is possible to create an initial DT for an individual by aggregating cross-platform data. In our previous research [34], we introduced a theoretical approach for linking user profiles across different platforms. This approach involved tracking and merging activities into DTs. In this study, we selected appropriate data points for profile matching, including names, usernames, location data, and images, using the previously created Linktree dataset.

Table 4 provides an overview of the data points available for profile matching and the number of data points included in the dataset. We use several techniques to assess the similarity of user profiles across social media platforms.

The degree of similarity between names and usernames is quantified using the Jaro–Winkler distance, a metric that calculates the proximity and shared characters between two strings, giving more weight to similarities near the beginning of the strings. This metric is used to assess the similarity of names and usernames because it effectively captures the partial similarity of shorter strings and giving more weight to initial characters [35], a feature particularly useful for handling variations in names and usernames. The thresholds of 0.75 for names and 0.60 for usernames are derived from previous research and empirical evaluations [36–38] and ensure a balance between precision and recall. These values were chosen based on studies suggesting that names

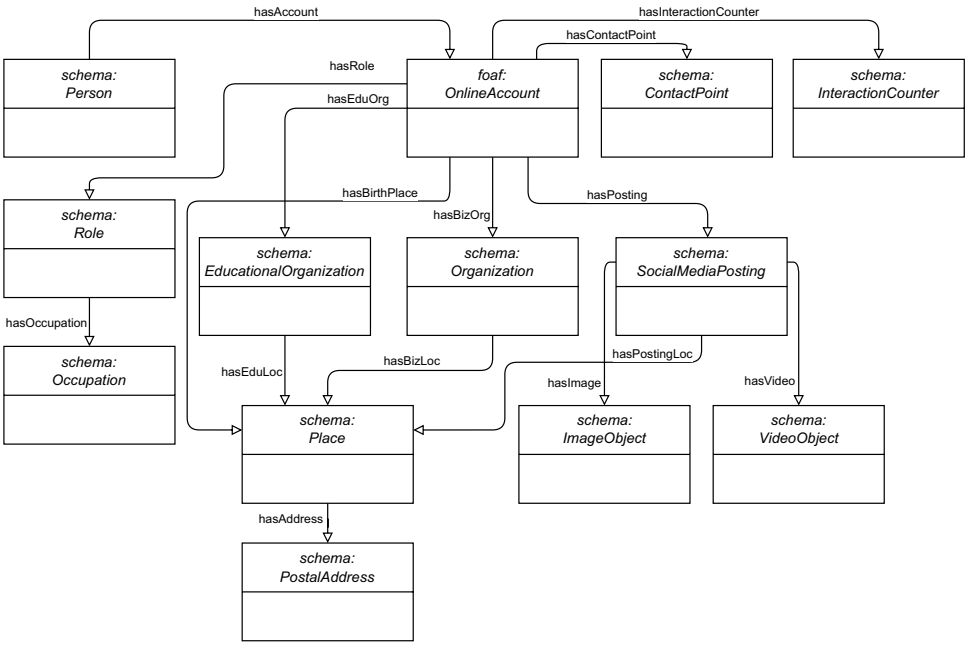

Fig. 3 Simplified model of the DT

Table 4 Available data points for profile matching

#	Twitter	LinkedIn	Facebook	Instagram
Name	4956	4264	4909	4630
Username	4956	4264	4909	4630
Location	4261	–	1369	–
Image	4562	4028	4908	4606

typically require a more stringent match than usernames, reflecting the greater variation found in usernames. As a result, the chosen thresholds strike an optimal balance, maintaining high precision while allowing for the necessary flexibility. This balance is critical to accurately identifying matches without overly constraining the algorithm, which would lead to missed legitimate matches.

For image similarity evaluation, both color and grayscale histograms as well as template matching techniques were implemented to improve the accuracy. We constructed a compact dataset of 100 entries to evaluate the image similarity metrics. The grayscale histogram threshold of 0.77, which achieved an F1 score of 91% and a precision of 96%, proved to be effective in discriminating similar images even in the presence of low-resolution noise. The stringent threshold of 0.22 for the color histogram is also consistent with this emphasis on precision, ensuring that false positives are minimized despite a lower F1 score. In addition, the template matching threshold of 22, while resulting in an F1 score of 68%, provides a very high precision of 99%, confirming the effectiveness of this technique for robust identity matching.

Location-based matching uses geographic distance with a threshold of 18.35 km based on empirical geocoding evaluation results, which is consistent with the accuracy generally observed in Twitter-based geocoding studies. This ensures a reliable, yet flexible approach to handling small geographic discrepancies due to variations in textual location data. Overall, these carefully tuned thresholds and metrics are designed to mitigate the noise and potential inaccuracies inherent in cross-platform data aggregation, ensuring a robust foundation for the construction of comprehensive DTs.

Data and Knowledge Engineering

To analyze the data and validate our hypotheses, we will apply the chosen methodology (Fig. 4). The instantiated variant of the DT presented in the previous section will be enriched with additional information. This data and knowledge engineering process can be divided into phases, each with specific challenges.

The initial phase focuses on collecting the necessary data, with an emphasis on determining appropriate entry points (**RQ1**). It is important to identify data repositories and sources that meet the research needs. With a clear understanding of the data, the next step is to build models that can simulate real-world behavior, protect personal information through pseudonymization (**RQ2**), and generate synthetic data (**RQ3**). This requires the selection of appropriate algorithms and tools that generate high-fidelity synthetic data while preserving the privacy of the original data sources. Finally, interpretation involves deriving actionable insights from the synthetic data. However, data enrichment to link the new information gained from profile extraction and

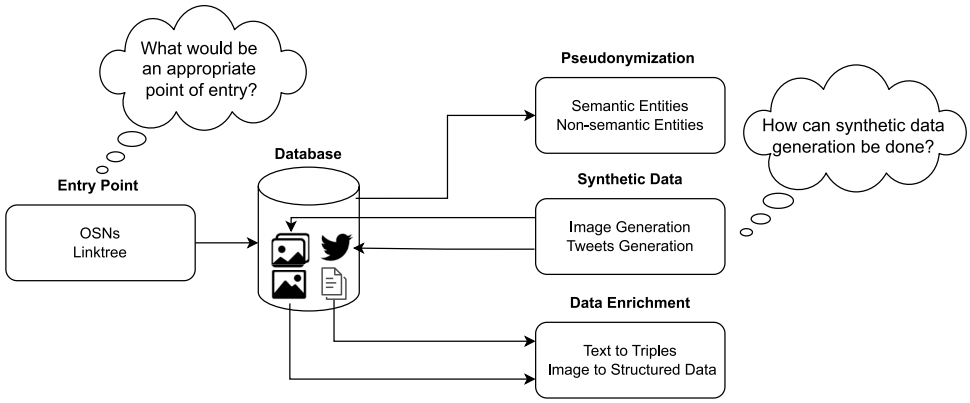

Fig. 4 Guiding questions in the data and knowledge engineering workflow

matching with open knowledge resources available on the Semantic Web is essential for the construction of the digital twin (**RQ4**). Therefore, we show how to obtain structured data from user-generated content and visual information from images for ontology integration.

Identify an Appropriate Entry Point for Collecting Data

To address our first research question (**RQ1**) experimentally, we test our hypothesis that Linktree is an appropriate starting point for a comprehensive search of linked personal information in different OSNs. Linktree is an SMRLP, which allows users to aggregate various links on a single page, thus requiring only one central link to access all of a user's profile links. We chose Linktree based on an initial analysis that showed it had a significantly higher link count in tweets compared to other existing platforms such as BrideURL, Linkin.bio, or ManyLink [39]. In addition, Linktree is the most prominent of the current SMRLPs, with over 30 million users worldwide.[3]

For this study, we obtained 540,830 links from our Twitter dataset (Table 2). In addition, we first collected data from all Linktree pages and extracted links to various platforms of interest, including Instagram, Facebook, Twitter, and LinkedIn (as shown in Fig. 5). We then analyzed the intersections of the provided links within each Linktree page. In Table 5, we can see that Instagram and Twitter contain the highest number of links, followed by Facebook, while LinkedIn has significantly fewer links. For further analysis, we limited our focus to personal links only. We were able to distinguish between personal and corporate accounts by examining URLs such as "linkedin.com/in" for individuals and "linkedin.com/company" for companies. There were 6,076 links to personal profiles and 6,237 links to company profiles in our dataset. After extracting the links, we applied a data collection and preprocessing pipeline (Fig. 2) to collect personal profiles from Twitter, LinkedIn, Facebook, and Instagram.

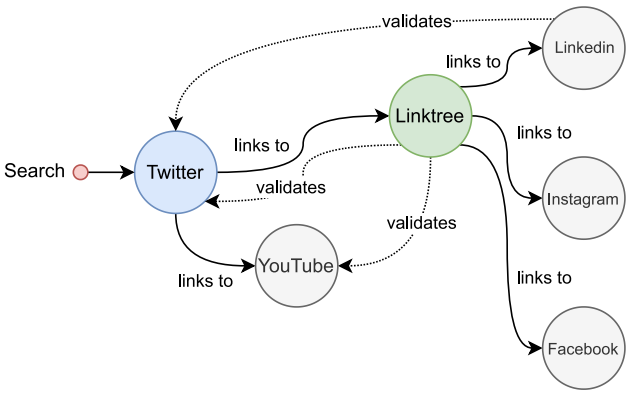

Fig. 5 Linktree data acquisition strategy

Table 5 Link overlaps within the resolved links

Platforms	# Links
Instagram	265,300
Facebook	136,743
Twitter	262,746
LinkedIn	37,626
Instagram ∩ Facebook	118,235
Instagram ∩ Twitter	201,791
Instagram ∩ LinkedIn	31,112
Facebook ∩ Twitter	103,817
Facebook ∩ LinkedIn	25,178
Twitter ∩ LinkedIn	20,262
Instagram ∩ Facebook ∩ Twitter	93,769
Instagram ∩ Facebook ∩ LinkedIn	22,747
Instagram ∩ Twitter ∩ LinkedIn	26,409
Facebook ∩ Twitter ∩ LinkedIn	21,504
Instagram ∩ Facebook ∩ Twitter ∩ LinkedIn	19,740

[3] https://productmint.com/linktree-statistics/, accessed: 2024-05-21.

Table 6 The top twelve domains on Linktree

Platform	# Links
Instagram	319,007
Telegram	298,828
Twitter	290,852
YouTube	285,464
Facebook	146,805
clicktotweet	104,856
TikTok	102,848
Spotify	68,038
bitly	61,354
Discord	47,236
Apple	41,988
Linkedin	40,476

Linktree is an appropriate entry point for data collection because the large number of links to different OSNs from a single profile indicates that Linktree can significantly improve the speed of user profile acquisition, matching, and validation (Table 5). For the platforms of interest here, the links include three of the top five OSNs. LinkedIn is still in the top twelve, as shown in Table 6.

Pseudonymization

In what follows, we answer our second research question (**RQ2**) by explaining what steps need to be taken to achieve compliance with the GDPR and with ethical principles. However, to avoid adversely affecting pre-trained profile matching algorithms and NLP approaches when using pseudonyms, we propose that semantic entities[4] should be treated differently from non-semantic entities when pseudonymizing.

Since the main purpose of pseudonymization is to connect users across different OSNs, the data points in Table 7 are pseudonymized.

Pseudonymization of Semantic Entities

Semantic entities were pseudonymized by semantically equivalent substitution, which was obtained by locally querying the LLM Llama 3.[5] This is possible because the OSN profile data is mostly structured and thus the named entity can be preidentified, but there are some exceptions that can be identified by preprocessing using simple regular expressions or applying NLP methods (e.g., worked as *Digital Marketing Lead* at *Microsoft*). Given the semantic entities, we build an additional information table by querying Llama 3 for the semantically equivalent entities. For all Llama 3 queries, we set the inference parameters as follows:

temperature: float = 0.1
top_p: float = 0.3
max_seq_len: int = 4,096
max_batch_size: int = 10

The use of an LLM to generate a semantically analogous substitution is very promising, as explicitly shown in Table 8. If we cluster two groups of person names using k-means[6] applied to vectors generated by e.g. GPT2,[7] we get two clearly distinct clusters per country as shown in Table 9. These are more potential replacements.

Table 7 Data points of profile data

	Data type	Entity type
User ID	DI	Non-semantic
Username	DI	Non-semantic
Cellphone	DI	(Non-)semantic
E-mail	DI	Non-semantic
Website	DI	Non-semantic
OSN ID	DI	Non-semantic
Display name	IA	Semantic
Short name	IA	Semantic
Nickname	IA	Semantic
Location	IA	Semantic
Organization	IA	Semantic
College	IA	Semantic
Secondary school	IA	Semantic
Date of birth	IA	Semantic
Address	IA	Semantic

DI direct identifier, *IA* indirect attribute

Pseudonymization of Non-Semantic Entities

There are two classes of non-semantic named entities. One is numeric only. This is usually the case with user IDs or other OSN IDs, for example. The other is alphanumeric. Examples include user names, email addresses, and web pages.

[4] A semantic entity is defined as a named entity with a natural, social, and geo-located reference in the real world. Thus, the typical named entities for person, organization, and location fall into this category. User ID and username, which exist only in the online world, are in the non-semantic category.

[5] https://llama.meta.com/llama3/.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[7] https://github.com/openai/gpt-2.

Table 8 Example of a prompt request and response by Llama 3

Request	
Role	System
Content	You always answer me with five person names that are as requent as the given person name in the same country. The generated names should have different first name and surname to each other. The output should follow the following json format: {"country": country, "input": input, "output": list_of_person_names}.
Role	User
Content	Alessandro Brofferio
Response	
Role	Assistant
Content	{country: Italy, input: Alessandro Brofferio, output: ["Andrea Bianchi", "Luca Ferrari", "Giovanni Rossi", "Marco Conti", "Francesco Marino"]}

Table 9 Cluster samples of generated names by country

	Cluster 1	Cluster 2
	Ahmed Benjelloun	Andrea Bianchi
	Omar Elghazi	Luca Ferrari
	Mohammed Amari	Giovanni Rossi
	Abdelhakim Rais	Marco Conti
	Youssef Ziani	Francesco Marino
Country	Morocco	Italy

In general, the pseudonymization of a non-semantic entity is performed by applying the addition and modulo operation to the given units. In the case of *digit units*, the operation is much simpler because the digit is replaced by another digit. To disguise this simple pseudonymization, we add a special digit to the beginning of the number so that the whole number can be changed. For addition and modulo operations, we use the prime number. The size of the prime number is easily adjustable. As shown in Table 10, the country code is treated as a special entity for the mobile phone and for WhatsApp. Although it has no direct referent in the real world, it denotes an artificial geolocation. Therefore, it should be treated as a kind of semantic entity.

For the pseudonymization operation on the *alphanumeric entities*, we have adapted Hill's cypher algorithm. However, while Hill's cryptography works on a single letter [40], we have developed a system that works on chunks of the alphabet. Besides the chunk size, our method differs from Hill's cryptography in two other aspects. First, the block size is not fixed, resulting in the variable block size. We determine the block size as the sentence size obtained by applying NLP to the given text. To specifically apply the variable block size to our profile data, it does not need a size limit due to its one-unit property. Second, we tokenize the given string according to our split mechanism. It divides the given string into two blocks consisting of consonants and vowels. The splitting is done by a simple regular expression that includes all vowels from the ISO-8859 language group. The alphabet 'y' is treated as a vowel in our method.

Table 10 Numeric examples and aliases

Attribute	Value	Pseudonym
User ID	100000656079902	24071379947782631
Mobile phone	+ 5546997087922	+ 552703024916093
WhatsApp	+ 4915175895942	+ 492402202724013

Using background knowledge and dictionary attack [25, 41], pseudonymized data can be exploited. In this paper, we present methods for pseudonymizing data that not only make pseudonymized data more real, but also make it more secure, since it is more difficult for attack models to associate pseudonymized data with original data. After applying the aforementioned pseudonymization strategies to our data collection, we were unable to measure any negative impact on profile matching or other NLP approaches. As a result, we have demonstrated that these pseudonymization steps are feasible for personal data on OSNs. They are also GDPR compliant and meet ethical concerns.

Synthetic Data Generation

Synthetic data is important for training machine learning methods and for research to compensate for rare data types or to study sensitive datasets. As mentioned in Sect. Making Trade-Offs with Synthetic Data, in the context of privacy risk mitigation and DT construction, it is important to recognize the potential biases and inaccuracies inherent in the use of synthetic data. One prominent limitation is that while synthetic data mimics real-world data, it does not always capture the full complexity, variability, and nuance of real human interactions and behaviors. This can result in models that are not fully representative of the real population, potentially introducing bias into both the data and the results of the analysis. However, despite these limitations, the use of synthetic data presents a necessary trade-off and compelling benefits, particularly when dealing with sensitive personal information. The ability to train and validate models without risking the exposure of individual identities is paramount to complying with privacy regulations, such as GDPR, and ensuring ethical research practices. In addition, as synthetic data generation techniques, such as those using GANs and other cutting-edge methods, advance, the fidelity and representativeness of synthetic datasets continue to improve.

While synthetic data are not perfect, they allow researchers to develop and refine privacy-preserving techniques and DT models more responsibly and ethically than would be possible with real data. It provides a pragmatic balance between the need for comprehensive data science and the imperative to protect individual privacy, making it an indispensable tool in current and future research efforts. In the following, we explore the extent to which synthetic data can be used to fine-tune AI models (**RQ3**). Below, we show how we use synthetic data in the context of image recognition and social media content.

Face Extraction and Generation

Investigating the ability of vision-language models (VLMs) to identify and extract sensitive personal information from images shared on OSNs is an important question that we have pursued in previous work [31]. To this end, we have evaluated new state-of-the-art VLMs, including BLIP-2 and InstructBLIP. In recent years, a number of VLMs have been introduced to advance multimodal deep learning, including vision transformer (ViT) [42], Contrastive Language-Image Pre-Training (CLIP) [43], and Bootstrapping Language-Image Pre-Training (BLIP) [44]. These models are capable of addressing a variety of challenges in both computer vision (CV) and NLP. BLIP introduces a novel approach for handling noisy web data through a method called Captioning and Filtering (CapFilt), which improves the quality of the training data. It also introduces a multi-modal mixture of encoder-decoders (MED), a multitask model that operates in three modes: unimodal encoder, image-based text encoder, and image-based text decoder [44]. The unimodal encoder for text and images is trained using an Image-Text Contrastive (ITC) loss, similar to the pre-training of the CLIP model. The image-grounded text encoder incorporates additional cross-attention layers to capture interactions between image and text, and is trained with an Image-Text Matching (ITM) loss to distinguish between positive and negative image-text pairs [44]. For image-based text decoders, it replaces bidirectional self-attention layers with causal self-attention layers, and uses the same cross-attention layers and feed-forward networks as encoders. The decoder uses LM loss to produce labels for given images.

BLIP-2 and InstructBLIP extend the integration of current VLMs with LLMs. BLIP-2 combines frozen image encoders with LLMs for pre-training, based on an architecture centered around the Querying Transformer (Q-Former), which effectively bridges the gap between visual and language modalities. Q-Former allows pre-trained vision and language models to be used for downstream tasks such as visual question answering and image-text generation without weight updates. This architecture's two-stage pre-training procedure results in outstanding performance across a variety of vision-language tasks. It supports zero-shot image-to-text generation with natural language instructions and has fewer trainable parameters than previous models. As a result, the model is capable of context-aware responses to text prompts. When using LLMs such as OPT and T5, BLIP-2 is limited to a context length of 512 tokens, which must be taken into account when creating detailed prompts and expected responses. InstructBLIP further refines BLIP-2 through instruction tuning, which uses an instruction-aware feature extraction method with Q-Former. This transforms data from 26 datasets into an instruction-based format and uses a balanced sampling strategy for the training dataset to optimize learning, improve zero-shot performance, and achieve state-of-the-art results when fine-tuned for specific tasks. InstructBLIP is compatible with models such as Vicuna [45], which in turn has been fine-tuned using the Llama base model [46].

We investigated the effectiveness of the model in following prompts to provide constrained and accurate responses. The methodology employed involves the use of an artificial dataset, derived from the Visual Privacy (VISPR) dataset and enriched with various privacy-related attributes, to extract relevant human attributes and sensitive information from images. In order to evaluate and improve our recognition methods, [31] we generate synthetic images of individuals with different attributes such as age and hair color (Fig. 6). These artificial images are generated using sophisticated algorithms incorporated into methods such as GANs [47], diffusion models, VAEs, and neural style transfer [48]. Tailored to specific goals and applications, each of these approaches offers unique advantages and presents different challenges.

We found that VLMs, especially those based on BLIP, are highly effective at recognizing people in images and identifying specific characteristics of people, such as age, gender, and eye color. However, challenges remain in recognizing documents, particularly country-specific documents such as driver's licenses. The models also showed shortcomings when dealing with synthetic images, suggesting that further refinements are needed to ensure accurate DT representations and to protect privacy.

Social Network Content Generation

We conducted experiments using current LLMs, such as GPT-4, to create synthetic tweets. We aimed to create a spectrum of tweets that could be used to build comprehensive user profiles, including profile data, images, and diverse content. Throughout our experiments, we carefully crafted prompts to elicit information across many categories, including family, health diagnoses, birthdays, age, job, employer, location, events, friends, and government. We used LLM's

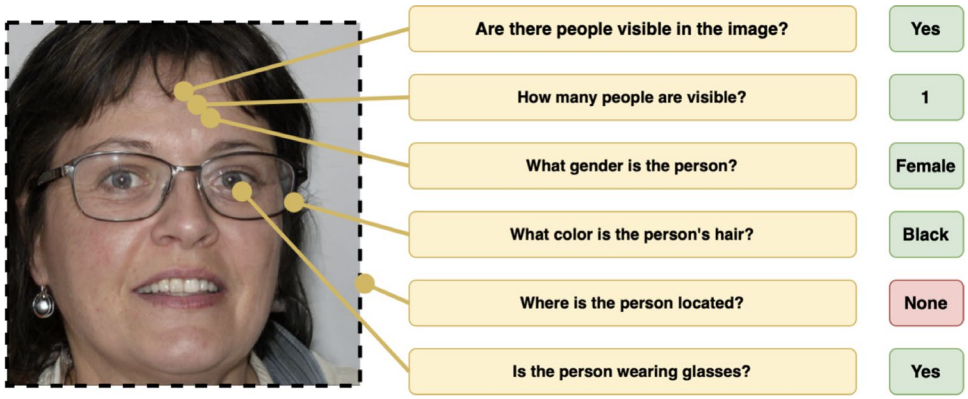

Fig. 6 Attribute extraction approach using VLMs [31]

sophisticated capabilities to provide tweets in JSON. The tweets were generated in batches of 50, and at the end of our experiment, we identified a significant challenge: many of the tweets had semantic similarities, and about 15% of the tweets were exact duplicates. Thus, RQ3 could not be adequately addressed here and will need to be explored further.

The potential solution to these challenges may lie in the introduction of additional variables into the tweet generation process. This could include incorporating randomized lists of topics for the LLM to latch onto, which could result in more varied content. Real-time data from a variety of sources, such as current events or trending popular culture, could also be added to the prompts to increase the relevance and authenticity of the generated tweets.

Personalization is the key to increasing authenticity. One solution could be to include user-specific characteristics in the prompts to instruct the LLM to generate content that reflects individual user identities. We are investigating various methods to verify the authenticity of these tweets, including NLP techniques, to ensure that the generated content truly simulates real-world tweets.

Building on this, we're also exploring the idea of using transfer learning mechanisms that could help the LLM learn from a large dataset of existing real-world tweets and apply this learning to the task of generating new tweets. By having the LLM extract optimal features from the authentic data pool, we can improve both the diversity and authenticity of our generated tweets.

Human Digital Twin Enrichment

To address our fourth research question (**RQ4**), how to construct a DT, and in particular how to enrich it, we conduct a series of experiments aimed at ensuring that the DT is as comprehensive as possible. Our approaches include fine-tuning current state-of-the-art LLMs and designing specific datasets to evaluate existing VQA techniques. We are also interested in using data mining techniques to transform unstructured data consisting of text and images into structured data. This transformation allows us to improve the usability and interoperability of data in the context of DT by integrating it with various ontologies, such as Schema.org and FOAF (Sect. Constructing Digital Twins).

Text to Triple Conversion

The huge amount of tweets represents a considerable amount of unstructured data in the form of text and images (Table 2). The goal of this work is to extract triples from texts such as tweets and to integrate them into an ontology-based representation. The information can then be used to construct a knowledge graph. We use GPT-4 and an open source LLM called Miqu-1-70B for instruction fine-tuning in this approach, shown in Fig. 7.

To create the dataset, a random sample of tweets from 300 users was selected from the database. Schema.org was used to define the properties of interest. The dataset was then created using a few-shot learning approach with GPT-4. The extracted triples were then manually evaluated in terms of the model's ability to correctly identify subject, predicate, and object [23]. The distribution is shown in Table 11 with respect to the properties of the final dataset.

Instruction fine-tuning is the process of adjusting a model so that it accurately follows and performs specific tasks as described in the instructions. For example, extracting a subject-predicate-object triple from a given tweet and defining the desired output to ensure the correct response and desired format in JSON. Using the Quantized Low-Rank Adaptation (QLoRA) technique in Fig. 7, the Miqu-1-70B model is fine-tuned to the dataset. We conducted experiments with different parameters to identify those that optimize performance in order to configure the training environment for our model. Table 12 shows the parameters that were identified.

A manual evaluation was performed on the extracted original Miqu-1-70B model as well as on the fine-tuned version. Table 13 shows the results of the evaluation with the fine-tuned (q4_k_m) model version.

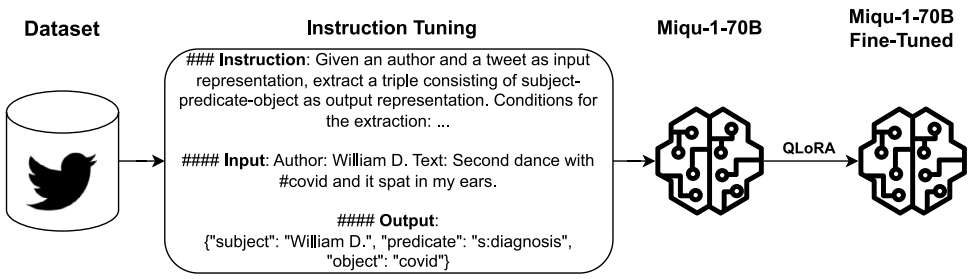

Fig. 7 Fine-tuning process

Table 11 Property frequencies in the Twitter dataset

Property	Frequency
s:location	298
s:spouse	227
s:attendee	225
s:colleague	192
s:worksFor	179
s:healthCondition	158
s:parent	139
s:knows	121
s:workLocation	104
s:jobTitle	89
s:sibling	70
s:alumniOf	63
s:diagnosis	52
s:children	51
s:nationality	42
s:birthDate	38
s:contactPoint	4
s:birthPlace	2

Table 12 Model training parameter configuration

Parameter	Value
Training batch size	4
Total training steps	700
Learning rate	2×10^{-4}
Weight decay	0.01
Optimizer	Paged AdamW
LoRA rank	64
LoRA scaling factor (*lora_alpha*)	16
LoRA dropout rate	0.1
LoRA modules	q_proj, k_proj, v_proj, gate_projo_proj up_proj, down_proj, lm_head

Improvements in the accuracy of relational triple extraction were observed after fine-tuning the Miqu-1-70B model. The micro and macro averages across all predicates also showed significant improvements, indicated in bold. This reflects a consistent improvement in the model's ability to accurately extract and categorize information from tweets. In both manual and automated evaluations, the fine-tuned model was shown to effectively and accurately extract triples in JSON format. The model can be used to analyze a large number of tweets to create a knowledge graph, as shown in Fig. 8.

Structured Data Extraction from Images

Images appear in both profiles and shared content. To further complete the DT, it is necessary to process images and relevant features within these images to facilitate further integration of ontology-compliant data. One approach we have identified as a potential solution is based on VQA, which allows the generation of answers to sequential questions and validations, including in-depth and control questions [49].

To evaluate these current VQA models, we developed our own dataset with different expressions for privacy attributes, based on the VISPR dataset [11, 12], which contains 68 image attributes suitable for classification tasks. To create the dataset, we selected relevant privacy-sensitive attributes, focusing on directly visible personal attributes and excluding textual information as shown in Table 14.

In addition, for the evaluation of the VQA models, we constructed different prompts and classes of interest. We categorize age into three groups based on [50] to define the values for the attributes and to maintain clarity and consistency in our image annotations: "child" (up to 16 years), "adult" (up to 55 years), and "elderly" (55 years and older), excluding more nuanced age groups to avoid subjective interpretations. Skin and hair color classifications are based on [51]. Simplicity is emphasized by avoiding overly specific color values that may lead to inconsistencies. Eye color classifications are based on [52], which stresses simple and accurate classifications. Overall, our approach is to use general categories for attributes such as age, skin, hair, and eye color. This minimizes complexity and increases the reliability of our annotations.

Personal attributes and documents were analyzed using three VQA models, BLIP [53], BLIP-2 [54], and InstructBLIP [55]. The methodology involved the testing of different prompt variations to optimize zero-shot performance and model accuracy in identifying details from visual input

Table 13 Precision of the Miqu-1-70B model before and after fine tuning

Property	Miqu-1-70B				Miqu-1-70B-FT			
	Subj.	Pred.	Obj.	#	Subj.	Pred.	Obj.	#
s:alumniOf	1.0000	0.5273	0.8182	55	1.0000	0.5556	0.8667	45
s:attendee	1.0000	0.9231	0.9231	26	1.0000	0.9565	0.9855	69
s:birthDate	0.9231	0.8462	0.7692	26	1.0000	0.9333	0.8667	15
s:birthPlace	1.0000	0.6667	0.6667	3	1.0000	1.0000	1.0000	4
s:children	1.0000	0.8929	0.8571	28	1.0000	0.9459	0.9189	37
s:colleague	1.0000	0.9412	0.9118	34	1.0000	0.8992	0.8992	119
s:contactPoint	1.0000	0.2000	0.2000	10	1.0000	0.3333	0.6667	3
s:diagnosis	1.0000	1.0000	1.0000	9	1.0000	1.0000	1.0000	11
s:healthCondition	0.9891	1.0000	1.0000	92	1.0000	0.9897	0.9588	97
s:jobTitle	0.8889	0.8889	0.8889	9	0.9565	0.8261	0.8261	23
s:knows	1.0000	0.7740	0.9231	208	1.0000	0.8716	0.9324	148
s:location	0.9834	0.8729	0.9282	181	1.0000	0.9318	0.9545	132
s:nationality	1.0000	0.8889	0.8889	9	1.0000	0.9167	0.9167	12
s:parent	1.0000	0.9722	0.9167	36	0.9811	0.9811	0.9623	53
s:sibling	1.0000	1.0000	1.0000	12	1.0000	0.9615	0.9231	26
s:spouse	1.0000	0.9853	0.9412	68	1.0000	0.9535	0.9302	86
s:workLocation	1.0000	0.7755	0.8367	49	0.9818	0.8727	0.9455	55
s:worksFor	0.9692	0.7077	0.6923	65	1.0000	0.8704	0.9074	54
Micro Avg	0.9902	0.8370	0.8913	920	**0.9970**	**0.9050**	**0.9312**	989
Macro Avg	0.9863	0.8257	0.8423	920	**0.9955**	**0.8777**	**0.9145**	989

A triple consists of Subj. = subject, Pred. = predicate, and Obj. = object

The hashtag indicates the number of supports

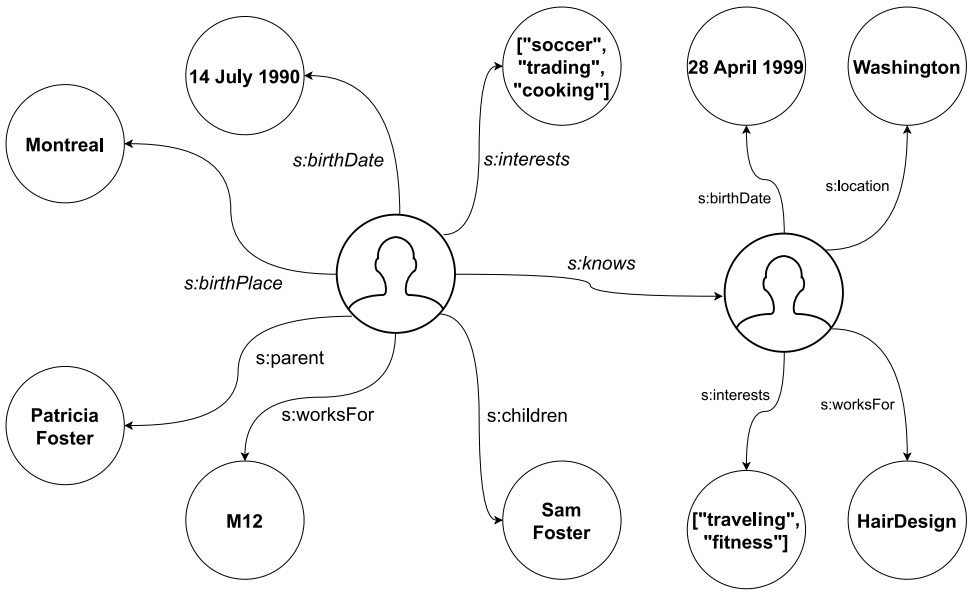

Fig. 8 Example of a knowledge graph constructed from triple extraction

(Table 15). To assess the privacy VQA performance of BLIP, BLIP-2, and InstructBLIP, we measured their precision, recall, and F_1 scores. Our experiments were conducted using an A6000 graphics card. BLIP is the most compact of the three, with a size of 1.54 GB. In contrast, BLIP-2 (flan-t5-xxl version) and the InstructBLIP model (Vicuna-13b version) are significantly larger, at 49.44 GB and 49.49 GB, respectively. When it comes to processing speed, BLIP outperformed the others by completing the task in 1:06 h for each attribute with three different prompts. BLIP-2 followed, requiring 2:26 h, while InstructBLIP lagged behind at 3:15 h. In our analysis of prompt effectiveness, BLIP-2

Table 14 Selected VISPR dataset attributes

Attribute	# of Img.
a1_age_approx	1711
a4_gender	1863
a5_eye_color	1348
a6_hair_color	1759
a11_tattoo	45
a12_semi_nudity	247
a13_full_nudity	11
a17_color	1914
a29_ausweis	47
a30_credit_card	97
a31_passport	263
a32_drivers_license	70
a33_student_id	70
a39_disability_physical	41

excelled with simpler and more concise prompts, whereas InstructBLIP yielded better results with detailed prompts. The results presented below are based on the highest F_1 score achieved with the optimal prompt. While BLIP is a small model that uses a text transformer that is initialized by the BERT model [56], it is worth noting that BLIP-2 and InstructBLIP use much larger LLMs, such as Flan-T5-XXL [57] and Vicuna-13b [45]. When evaluating the prompts, we found that BLIP-2 performed better with simple and short prompts, while InstructBLIP performed better with more detailed prompts. The prompt with the highest F_1-score is used as the basis for the following results. The results of person recognition and attribute extraction are shown in Table 16. The performance of the top model in each category is shown in bold.

When we analyzed documents, we used a dataset that contained 536 different document types. Table 17 shows the results for the BLIP-2 model, the best performing model in our study. This table provides a detailed breakdown, highlighting the progress made in analyzing documents.

Our person detection dataset contained 1000 images, 46 of which were removed due to ambiguities, such as unrecognizable in certain situations, like racing. As shown by the InstructBLIP model, person detection was highly reliable. However, for different personal attributes and documents, the results of the study were inconsistent. For personal attributes such as "age" and "hair color", the results were generally good, even though these attributes are difficult to annotate accurately. Without additional context, it was difficult to distinguish between age groups and specific hair colors. The "eye color" attribute was particularly challenging, with annotators having to skip over numerous images in which it was impossible to accurately determine the color, even when using high-resolution images and zoom functionality. In contrast, very good or fair results were obtained for attributes that required yes/no answers, such as "nudity" and "tattoos". These were easier to annotate. However, the annotation process was complicated by the fact that the definition of "semi-nudity" turned out to be ambiguous. As a result, cases such as a bare male torso were often not recognized by the BLIP model. Due to their standardized appearance across countries, "passports" and "credit cards" were reliably recognized for document identification. However, due to significant variance in their appearance across the world, "driver's licenses" and "national identification cards" showed poor recognition. The development of an approach specifically for German documents is aimed at creating the basis for the possible development of tailored approaches for other countries. In order to effectively handle the wide variety of document formats, the models need to be fine-tuned. So there is still a lot of work to be done. Finally, the structured data extracted from the DT can now be validated, further enriched and used within ontologies.

Table 15 Prompt examples and defined class labels

Attribute	Prompt	Class Labels
a1_age_approx	How old is the person?	[Child, adult, elderly]
a4_gender	What is the gender of the person?	[Male, female]
a5_eye_color	What color are the person's eyes?	[Blue, green, gray, brown]
a6_hair_color	Which color is the hair of the person?	[Black, blond, brown, gray, red]
a11_tattoo	Does the person have a tattoo?	[Yes, no]
a12_semi_nudity	Is the person partially nude?	[Yes, no]
a13_full_nudity	Is the person fully nude?	[Yes, no]
a17_color	What is the skin color of the person?	[Black, brown, white]
a29_ausweis, a30_credit_card, a31_passport, a32_drivers_license, a33_student_id	Which document is in this picture?	[National identification card, credit card, passport, driver's license, student ID]
a39_disability_physical	Does the person have a disability?	[Yes, no]

Table 16 Person detection and attribute results

	Precision	Recall	F_1-score	Support
Person detection				
BLIP	0.9602	0.9602	0.9602	954
BLIP-2	0.9503	0.9599	0.9551	954
InstructBLIP	**0.9608**	**0.9707**	**0.9658**	954
Age				
BLIP	**0.9137**	**0.9345**	**0.9240**	1,666
BLIP-2	0.9079	0.9286	0.9181	1,666
InstructBLIP	0.8838	0.9040	0.8937	1,666
Gender				
BLIP	**0.9725**	**0.9824**	**0.9774**	1,766
BLIP-2	0.9719	0.9807	0.9763	1,766
InstructBLIP	0.9697	0.9796	0.9746	1,766
Eye color				
BLIP	**0.8132**	**0.8391**	**0.8260**	628
BLIP-2	0.7708	0.7879	0.7792	628
InstructBLIP	0.7404	0.7608	0.7504	628
Hair color				
BLIP	**0.8798**	**0.8865**	**0.8831**	1,577
BLIP-2	0.7202	0.7231	0.7216	1,577
InstructBLIP	0.7988	0.8032	0.8010	1,577
Skin color				
BLIP	**0.9501**	**0.9645**	**0.9573**	1,858
BLIP-2	0.8787	0.8889	0.8838	1,858
InstructBLIP	0.7637	0.7692	0.7665	1,858
Tattoo				
BLIP	0.8222	0.8222	0.8222	90
BLIP-2	0.8222	0.8222	0.8222	90
InstructBLIP	**0.8555**	**0.8555**	**0.8555**	90
Semi nudity				
BLIP	0.7974	0.8009	0.7991	462
BLIP-2	**0.8297**	**0.8333**	**0.8315**	462
InstructBLIP	0.7780	0.7814	0.7797	462
Full nudity				
BLIP	**0.9545**	**0.9545**	**0.9545**	22
BLIP-2	0.9090	0.9090	0.9090	22
InstructBLIP	**0.9545**	**0.9545**	**0.9545**	22
Disability physical				
BLIP	0.7439	0.7439	0.7439	82
BLIP-2	0.8048	0.8048	0.8048	82
InstructBLIP	**0.8293**	**0.8293**	**0.8293**	82

Findings and Discussion

Creating a DT will be a source of sensitive data, and being able to correlate this data will become a key enabler of data-driven applications. Our work aims to reduce this risk by raising awareness of the interrelationships between disparate pieces of public information. We monitor selected OSNs, analyze the data they collect, correlate it with other social media profiles, and build person-centric data networks (i.e., DTs). The goal is to identify potential targets of cyber threats and classify their potential risk based on the available data. This requires pseudonymization and the use of synthetic data, as personal data plays a central role. But where to start collecting data? What is the right starting point to find all the relevant pieces of the puzzle before they are assembled into a DT and a (perhaps incomplete) picture is created? What does it mean to handle personal data according to GDPR? What privacy and ethical measures need to be taken? How can synthetic data be used in AI models? And how can we complete the DT puzzle?

Table 17 Detailed results for documents by BLIP model

	Precision	Recall	F1 score	Support
Credit card	0.9773	0.9053	0.9399	99
Driver's license	1.0000	0.6418	0.7818	94
Nat. ident. card	0.2866	0.9574	0.4412	46
Passport	0.9951	0.7739	0.8707	213
Student ID	1.0000	0.6818	0.8108	95

Research Contributions

In Sect. Research Aim and Questions, we asked the following questions to examine:

1. To efficiently find all relevant pieces of information before assembling them into a DT, what is the appropriate starting point in the Social Web? [**RQ1**]
2. What pseudonymization steps must be taken to comply with privacy regulations and ethical concerns? [**RQ2**]
3. To what extent can synthetic data be used as a substitute or complement for (re)training or fine-tuning AI models? [**RQ3**]
4. How to construct (i.e., model, instantiate, and enrich) a DT from OSNs? [**RQ4**]

Although we were able to provide answers to the aforementioned research questions in Sect. Data and Knowledge Engineering, there are significant challenges to implementing DTs while ensuring privacy and complying with regulations. The fact that not all data can be used directly in the creation of DTs is one of the main challenges. It is important to protect the privacy of the individuals represented by the DTs as well as the research process. This requires the development of techniques for the protection of sensitive information while still allowing for meaningful analysis and insight. Promising solutions to these challenges are synthetic data and pseudonymization, as shown. In our work, we have developed tools and techniques for

processing and analyzing text and images from social media to create DTs. We have also developed a framework to guide this process, from data acquisition to the creation of the final DT. Our results show that these approaches can extract relevant information and create meaningful representations of web users. Thus, the creation of DTs from social media data offers a powerful tool for understanding and mitigating privacy risks on the Web. However, realizing this potential will require ongoing research and development to address the significant challenges of privacy, data quality, and ethical use. By continuing to refine our techniques and frameworks, and by engaging in multidisciplinary collaboration and public dialogue, we can work toward a future where DTs are used to empower and protect Web users, rather than to exploit or harm them.

Limitations

However, there are limitations to our current approach. The accuracy and completeness of DTs depend on the quality and quantity of data available. Not all users have the same amount or type of information, which can lead to gaps or biases in the resulting DTs. Furthermore, while our techniques aim to protect privacy, there is always a risk that sensitive information could be inadvertently exposed or that the DTs could be misused for malicious purposes. To address these limitations, future work should focus on refining the techniques for data collection and analysis, particularly in the areas of text and image processing. This could include the development of more advanced NLP and CV algorithms that can better understand the context and meaning of social media data.

While valuable for research, there are limitations to the use of synthetic data [9]. One major concern is that synthetic data may not capture the complexity and nuance of real data [10]. Synthetic data can mimic the statistical properties of real data. However, it may not accurately represent the diversity and variability present in real data sets. This can introduce bias or inaccuracy into models trained on synthetic data, potentially compromising analytical effectiveness. However, research shows that it is possible to synthesize data with minimal domain gap so that trained models can generalize to real, in-the-wild data [10, 14].

Using background knowledge and dictionary attack [25, 41], pseudonymized data can be exploited. Therefore, pseudonymized data is stored on encrypted disks and accessible only to selected researchers.

Despite the strengths of the evaluation metrics and profile matching thresholds, several open questions remain. For example, the impact of different linguistic and cultural name variations on the effectiveness of the Jaro-Winkler distance warrants further investigation. While the current threshold settings work well within the dataset used, extending this research to a broader international dataset may reveal the need for additional adjustments. Another open question concerns the scalability of the image similarity techniques. The effectiveness observed with the compact dataset may face challenges when applied to larger datasets or images with greater variability in resolution and quality. Finally, the location-based matching threshold, while based on empirical research, could benefit from contextual adjustments depending on urban density or variations in data quality across regions, which may require dynamic thresholds rather than a static distance value. Addressing these open questions will be critical to refining our approach and ensuring the adaptability and robustness of the digital twin model as it scales and evolves.

Privacy and Ethical Concerns

In the case of mass quantitative analysis of publicly accessible *personal data*, for example from OSNs, it will not be practically possible to inform all data subjects. Moreover, we do not process data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a natural person, health data or data concerning the sex life or sexual orientation of a natural person. However, due to ethical concerns, we pseudonymize the names of individuals and their identifying attributes (Sect. Pseudonymization). Since the GDPR and the requirement of data protection by design and by default [18], the privacy-preserving data process is an essential component of a software development aimed at analyzing the personal information. In this paper, we presented methods for pseudonymizing data that not only make pseudonymized data more real, but also make it more secure, since it is more difficult for attack models to associate pseudonymized data with original data. They are also GDPR compliant and meet ethical concerns.

Synthetic data is useful for training machine learning models from a privacy perspective. Similarly, when dealing with online threats, synthetic data can be used to simulate attacks such as spearfishing and test the effectiveness of security measures without exposing real data. But it can even play an important role in cyber attacks. There are AI-powered tools that specialize in social media intelligence, using advanced facial recognition algorithms to extract personal information from social media platforms [58]. In addition, an AI-powered bot operated undetected for an entire week on Reddit [59], using synthetically generated data to post comments and engage in conversations with users. Remarkably, it was able to convince several people of its human identity. For example, automated chatbots can be used to trick users and steal personal information. By convincingly mimicking human interactions, these AI-powered entities can gain

trust and encourage users to reveal sensitive data such as passwords, financial information, or personal identifying information [2]. This capability is a serious risk. For this reason, our synthetic privacy policy is very strict that data is only used for predefined purposes and deleted after use. However, despite these limitations, the use of synthetic data presents a necessary trade-off and compelling benefits, especially when dealing with sensitive personal data. To comply with privacy regulations, such as GDPR, and ensure ethical research practices, the ability to train and validate models without risking the exposure of individual identities is crucial. While not perfect, synthetic data allows researchers to develop and refine privacy-preserving techniques and DT models more responsibly and ethically than with real data.

In *combination with other data*, even small quantities of sensitive data may become hazardous. Sharing sensitive content can put users at risk of having their personal information exposed or misused, which can lead to various threats such as deanonymization or doxing [60]. Doxing occurs when previously private information about individuals is made public, often with the intent to harm, humiliate, or harass them. By exposing social media users to a mirror (i.e., a DT) that shows the extent to which they voluntarily, and sometimes unknowingly, share personal information without protection, and by highlighting the potential for misuse, we create an awareness of how to be more careful with data in the future.

Conclusion and Outlook

Our research focuses on developing DTs that replicate web users and their behaviors. This method can make privacy risks more visible and help prevent doxing by highlighting already available web data. Key challenges include maintaining privacy and regulatory compliance while building DTs. The use of synthetic data, which mimics real data without compromising privacy, and pseudonymization, which replaces identifiable information with pseudonyms, are promising solutions.

We have developed tools for processing and analyzing social media text and images to build DTs, resulting in detailed knowledge graphs and enriched image data. Our datasets, collected from different platforms and protected by pseudonymization, were used to refine our models. Despite demonstrating effectiveness, our approach has limitations in terms of data quality and completeness, as well as potential privacy risks.

Future work should focus on improving data collection and analysis methods using advanced NLP and computer vision algorithms. Further research should develop robust methods for quantifying privacy risks and ensuring that they are effectively communicated and mitigated. Ongoing advances in pseudonymization, synthetic data generation, large-scale language models, and computer vision could improve DT capabilities. In addition, a customized threat detection scoring system based on the ENISA framework could standardize privacy risk assessment.

In conclusion, the creation of DTs from social media data is a promising tool for understanding and mitigating privacy risks, which requires continuous research and ethical considerations to empower and protect web users.

Acknowledgements This research is funded by dtec.bw - Digitalization and Technology Research Center of the Bundeswehr which we gratefully acknowledge. dtec.bw is funded by the European Union - NextGenerationEU.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Iordanou C, Smaragdakis G, Poese I, Laoutaris N. Tracing cross border web tracking. In: Proceedings of the Internet Measurement Conference 2018. IMC '18, pp. 329–342. Association for Computing Machinery, New York, NY, USA. 2018. https://doi.org/10.1145/3278532.3278561.
2. Bäumer FS, Denisov S, Su Lee, Y, Geierhos M. Towards authority-dependent risk identification and analysis in online networks. In: Halimi A, Ayday E (eds) Proceedings of the IST-190 Research Symposium (RSY) on AI, ML and BD for Hybrid Military Operations (AI4HMO). 2021.
3. Barricelli BR, Casiraghi E, Fogli D. A survey on digital twin: definitions, characteristics, applications, and design implications. IEEE Access. 2019;7:167653–71. https://doi.org/10.1109/ACCESS.2019.2953499.
4. Schultenkämper S, Bäumer FS. Privacy risks in german patient forums: a NER-based approach to enrich digital twins. In: Lopata A, Gudonienė D, Butkienė R (eds) Information and software technologies, pp. 113–123. Springer, Cham. 2024. https://doi.org/10.1007/978-3-031-48981-5_9.
5. Lauer-Schmaltz MW, Cash P, Hansen JP, Maier A. Towards the human digital twin: definition and design–a survey. 2024.https://doi.org/10.48550/arXiv.2402.07922.

6. Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Commun ACM. 2016;59(2):44–51. https://doi.org/10.1145/2844544.
7. Pankong N, Prakancharoen S, Buranarach M. A combined semantic social network analysis framework to integrate social media data. In: Knowledge and Smart Technology (KST). 2012:37–42. https://doi.org/10.1109/KST.2012.6287736.
8. Bäumer FS, Grote N, Kersting J, Geierhos M. Privacy matters: detecting nocuous patient data exposure in online physician reviews. In: Damaševičius R, Mikašytė V (eds) Information and Software Technologies. 2017:77–89. Springer, Cham. https://doi.org/10.1007/978-3-319-67642-5_7.
9. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, Cohen SN, Weller A. Synthetic data—what, why and how?. 2022.https://doi.org/10.48550/arXiv.2205.03257.
10. Wood E, Baltrušaitis T, Hewitt C, Dziadzio S, Cashman TJ, Shotton J. Fake it till you make it: face analysis in the wild using synthetic data alone. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021;3661–3671. https://doi.org/10.1109/ICCV48922.2021.00366.
11. Orekondy T, Schiele B, Fritz M. Towards a visual privacy advisor: understanding and predicting privacy risks in images. In: 2017 IEEE International Conference on Computer Vision (ICCV). 2017;3706–3715. https://doi.org/10.1109/ICCV.2017.398.
12. Orekondy T, Fritz M, Schiele B. Connecting pixels to privacy and utility: automatic redaction of private information in images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018;8466–8475. https://doi.org/10.1109/CVPR.2018.00883.
13. Yamin MM, Ullah M, Ullah H, Katt B. Weaponized AI for cyber attacks. J Inform Secur Appl. 2021;57: 102722. https://doi.org/10.1016/j.jisa.2020.102722.
14. Tremblay J, Prakash A, Acuna D, Brophy M, Jampani V, Anil C, To T, Cameracci E, Boochoon S, Birchfield S. Training deep networks with synthetic data: bridging the reality gap by domain randomization. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018;1082–10828. https://doi.org/10.1109/CVPRW.2018.00143
15. Shengli W. Is human digital twin possible? Comput Methods Prog Biomed Update. 2021;1: 100014. https://doi.org/10.1016/j.cmpbup.2021.100014.
16. Karabulut E, Pileggi SF, Groth P, Degeler V. Ontologies in digital twins: a systematic literature review. Fut Gen Comput Syst. 2024;153:442–56. https://doi.org/10.1016/j.future.2023.12.013.
17. Lison P, Pilán I, Sánchez D, Batet M. Øvrelid, L. Anonymisation models for text data: state of the art, challenges and future directions. 2021. https://doi.org/10.18653/v1/2021.acl-long.323.
18. Commission E. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016.
19. ENISA C, Limniotis K, Hansen M, Jensen M, Eftasthopoulos P, Drogkaris P, Bourka A. Data pseudonymisation–advanced techniques and use cases–technical analysis of cybersecurity measures in data protection and privacy. 2021. https://doi.org/10.2824/860099.
20. ENISA Guasconi F, Angelidis P, Drogkaris P. Deploying pseudonymisation techniques–the case of health sector. european union agency for cybersecurity. Athens. 2022. https://doi.org/10.2824/092874.
21. Yermilov O, Raheja V, Chernodub A. Privacy- and utility-preserving NLP with anonymized data: a case study of pseudonymization. In: Proceedings of the Annual Meeting of the ACL. 2023. https://doi.org/10.18653/v1/2023.trustnlp-1.20.
22. Liu Z, et al. DeID-GPT: zero-shot medical text de-identification by GPT-4. 2023. https://arxiv.org/pdf/2303.11032.pdf.
23. Schultenkämper S, Bäumer F, Geierhos M, Lee YS. From unstructured data to digital twins: from tweets to structured knowledge. In: Proceedings of the Thirteenth International Conference on Social Media Technologies, Communication, and Informatics, SOTICS 2023, pp. 6–11. IARIA, Valencia. 2023.
24. Liu Z, Li Y, Shu P, Zhong A, Yang L, Ju C, Wu Z, Ma C, Luo J, Chen C, Kim S, Hu J, Dai H, Zhao L, Zhu D, Liu J, Liu W, Shen D, Liu T, Li Q, Li X. Radiology-Llama2: best-in-class large language model for radiology. 2023. https://doi.org/10.48550/arXiv.2309.06419.
25. Watanabe C, Amagasa T, Liu L. PrIvacy risks and countermeasures in publishing and mining social network data. In: 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom). 2011:55–66. https://doi.org/10.4108/icst.collaboratecom.2011.247177.
26. Majeed A, Lee S. Anonymization techniques for privacy preserving data publishing: a comprehensive survey. IEEE Access. 2021;9(9):8512–45. https://doi.org/10.1109/ACCESS.2020.3045700.
27. Zhang Y, Gan Z, Carin L. Generating text via adversarial training. In: NIPS Workshop on Adversarial Training, Academia. edu. 2016;21:21–32.
28. Yang L-C, Chou S-Y, Yang Y-H. MidiNet: a convolutional generative adversarial network for symbolic-domain music generation. 2017. https://doi.org/10.48550/arXiv.1703.10847.
29. Antipov G, Baccouche M, Dugelay J-L. Face aging with conditional generative adversarial networks. In: 2017 IEEE International Conference on Image Processing (ICIP), 2017;2089–2093. https://doi.org/10.1109/ICIP.2017.8296650.
30. Bao J, Chen D, Wen F, Li H, Hua G. Cvae-gan: fine-grained image generation through asymmetric training. In: 2017 IEEE International Conference on Computer Vision (ICCV). 2017:2764–2773. https://doi.org/10.1109/ICCV.2017.299.
31. Schultenkämper S, Bäumer FS. Pixels versus privacy: leveraging vision-language models for sensitive information extraction. Int J Adv Secur. 2024;17 (In Press).
32. Dineva K, Atanasova T. Osemn process for working over data acquired by iot devices mounted in beehives. Curr Trends Natl Sci. 2018;7(13):47–53.
33. Denisov S, Bäumer FS. The only link you'll ever need: how social media reference landing pages speed up profile matching. In: Lopata A, Gudonienė D, Butkienė R (eds) Information and software technologies. 2022:136–147. Springer, Cham. https://doi.org/10.1007/978-3-031-16302-9_10.
34. Schultenkämper S, Bäumer FS, Bellgrau B, Lee YS, Geierhos M. From digital tracks to digital twins: on the path to cross-platform profile linking. In: Sales TP, Kinderen S, Proper HA, Pufahl L, Karastoyanova D, Sinderen M (eds) Enterprise design, operations, and computing. EDOC 2023 Workshops, pp. 158–171. Springer, Cham. 2024.
35. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. 1990.
36. Karakasidis A, Pitoura E. Identifying bias in name matching tasks. In: International Conference on Extending Database Technology. 2019.
37. Li X, Guttmann A, Cipière S, Maigne L, Demongeot J, Boire J-Y, Ouchchane L. Implementation of an extended fellegi-sunter probabilistic record linkage method using the jaro-winkler string comparator. In: IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). 2014;375–379. https://doi.org/10.1109/BHI.2014.6864381.
38. Treeratpituk P, Giles CL. Name-ethnicity classification and ethnicity-sensitive name matching. Proc AAAI Conf Artif Intell. 2021;26(1):1141–7. https://doi.org/10.1609/aaai.v26i1.8324.

39. Kammakomati M, Battula SV. MergeURL: an effective url merging and shortening service. 2020;9:63–69.
40. Hill LS. Cryptography in an algebraic alphabet. Am Math Mon. 1929;36(6):306–12. https://doi.org/10.1080/00029890.1929.11986963.
41. Desai N, Das ML, Chaudhari P, Kumar N. Background knowledge attacks in privacy-preserving data publishing models. Comput Secur. 2022;122.
42. Dosovitskiy A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations. 2021.
43. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, 2021;39:8748–8763. PMLR, Virtual Event.
44. Jin W, Cheng Y, Shen Y, Chen W, Ren X. A Good prompt is worth millions of parameters? Low-resource prompt-based learning for vision-language models. In: Proceedings of the 60th Annual Meeting of the ACL (Volume 1: Long Papers), 2022;2763–2775. ACL, Dublin, Ireland.
45. Zheng L, Chiang W-L, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing EP, Zhang H, Gonzalez JE, Stoica I. Judging LLM-as-a-judge with MT-bench and chatbot arena. 2023. https://doi.org/10.48550/arXiv.2306.05685.
46. Touvron H, et al. Llama 2: open foundation and fine-tuned chat models. 2023.
47. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019;4396–4405.
48. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016;2414–2423. https://doi.org/10.1109/CVPR.2016.265.
49. Schultenkämper S, Bäumer FS. Looking for a needle in a haystack: how can vision-language understanding help to identify privacy-threatening images on the web. In: The Eighteenth International Conference on Internet and Web Applications and Services (ICIW 2023), 2023;1–6. IARIA, Venice.
50. Geifman N, Rubin E. Towards an age-phenome knowledgebase. BMC Bioinform. 2011;12(1):229. https://doi.org/10.1186/1471-2105-12-229.
51. Jablonski NG. The evolution of human skin and skin color. Annu Rev Anthropol. 2004;33(1):585–623. https://doi.org/10.1146/annurev.anthro.33.070203.143955.
52. Frost P. European hair and eye color: a case of frequency-dependent sexual selection? Evolut Hum Behav. 2006;27(2):85–103. https://doi.org/10.1016/j.evolhumbehav.2005.07.002.
53. Li J, Li D, Xiong C, Hoi S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. 2022;12888–12900. PMLR.
54. Li J, Li D, Savarese S, Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. ICML'23. J Mach Learn Res Honolulu Hawaii USA.
55. Dai W, Li J, Li D, Tiong AMH, Zhao J, Wang W, Li B, Fung PN, Hoi S. InstructBLIP: towards general-purpose vision-language models with instruction tuning. Adv Neural Inform Process Syst. 2024;36.
56. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers), 2019;4171–4186. ACL, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423.
57. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y, Wang X, Dehghani M, Brahma S, Webson A, Gu SS, Dai Z, Suzgun M, Chen X, Chowdhery A, Castro-Ros A, Pellat M, Robinson K, Valter D, Narang S, Mishra G, Yu A, Zhao V, Huang Y, Dai A, Yu H, Petrov S, Chi EH, Dean J, Devlin J, Roberts A, Zhou D, Le QV, Wei J. Scaling Instruction-Finetuned Language Models. 2022. https://doi.org/10.48550/arXiv.2210.11416.
58. ThoughtfulDev. EagleEye: stalk your friends. Find their Instagram, FB, and Twitter Profiles using Image Recognition and Reverse Image Search. https://github.com/ThoughtfulDev/EagleEye. Accessed 2024-05-28.
59. MIT Technology Review: A GPT-3 bot posted comments on reddit for a week and no one noticed. https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/. Accessed: 2024-05-28.
60. Douglas DM. Doxing: a conceptual analysis. Ethics Inform Technol. 2016;18(3):199–210. https://doi.org/10.1007/s10676-016-9406-0.