From Unstructured Data to Digital Twins: From Tweets to Structured Knowledge

Sergej Schultenkämper[#1], Frederik S. Bäumer[#2], Yeong Su Lee[*3], Michaela Geierhos[*4]
[#]Bielefeld University of Applied Sciences and Arts
Interaktion 1, 33619 Bielefeld, Germany
[1]sergej.schultenkaemper@hsbi.de
[2]frederik.baeumer@hsbi.de
[*]University of the Bundeswehr Munich
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
[3]yeongsu.lee@unibw.de
[4]michaela.geierhos@unibw.de

***Abstract*—This paper focuses on extracting relevant information from unstructured data, specifically analyzing text shared by users on Twitter. The goal is to build a comprehensive knowledge graph by extracting implicit personal information from tweets, including interests, activities, events, family, health, relationships, and professional information. The extracted information is used to instantiate a Digital Twin and develop a personalized alert system to protect users from threats, such as social engineering or doxing. The paper evaluates the effectiveness of state-of-the-art large language models, such as GPT-4, for extracting relevant triples from tweets. The study also explores the notion of Digital Twins in the context of cyber threats and presents related work in information extraction. The approach includes data collection, multi-label classification, relational triple extraction, and evaluation of the results. The dataset used is from Twitter, and the study analyzes the challenges posed by user-generated data. The results show the accuracy of the extracted triples and the personal characteristics that can be identified from tweets for the development of the Digital Twin. The results contribute to the ADRIAN research project, which focuses on machine learning-based methods for detecting potential threats to people's privacy.**

***Index Terms*—*Digital Twin*; *Data Privacy*; *Semantic Triple*.**

I. Introduction

Through almost every activity on the Web, users leave footprints, both active and passive [1]. This includes obvious information, such as images, text, and video that users knowingly upload, as well as information that is transmitted without user intervention, such as the device's IP address or user agent. Moreover, information hidden in text and images that are unknowingly posted is difficult for users to keep track of. This has already been demonstrated in an impressive and media-effective way, e.g., by the automatic identification of vacation announcements and the extraction of hidden Global Positioning System (GPS) image data from Twitter (aka "X"), which could be used, for example, to scout vacant properties for burglaries [2] or to reveal the running routes of soldiers on secret army bases, whose publication on sports portals revealed the exact location of the military installations [3]. Even small amounts of information, when combined with other information, can be a threat, as research has shown [4].

In this paper, we focus on texts shared by users on Twitter to extract relevant information from unstructured, noisy data. According to [5], over 500 million tweets are shared every day for various reasons, such as expressing personal views, sharing news, or discussing current events. Data analysis can help extract both implicit and explicit personal information [6]. For this reason, this study focuses on analyzing tweets regarding *personal interests*, *activities*, *events*, *family*, *health*, *relationships*, and *professional information*. The goal is to construct a robust knowledge graph by retrieving information from unstructured sources. The instantiation of the Digital Twin (DT) is based on this graph. In this context, we plan to develop a personalized early warning system that will be centered around the DT. In case of excessive disclosure of sensitive information, this system can notify users in Online Social Networks (OSNs). Providing early warnings of potential threats, such as social engineering or doxing, can mitigate risks to users. To address the challenge of information extraction from tweets, we investigate the suitability of current state-of-the-art Large Language Models (LLMs), such as Generative Pretrained Transformer 4 (GPT-4), for extracting significant triples in the form of subject, predicate, and object. The purpose of this study is to evaluate the effectiveness of LLMs for information extraction. The objective is to identify significant triples from unstructured and minimally informative Twitter data of varying lengths [7].

All of these considerations are part of the research project ADRIAN, which stands for "Authority-Dependent Risk Identification and Analysis in online Networks". For this purpose, we discuss related work in Section II and describe the dataset and our approach in Section III. Subsequently, we present our modeling strategies and results in Section IV. Finally, we discuss our findings (Section V) and draw conclusions in Section VI.

II. Related Work

In this section, we discuss the notion of DTs in the context of cyber threats and the related work on information extraction.

A. Digital Twins in the Context of Cyber Threats

The term DT is ambiguous and finds application in various research and practical domains, including medicine and computer science [3] [8] [9]. The advancement of artificial intelligence has broadened the scope of this term, and in a broader sense, DTs can be defined as computer-based models or (physical and/or virtual) machines that simulate, emulate, mirror, or act as a "twin" of a real-world entity, which could be an object, a process or a human [8]. In our context, we identify three distinct levels of integration for DTs, according to [8]: (a) *Digital Model*, (b) *Digital Shadow*, and (c) *Digital Twin*. A digital model is a basic virtual representation of a physical object or system, without any automatic exchange of information between the virtual and physical worlds. Any changes to the physical object must be manually updated in the digital model. In the future, a digital shadow builds on this concept by enabling a unidirectional, automatic flow of information from the physical to the virtual world. Sensors capture information from the physical entity and transmit signals to the virtual model. Finally, a complete DT exists when bidirectional communication is established between the virtual and physical environments, facilitating the automatic exchange of information. This allows the DT to accurately reflect the real-time state and evolution of its physical counterpart. However, when considering socio-technical systems, the dynamics change as these systems include both human and machine components.

Therefore, it becomes relevant to explore the notion of a *Human DT* [10]. Despite its growing importance, a standardized definition or understanding of this concept has not yet been achieved [9] [11] . The digital information available about individuals is often referred to as the *Digital Footprint* or *Digital Representation*, with these terms often used interchangeably. These terms refer to the data left behind by users on the Internet, often unknowingly, without identifying or linking to a specific individual. The concepts of Digital Footprint, Digital Shadow, and Digital Twin can be distinguished on the basis of several aspects: Identifiability, active or passive data collection, individualized or aggregated evaluation, real-time or delayed analysis, decision-making authority, and comprehensive representation [10]. The Human DT aims to store and analyze relevant characteristics of an individual in a given situation. This may include demographic or physiological data, skill or activity profiles, or health status [9] [10].

In the ADRIAN project, we define a DT as a digital representation of a real person, instantiated using publicly available information from the web [3] [12] . It is important to note that a DT can never fully capture the complexity of a real person, but rather reproduces specific characteristics that, either alone or in combination with other characteristics, may pose a threat to the individual [4].

B. Relational Triple Extraction

Relation Extraction (RE) is the identification of pairs of entities and their relations, expressed as (HEAD, RELATION, TAIL), from unstructured text. Traditional approaches to RE are divided into two separate tasks: Named Entity Recognition (NER) and Relationship Classification (RC) [13] [14]. However, this approach is prone to the problem of error propagation. Therefore, recent studies have aimed to overcome this problem by exploring common models for extracting relational triples in an end-to-end manner [15]. For example, one approach is the text generation technique, which treats a triple as a series of tokens, and uses the encoder-decoder architecture to generate triple components, similar to machine translation [16] [17].

Recent studies on text generation have highlighted In-Context Learning (ICL) as an important feature of GPT-3.5, GPT-4, and other transformer-based models [18] [19]. Unlike traditional machine learning models that require explicit and task-specific training datasets, models with ICL capabilities can learn and adapt to new tasks by using the context provided during inference. For example, GPT-3.5 or GPT-4 achieve ICL by providing a set of contextual examples at the beginning of a prompt. These examples help guide the model's responses. As a result of this capability, these models can perform a wide range of tasks without the need for task-specific fine-tuning, resulting in highly flexible and adaptable models.

Xu et al. (2023) [19] explore the use of LLMs for few-shot RE. The paper focuses on the approach of ICL, which involves designing prompts of varying complexity to help LLMs understand the task of RE. Two types of prompts are used: (1) The text prompt contains only the essential elements for requirements engineering. (2) The instruction prompt includes a task-related instruction that describes the requirements engineering task as well as the essential elements. The results indicate that instructions and schemes in ICL are crucial for RE with LLMs. In general, the instruct prompt model outperformed the text prompt model. According to the study, the inclusion of task-related information, such as instructions or schemes, is essential for effective ICL with LLMs.

Wan et al. (2023) [18] present GPT-RE, an innovative method for Relational Extraction (RE) that utilizes the ICL abilities of GPT-3. The method consists of two primary components. The first component is entity-aware demonstration retrieval, which reconstructs the context by incorporating information about the entity pair that is critical for RE. The second component is inferential demonstration, which enhances the demonstrations with inferences derived from the ground truth relationship labels. This facilitates GPT-3 to gain better understanding of the demonstrations and enhance its performance. The study's results indicate that GPT-RE surpasses fine-tuning on three datasets, implying that GPT-3 possesses the potential to perform outstandingly when the retriever has prior task knowledge. It is observed that the quality of demonstrations holds greater significance than their quantity. Furthermore, demonstrations enriched with reasoning present consistent improvement across all k-shot settings, implying that GPT-3's reasoning ability could be successfully unlocked by employing reasoning based on ground truth relational labels, thus enhancing ICL. The proposed GPT-RE method attains the highest scores in the SemEval 2010 and SciERC datasets, exhibiting its efficacy in RE.

III. Approach and Dataset

This section presents our approach (Fig. 1), which includes (1) data collection, (2) multi-label classification, (3) relational triple extraction, and (4) evaluation of the extraction. Selecting an appropriate data source is the first step, and we chose Twitter as our data source. Twitter contains a significant amount of structured and unstructured data, making it a suitable candidate for instantiating DTs and performing threat analysis as part of the ADRIAN project. Twitter's Application Programming Interface (API) provides extensive and easily accessible data, adding to its attractiveness as a data source. Tweets, constrained by character limits, present a significant challenge. Furthermore, hashtags, user mentions, URLs, and emoticons in tweets generate substantial noise, making it difficult to extract reliable information for meaningful insights [7] [20]. In our data collection, we randomly selected 300 users from our database to determine the frequency of tweets. Of these selected users, 246 had posted tweets, with an average frequency of 3,532 tweets (cf. Table I).

TABLE I
Descriptive Statistics of the Twitter Dataset

Dataset Feature	Count
No. of Users	246
Avg. Tweets/User	3,532
Median Tweets/User	546
Min. Tweets/User	1
Max. Tweets/User	80,689
Total Tweets	869,069
Top Languages	EN, DE, FR, ES, TR
No. of Reply Tweets	274,504
No. with Attachments	106,997
No. with Geolocation	43,138
No. of Retweets	236,553

It should be noted that a user can have a very large number of tweets, with our approach it is not possible to process such a large number of tweets because the extraction with GPT-4 consumes immense cost, and therefore in this work we limit ourselves to a number of 5,000 tweets for the extraction of the triples. Therefore, we use pre-filtering to identify relevant tweets at the initial stage. For this purpose, we chose the OffMyChest dataset [21], since it is one of the few available datasets with personal information. The dataset classification varies from 'little personal information', which includes basic author information, such as *age*, *occupation*, or *location*, to information about *family*, *interests*, or *hobbies*. Conversely, 'much personal information' consists of sensitive data that has the potential to expose individuals, such as their *health conditions*, *physical attributes*, *behavior*, or *personal experiences*, as explicitly stated in the record description. However, the dataset description does not distinguish between these two categories of information. So, based on our requirements, we annotated the information with the following labels: *Event*, *Family*, *Health*, *Interests*, *Personal Information*, *Relationships*, and *Work/School*, as shown in Table III. We use the annotated dataset to train the Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach (XLM-RoBERTa) [22] for multi-label classification. After pre-filtering with this model, 50 tweets from 100 users are randomly selected for relational triple extraction. To structure the triples, we use properties from Schema.org [23], as shown in Table II.

TABLE II
Defined categories and associated properties

Category	Properties
Event	s:attendee
Family	s:children, s:parent, s:sibling, s:spouse
Health	s:diagnosis, s:drug, s:healthCondition
Interests	s:interests
Personal Information	s:birthDate, s:birthPlace, s:email, s:gender, s:location, s:nationality
Relationships	s:colleague, s:knows
Work/School	s:alumniOf, s:jobTitle, s:workLocation, s:worksFor

All of the entities used are associated with Schema.org's Person class, and their structure enhances the clarity and interpretability of the data. Furthermore, the targeted triples can be directly integrated into our current applications. Thus, the DT can be instantiated directly. The generated dataset can be used to fine-tune available open-source models, such as LLaMa-2 [24] for triple extraction.

IV. Modeling and Results

The modeling phase consists of two main tasks: (1) multi-label classification and (2) few-shot prompt relational triple extraction. The first task involves annotating data from the OffMyChest dataset and then training an XLM-RoBERTa model based on the annotated data. The second task focuses on relational triple extraction, with core work including prompt engineering with examples for the few-shot approach and defining an output scheme.

A. Multi-Label Classification

We use XLM-RoBERTa, a multilingual language model, to train the multi-label classification model. This model uses the Masked Language Model (MLM) technique, like Bidirectional Transformers for Language Understanding (BERT) [25], but was trained on monolingual content from 100 different languages [22]. Our approach can be used in multiple languages due to the multilingual capabilities of XLM-RoBERTa and GPT-4. Our model is trained and evaluated on 1,803 annotated sentences from the OffMyChest dataset. The dataset was partitioned for training (80 %, 1,442 sentences) and evaluation (20 %, 361 sentences). The Weights & Biases library [26] is used to facilitate hyperparameter optimization and to monitor the progress of the experiment. Optimal performance was achieved on XLM-RoBERTa using a learning rate of 5e-5, over five training epochs, and with a training batch size of 16. The evaluation results are shown in Table III.

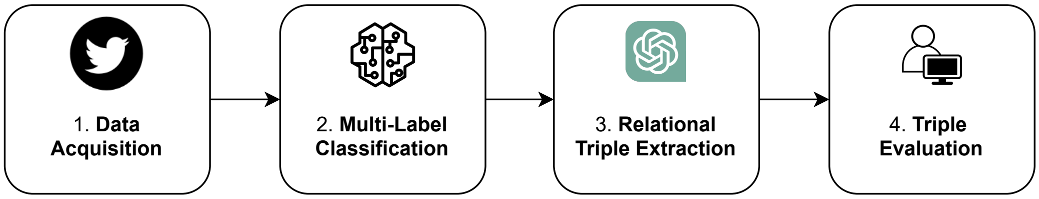
1. Data Acquisition
2. Multi-Label Classification
3. Relational Triple Extraction
4. Triple Evaluation

Fig. 1. Proposed approach for relational triple extraction

TABLE III
XLM-RoBERTa Classification Results

Class	Precision	Recall	F_1-score	Support
Event	0.7736	0.8542	0.8119	48
Family	0.8505	0.8349	0.8426	109
Health	0.7969	0.7846	0.7907	65
Interests	0.8333	0.6604	0.7368	53
Personal Information	0.7093	0.7871	0.7462	155
Relationships	0.7795	0.9000	0.8354	110
Work/School	0.7963	0.8600	0.8269	50

The results show the performance of the XLM-RoBERTa model across different classes. The highest precision score is observed in the "*Family*" class, while "*Relationships*" yields the highest recall. Regarding the F_1-score, the "*Family*" class presents the highest value. On the contrary, the "*Interests*" class has the lowest recall and F_1-score, while the "Personal Information" class has the lowest precision.

B. Few-Shot-Prompt Relational Triple Extraction

The prompt is designed to use the standard message and function call feature to create a scheme with appropriate descriptions. The prompt specifies that subject-predicate-object triples must be extracted from the provided tweets. Each tweet contains the author's name and the message content. It is important to note that the subject does not have to be the author of the tweet; other subjects should be identified and extracted. As recent studies have shown positive results with ICL [18] [19], we will provide examples in the form of triples.

Designing an output scheme involves using the function call feature of GPT-4. The function takes a scheme as input and generates a corresponding JavaScript Object Notation (JSON) object as its output. This JSON object can then be used to interact with external APIs or databases. The scheme is designed to extract subject-predicate-object triples from the input data. The subject refers to the entity that performs an action or the entity about which a statement is made. The predicate describes either the action performed by the subject or the state of the subject; a predefined set of values defines this attribute (cf. Table II). The object represents either the entity that is the subject of the action performed by the subject or the entity that is in some way involved in the action expressed by the predicate. Table IV illustrates an instance of the expected output of the function, where each row corresponds to a triple.

TABLE IV
Example output for the defined scheme

Subject	Predicate	Object
John	s:worksFor	Microsoft
Mary	s:location	New York

The next step is to validate the results of the GPT-4 model. For this task, we use the author, the tweets, and the extracted triples. To validate the results, we use LabelStudio [27] to annotate the correctness of the extracted subject, predicate, and object. A total of 1,288 triples were extracted from 5,000 tweets using GPT-4. The evaluation score is calculated based on the accuracy of each extracted property (cf. Table V).

TABLE V
Results for the GPT-4 Triple Extraction

Predicate Class	Subject	Predicate	Object	Support
s:alumniOf	1.0000	0.6897	0.8966	30
s:attendee	0.9760	0.9162	0.9401	167
s:birthDate	1.0000	0.8235	0.8824	17
s:birthPlace	1.0000	0.5000	1.0000	2
s:children	1.0000	1.0000	1.0000	9
s:colleague	1.0000	0.9605	0.9605	76
s:diagnosis	1.0000	0.6923	0.9231	14
s:email	1.0000	0.6667	1.0000	3
s:healthCondition	0.9091	0.8636	0.9545	22
s:interests	0.9942	0.9796	0.9650	352
s:jobTitle	0.9800	0.7800	0.8400	101
s:knows	1.0000	0.9643	1.0000	29
s:location	0.9908	0.9495	0.9312	219
s:nationality	1.0000	1.0000	0.8333	6
s:parent	0.9516	0.9032	0.9355	63
s:sibling	1.0000	0.8696	0.9565	23
s:spouse	1.0000	0.9000	0.9000	10
s:workLocation	1.0000	0.8889	0.8889	18
s:worksFor	0.9840	0.8560	0.8960	127
Micro Avg.	0.9866	0.9142	0.9331	Σ 1,288
Macro Avg.	0.9887	0.8528	0.9318	Σ 1,288
Weighted Avg.	0.9867	0.9142	0.9332	Σ 1,288

The high accuracy for the subject should be taken with caution, as it mostly concerns the tweeter. The model performs well in extracting predicates and objects from the given tweets. The model achieves average accuracies of 0.9142 (micro), 0.8528 (macro), and 0.9142 (weighted) for predicates, and 0.9331 (micro), 0.9318 (macro), and 0.9332 (weighted) for objects. Predicates with low counts, such as "s:birthPlace", "s:children", and "s:nationality", are underrepresented and require more data for a reliable measure of model performance.

V. DISCUSSION

The multi-label classification gave satisfactory results for all categories except "*Interests*". It is reasonable to hypothesize that the ambiguity associated with this particular class may be due to the broad and diverse nature of interests, covering a variety of areas, such as hobbies, occupations, artistic pursuits, sports, and other domains. The wide range may make it difficult to achieve accurate and consistent classifications. Intent classification could serve as an additional approach to pre-selecting tweets. Understanding the intent behind a tweet is as important as identifying its topic. Determining whether the user is self-disclosing, sharing information, or commenting on others is critical. Such insights can lead to a nuanced interpretation of the tweet's content and context, helping to develop more accurate classification strategies. Our method is effective in quickly extracting triples that can be transformed into a graphical representation (Fig. 2).

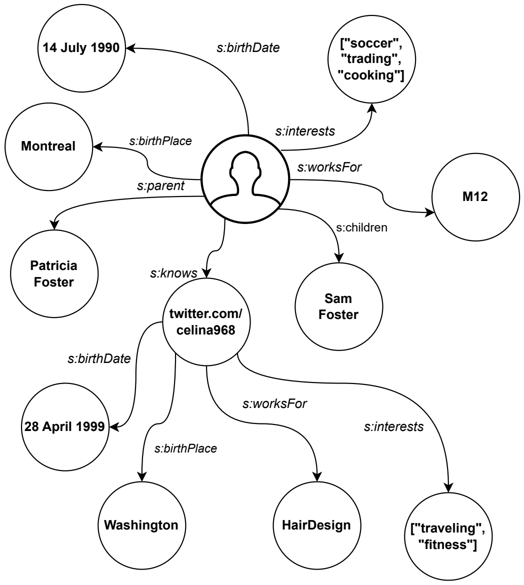

Fig. 2. An example of a knowledge graph constructed from extracted triples

The knowledge graph visually represents the relationships between extracted entities, providing a structured format that facilitates interpretation and analysis of the data. The nodes represent entities, while the edges represent the relationships between them, allowing the interconnectedness and hierarchical relationships of the extracted information to be explored.

However, the approach is not flawless due to the significant variance in the data, which requires normalization. Although the subject and predicate require minimal normalization, the object requires a different method due to the high variance during extraction. A common problem is the extraction of multiple individuals, usually seen with the target predicates "s:colleague" and "s:knows". Moreover, the existence of Twitter handles is a common problem that needs to be solved. Locations, represented by flags or abbreviations, are also extracted and need to be processed. Job titles sometimes include the company name, as in the case of "CTO of OpenAI". Although normalization in this particular case is straightforward, finding a solution for the following examples is much more challenging. It is necessary to analyze how to handle cases where there is no precise data, but statements like "*Father of 3 children*" or "*Photo of my dad and me ...*" that do not contain information about a specific person, but still have relevant information when combining the extracted information. This data should be considered as it may help to find additional family-related information in subsequent tweets or to enrich the dataset with information from the photo. For example, some information was labeled as "s:interest", but to better reflect its purpose, it could be categorized using additional predicates, such as "s:searchAction" or "s:skills". This would be useful if a user was searching for an employee or writing a tweet about specific skills. Furthermore, the "s:email" category seems too narrow in scope, and replacing it with "s:contactPoint" would allow for the inclusion of contact information from OSNs, such as LinkedIn links.

Closed LLMs, such as GPT-4 should be viewed critically in research, as they are not open, require payment, and lack transparency. In addition, they are not always robust, and their APIs are often unreliable and inaccessible. Therefore, open-source LLMs should be considered as an alternative. Open-source models offer transparency, flexibility, and community collaboration, allowing for review and improvement by a wider range of researchers. They are often more robust and adaptable, making them a preferable choice for research efforts. Reducing reliance on closed and expensive LLMs and encouraging the use of open-source alternatives is crucial to promoting openness, transparency, and progress in language modeling research. Of course, this requires that the necessary computing power (e.g., graphics cards) be made available. To be fair, this also comes at a cost, but it reduces dependency and increases transparency.

VI. CONCLUSION

Finally, this paper addressed the challenge of extracting relevant information from unstructured data, specifically tweets. The study demonstrated the potential of LLMs, such as GPT-4, to extract triples and build a comprehensive knowledge graph. By instantiating a DT and developing a personalized early warning system, the study aimed to protect users from potential threats arising from the disclosure of sensitive information. The results of the study demonstrate the effectiveness of the XLM-RoBERTa model in multi-label classification and provide insights into the personal characteristics expressed in tweets. In addition, the few-shot prompt relational triple extraction approach demonstrates the potential of GPT-4 to extract structured information from unstructured data. The designed prompts and output scheme enable the identification and representation of subject-predicate-object triples, contributing to the construction of a knowledge graph.

These modeling and extraction techniques lay the foundation for further advances in the field. With the availability of powerful tools and frameworks, there is an opportunity to train custom LLMs tailored to specific domains or datasets. This opens up new possibilities for improved accuracy and domain-specific insights in information extraction tasks. Looking ahead, training custom LLMs has several advantages. First, it allows for better adaptation to specific domains and datasets, leading to improved extraction accuracy and relevance. Custom LLMs can be fine-tuned for specialized datasets, including those containing personal information, thereby improving the performance and applicability of the extraction process. Moreover, training custom LLMs enables greater control over privacy and data security. By using in-house models, organizations can ensure that sensitive information remains within their infrastructure, reducing the risk of data breaches or unauthorized access. This approach addresses the growing concern about privacy and the need to protect personal data in today's digital landscape. In addition, custom LLMs provide the opportunity for continuous learning and refinement. By training models on evolving datasets and incorporating feedback from user interactions, the accuracy and performance of the extraction process can be continually improved. This adaptability is essential to keep pace with emerging trends, language variations, and evolving threats in online platforms.

In summary, training custom LLMs brings several benefits, including improved extraction accuracy, enhanced privacy and data security, and the potential for continuous learning and refinement. These benefits open up new avenues for research and development in the field of information extraction, providing valuable insights and actionable intelligence to mitigate threats and protect users in online environments.

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] C. Iordanou, G. Smaragdakis, I. Poese, and N. Laoutaris, "Tracing Cross Border Web Tracking," in *Proceedings of the Internet Measurement Conference 2018*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 329–342.

[2] M. B. Flinn, C. J. Teodorski, and K. L. Paullet, "Raising Awareness: An Examination of Embedded GPS Data in Images Posted to the Social Networking Site Twitter," *Issues in Information Systems*, vol. 11, no. 1, pp. 432–438, 2010.

[3] 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., 2021.

[4] F. S. Bäumer, N. Grote, J. Kersting, and M. Geierhos, "Privacy Matters: Detecting Nocuous Patient Data Exposure in Online Physician Reviews," in *International Conference on Information and Software Technologies*. Springer, 2017, pp. 77–89.

[5] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *International Journal of Information Technology*, vol. 15, no. 2, pp. 965–980, January 2020.

[6] M. Mazza, G. Cola, and M. Tesconi, "Ready-to-(ab)use: From fake account trafficking to coordinated inauthentic behavior on Twitter," *Online Social Networks and Media*, vol. 31, p. 100224, 2022.

[7] W. Ahmed, P. A. Bath, and G. Demartini, "Using Twitter as a data source: An overview of ethical, legal, and methodological challenges," *The ethics of online research*, vol. 2, pp. 79–107, 2017.

[8] 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.

[9] S. Schultenkämper and F. Bäumer, "Privacy Risks in German Patient Forums: A NER-based Approach to Enrich Digital Twins," in *Information and Software Technologies*. Cham: Springer International Publishing, 2023, In press.

[10] 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, August 2020.

[11] 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, 2019.

[12] F. S. Bäumer, J. Kersting, M. Orlikowski, and M. Geierhos, "Towards a Multi-Stage Approach to Detect Privacy Breaches in Physician Reviews." in *SEMANTICS Posters&Demos*, 2018.

[13] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction," *Journal of Machine Learning Research*, vol. 3, pp. 1083–1106, 2003.

[14] Y. S. Chan and D. Roth, "Exploiting syntactico-semantic structures for relation extraction," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 551–560.

[15] Y.-M. Shang, H. Huang, X. Sun, W. Wei, and X.-L. Mao, "Relational Triple Extraction: One Step is Enough," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, July 2022, pp. 4360–4366.

[16] X. Zeng, D. Zeng, S. He, K. Liu, and J. Zhao, "Extracting relational facts by an end-to-end neural model with copy mechanism," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 506–514.

[17] D. Zeng, H. Zhang, and Q. Liu, "Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 9507–9514.

[18] Z. Wan *et al.*, "GPT-RE: In-context Learning for Relation Extraction using Large Language Models," 2023.

[19] X. Xu, Y. Zhu, X. Wang, and N. Zhang, "How to Unleash the Power of Large Language Models for Few-shot Relation Extraction?" 2023.

[20] J. T. B. Jafar, "Information extraction from user generated noisy texts," Ph.D. dissertation, The Ohio State University, 2020.

[21] K. Jaidka, I. Singh, J. Lu, N. Chhaya, and L. Ungar, "A report of the CL-Aff OffMyChest Shared Task: Modeling Supportiveness and Disclosure," in *Proceedings of the AAAI-20 Workshop on Affective Content Analysis*. New York, USA: AAAI, 2020, pp. 118–129.

[22] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451.

[23] Schema.org, "Schema.org - Structured Data for the Web," 2023, Available: https://schema.org, retrieved 2023/10/02.

[24] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: ACL, June 2019, pp. 4171–4186.

[26] L. Biewald, "Experiment Tracking with Weights and Biases," 2020, Available: https://www.wandb.com/, retrieved 2023/10/02.

[27] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020-2023, Available: https://github.com/heartexlabs/label-studio, retrieved: 2023/10/02.