Hochschule
Bielefeld

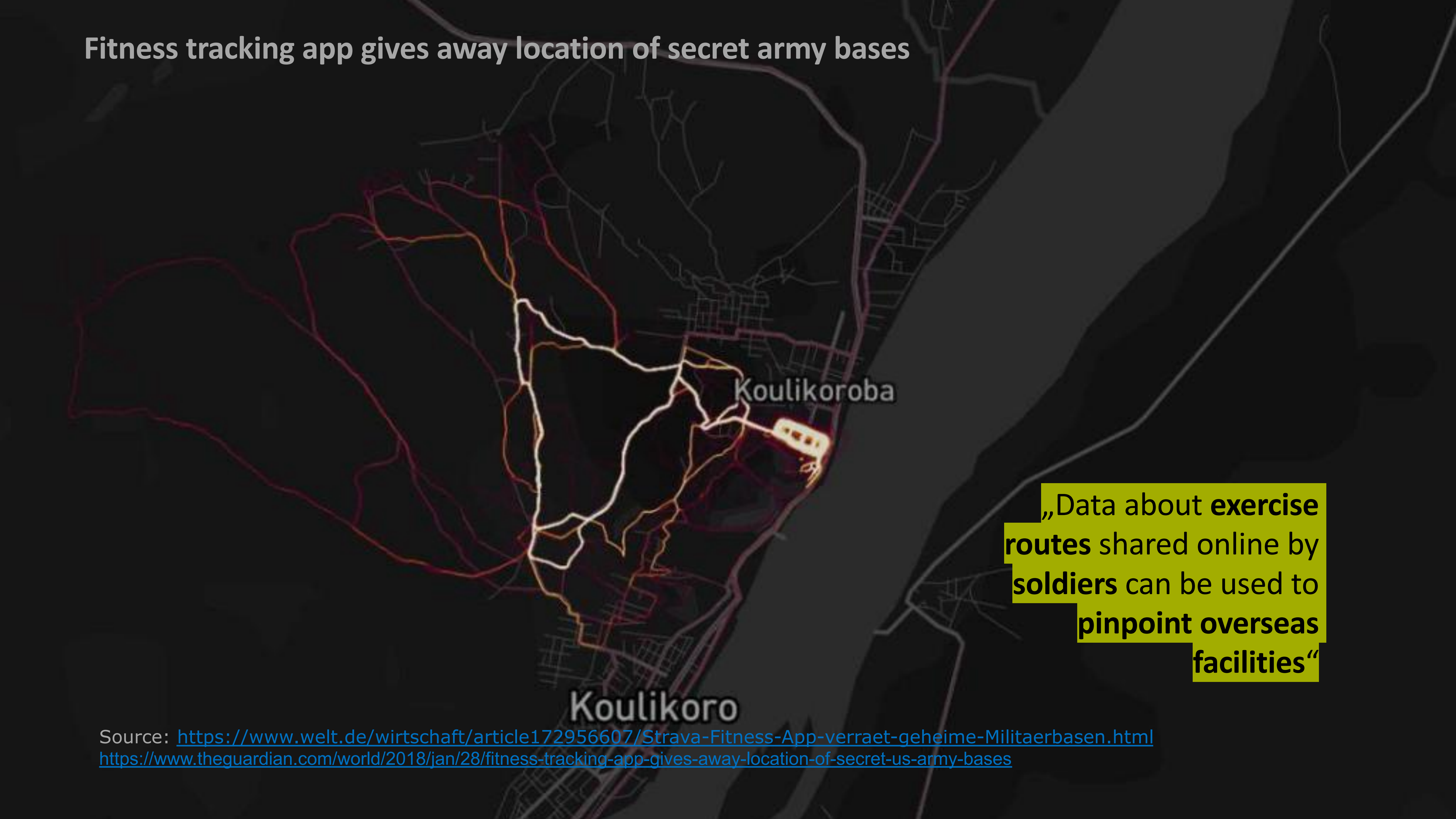
University of
Applied Sciences
and Arts

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

Sergej Schultenkämper, M.Sc.; Bielefeld University of Applied Sciences and Arts
Dr. Frederik Simon Bäumer; Bielefeld University of Applied Sciences and Arts
Dr. Yeong Su Lee; University of the Bundeswehr Munich
Prof. Dr. Michaela Geierhos; University of the Bundeswehr Munich

Funded by
the European Union
NextGenerationEU

Fitness tracking app gives away location of secret army bases

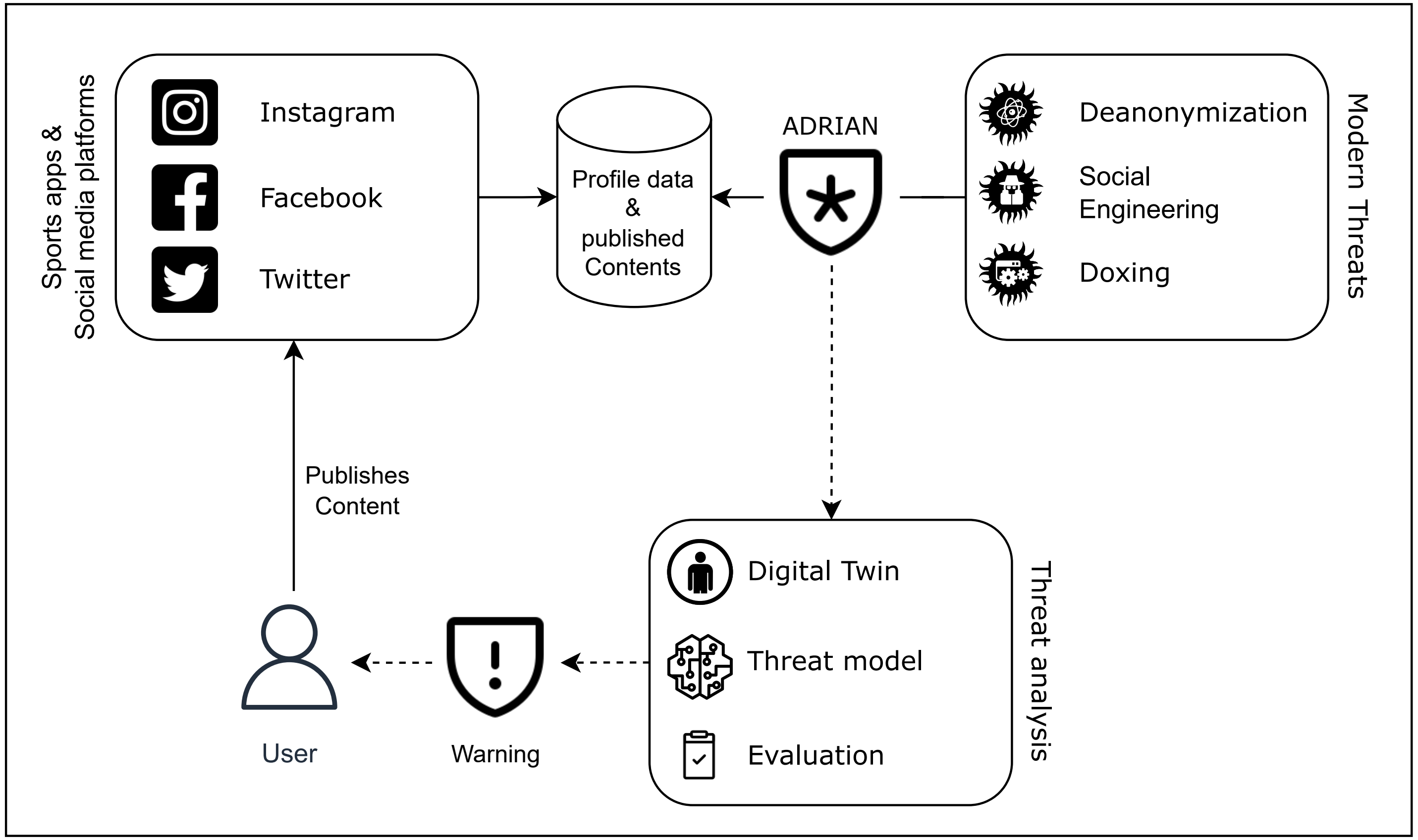


„Data about **exercise routes** shared online by **soldiers** can be used to **pinpoint overseas facilities**"

Source: https://www.welt.de/wirtschaft/article172956607/Strava-Fitness-App-verraet-geheime-Militaerbasen.html
https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

Hochschule
Bielefeld
University of
Applied Sciences
and Arts

ADRIAN RESEARCH PROJECT

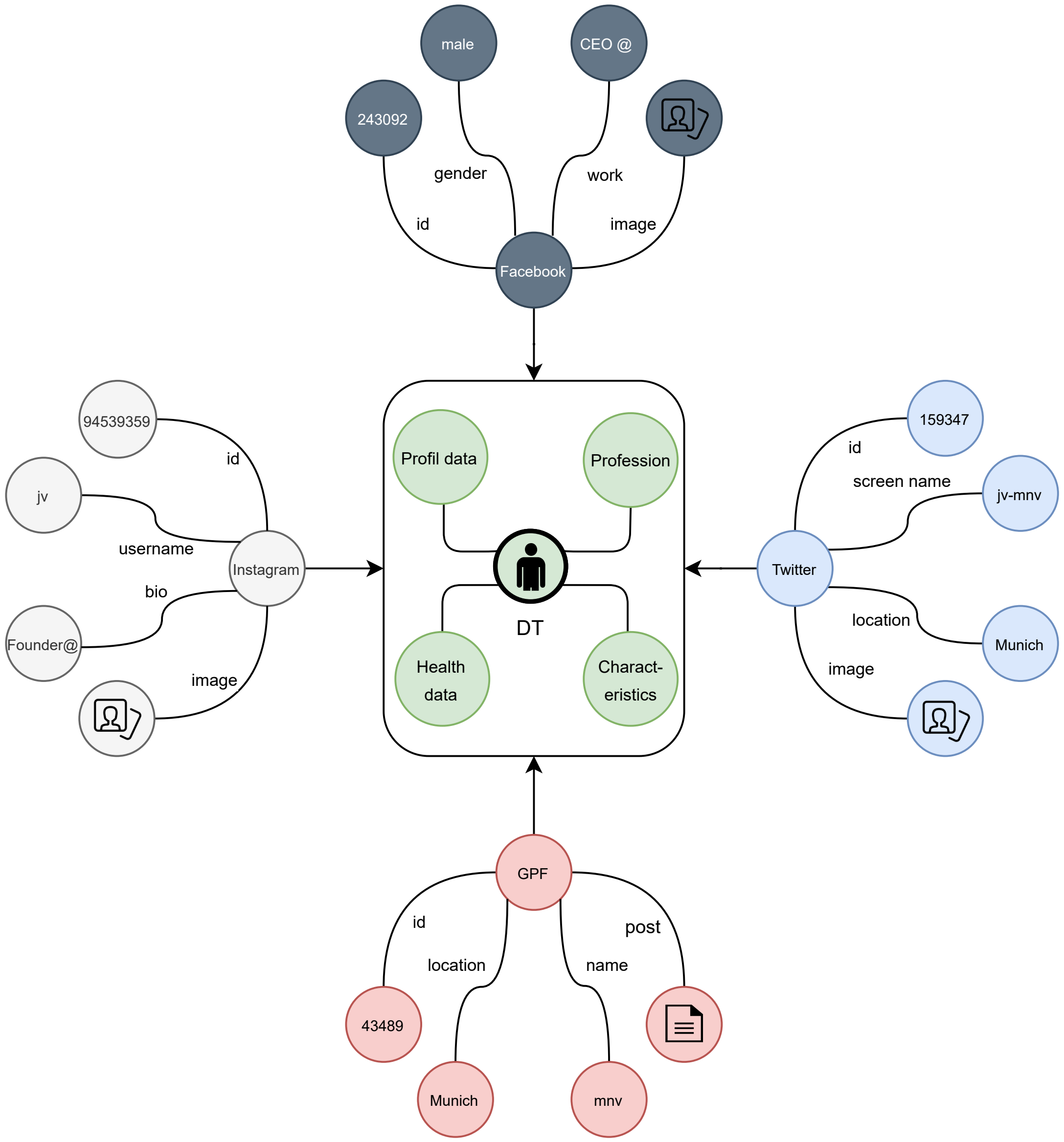


RELATED WORK

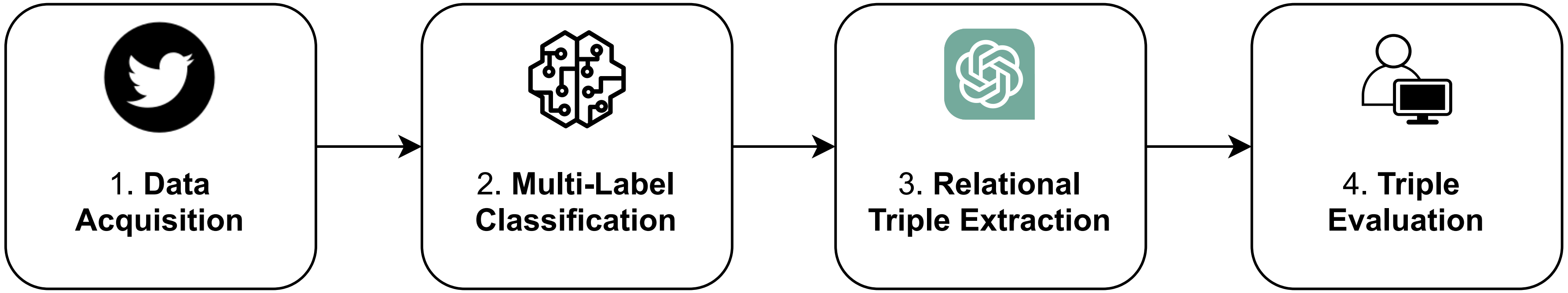
- **Traditional Methods**: Rule-Based Systems, and Language Model Techniques (Named Entity Recognition & Relationship Classification)
- **Latest Research**: Large Language Models (LLMs) combined with In-Context Learning (ICL) capabilities offer promising ways to extract relational data as a single task
- **Recent Studies**: Xu et al. (2023) emphasized instructional prompts in ICL. Wan et al. (2023) introduced GPT-RE, emphasizing quality demonstrations and improved reasoning

HUMAN DIGITAL TWIN (HDT)

- There is **no standard definition** for HDT
- In the ADRIAN project, the HDT is defined as a **digital representation** of a **real person**, **instantiated** through available **web-based information**
- The HDT is used to **store** and **analyze** relevant **characteristics** of an individual
- The **vulnerability** of a person can be modeled and measured



APPROACH



- **Data Acquisition**: 870,000 Tweets from 246 Users
- **Multi-Label Classification:** XLM-RoBERTa Training for 7 classes of interest
- **Relational Triple Extraction:** GPT-4 to analyze 5,000 tweets from 100 users
- **Triple Evaluation:** Manual Triple Evaluation

DATASET

- **Data Source:** Twitter API
- **Collection Date**: May 2023
- **Data Challenges:** Tweets have character restrictions and often contain very limited information
- **User-generated content**: Suffers from spelling and grammar errors, lacks context

Dataset Feature	Count
No. of Users Searched	300
No. of Users Found	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

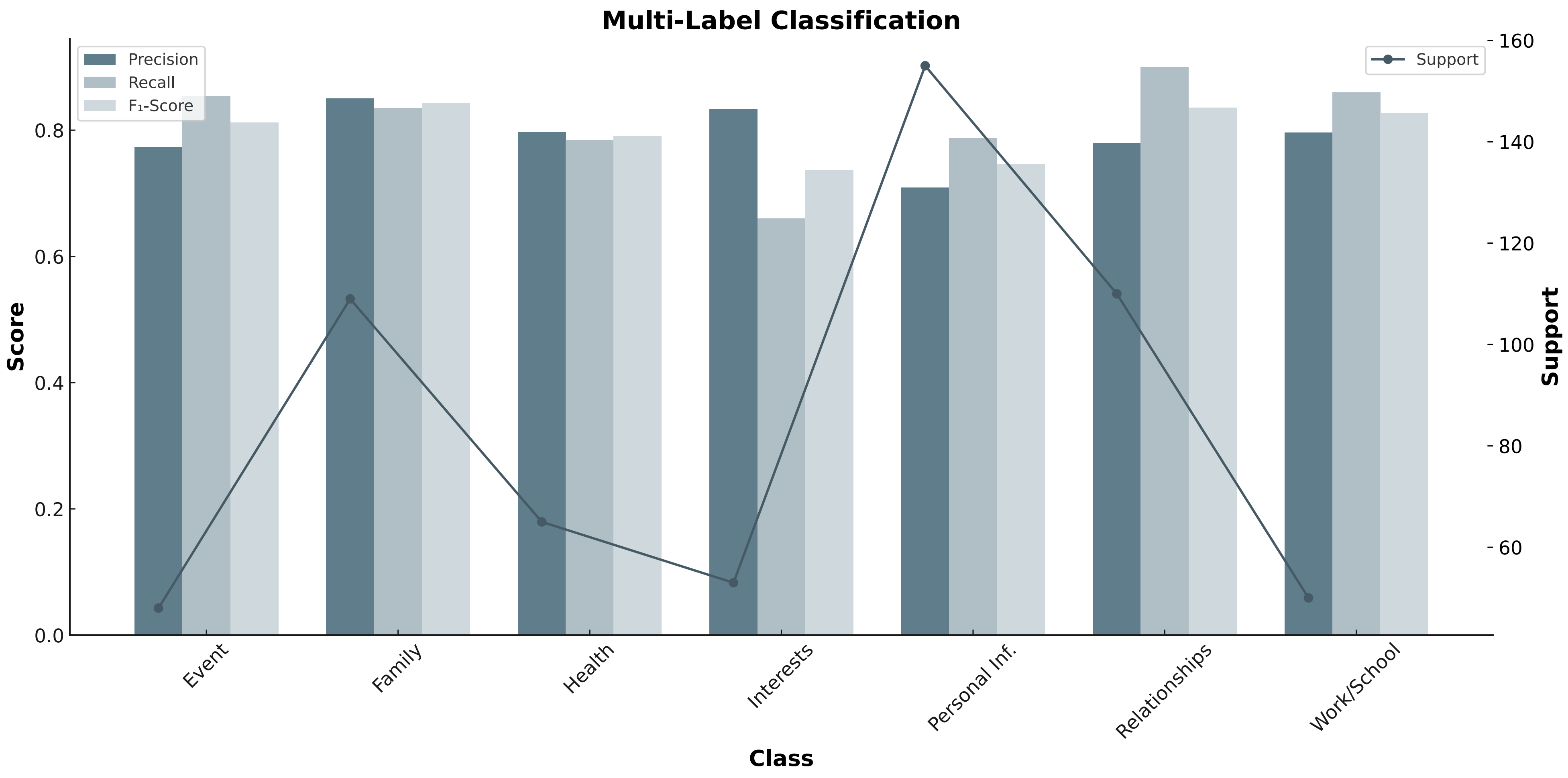
MULTI-LABEL CLASSIFICATION

- **Tweet Pre-Filtering:** Essential given the large number of tweets and the costs of GPT-4
- **Sentence Labeling:** Used „OffMyChest" dataset (Jaidka, K. et al., 2020) for annotation
- **Model Training:** Trained XLM-RoBERTa (Conneau, A. et al., 2020) with 1,438 sentences

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

HS'BI Hochschule Bielefeld University of Applied Sciences and Arts

RESULTS (XLM-ROBERTA)



RELATIONAL TRIPLE EXTRACTION

- **Method:** Extraction with GPT-4
- **Ontology-based:** Use Schema.org properties for knowledge graph
- **GPT-4 Features:** Leverage ICL features and function calling capabilities

Subject	Predicate	Object
John	s:worksFor	Microsoft
OpenAI	s:location	San Francisco, CA
Albert Einstein	s:spouse	Elsa Einstein

Person

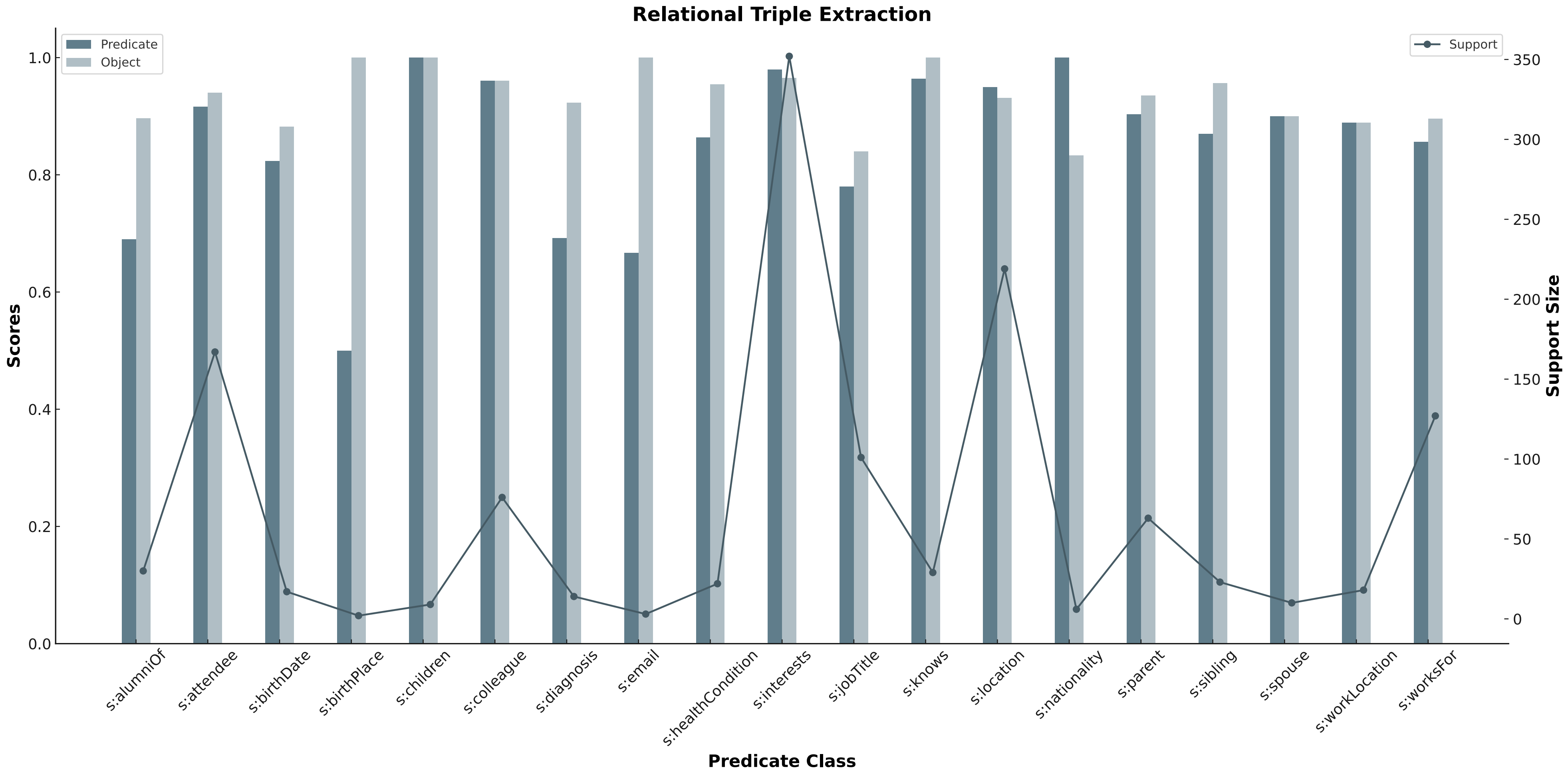
A Schema.org Type

Thing > Person

A person (alive, dead, undead, or fictional).

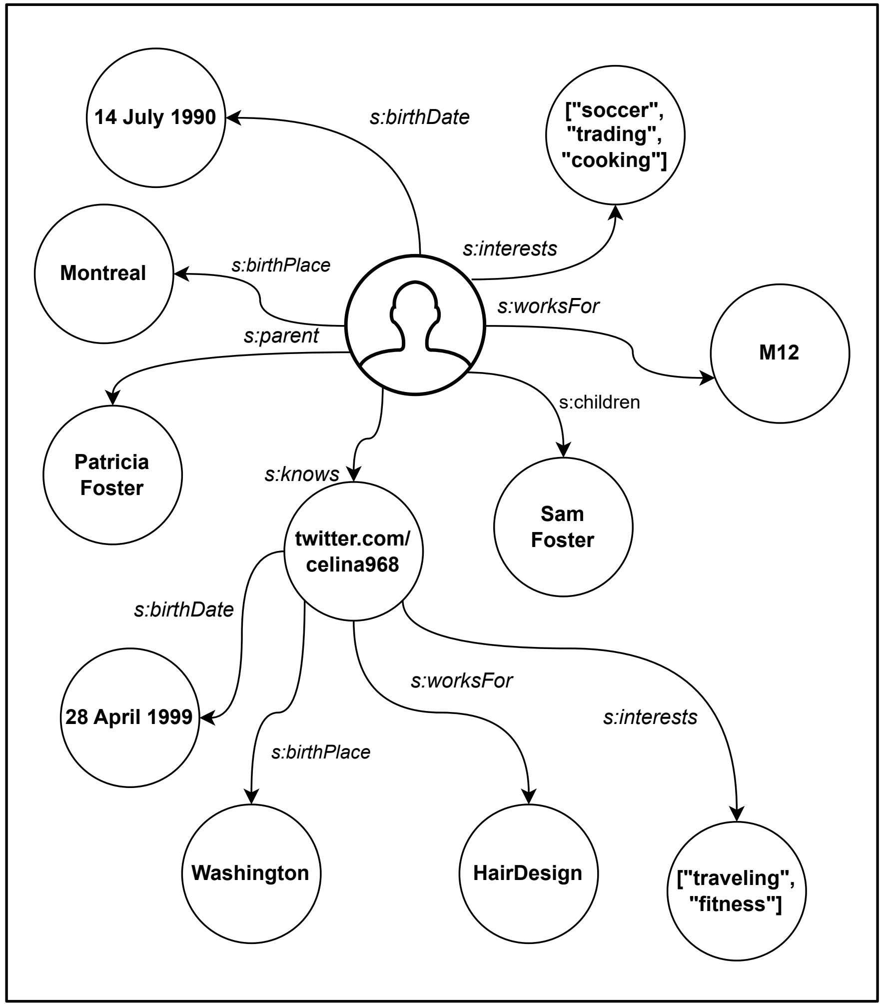
Property	Expected Type
Properties from Person	
additionalName	Text
address	PostalAddress or Text
affiliation	Organization
alumniOf	EducationalOrganization or Organization
award	Text
birthDate	Date
birthPlace	Place
brand	Brand or Organization
callSign	Text
children	Person

RESULTS (GPT-4)



DISCUSSION

- **RTE**: Extraction of **1,288 triples** from **5,000 tweets** with good results
- **ADRIAN**: The extracted triples can be leveraged to **expand** the **HDT** within the project
- **Knowledge Graph**: Enabling the **exploration** of **interconnected relationships** within extracted information



CONCLUSION & FUTURE WORK

- **GPT-4**: Provides the ability to extract information and insights from user-generated content
- **GPT-4 Features**: Applied function calling and ICL capabilities to achieve good results for extracting triples from tweets
- **Closed-Source LLMs**: Not cost-effective for processing large amounts of data
- **Open-Source LLMs**: Generate additional data to fine-tune leading open-source models such as Llama-2 (Touvron, H. et al., 2023)

REFERENCES

- Conneau, A. et al. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440-8451). Association for Computational Linguistics.
- Jaidka, K., Singh, I., Lu, J., Chhaya, N., & Ungar, L. (2020). A report of the CL-Aff OffMyChest Shared Task: Modeling supportiveness and disclosure. In Proceedings of the AAAI-20 Workshop on Affective Content Analysis (pp. 118-129). AAAI.
- Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
- Xu, X., Zhu, Y., Wang, X., & Zhang, N. (2023). How to unleash the power of large language models for few-shot relation extraction? arXiv preprint arXiv:2305.01555.
- Wan, Z., et al. (2023). GPT-RE: In-context learning for relation extraction using large language models. arXiv preprint arXiv:2305.02105.

HS'BI

Hochschule
Bielefeld

University of
Applied Sciences
and Arts

Thank you
for your attention!