h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

DISINFORMATION DETECTION:

AN EXPLAINABLE TRANSFER LEARNING APPROACH

Mina Schütz
Austrian Institute of Technology GmbH
Darmstadt University for Applied Sciences
Mina.schuetz@ait.ac.at

Alexander Schindler
Austrian Institute of Technology GmbH
alexander.schindler@ait.ac.at

Melanie Siegel
Darmstadt University for Applied Sciences
melanie.siegel@h-da.de

CODE 2021 – Science Track - Ph.D.Proposal

22.07.2021

ABOUT ME

Mina Schütz

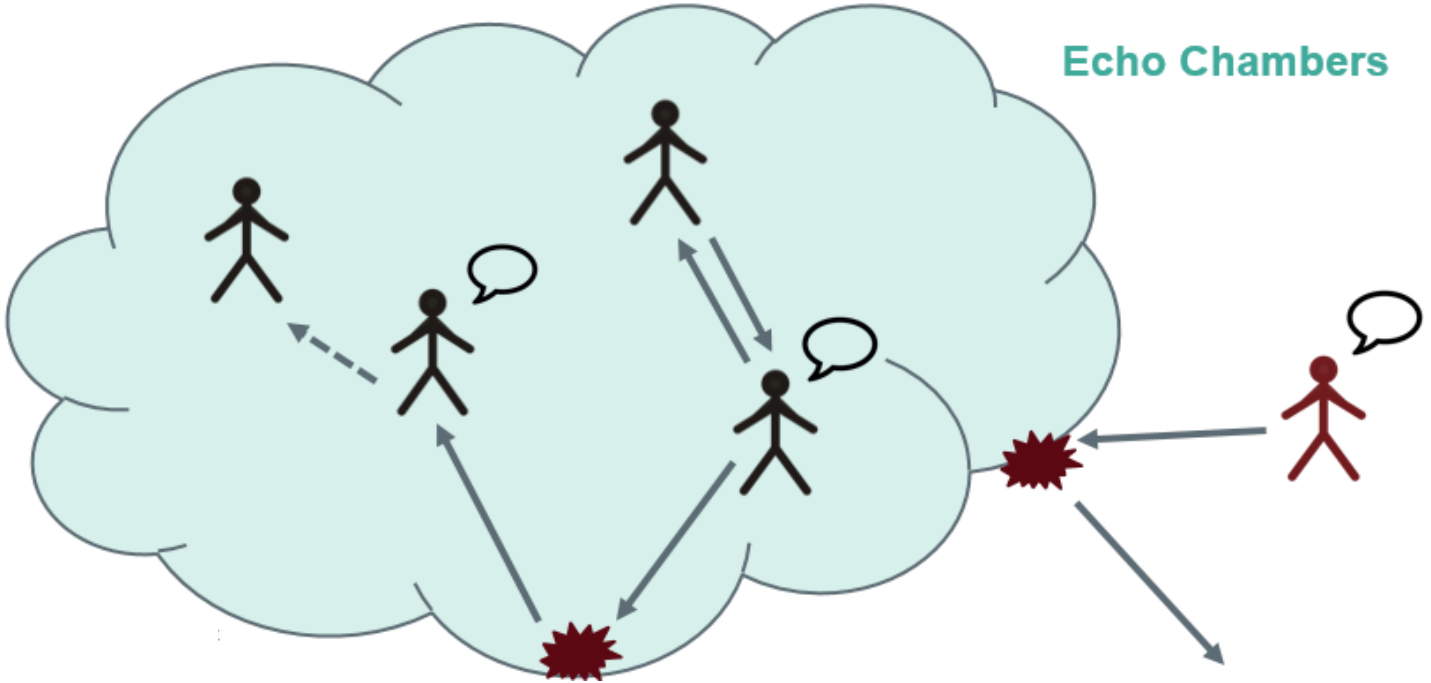
2020:	M.Sc. in Information Science (Darmstadt, University for Applied Sciences)
2019-2023:	M.Sc. + Ph.D. Student at Austrian Institute of Technology GmbH
2020-2025:	Doctorate at Darmstadt University (Applied Computer Sciences)

OUTLINE & KNOWLEDGE GAP

- **Disinformation detection** has **gained inreased focus** by research community
  - Fake News has an impact on **political** processes, **opinion** mining and **journalism**
  - **News consumption** mostly on **social media**
  - **Information overload**

- **Manual fact-checking** is **expensive** and **slow**
  - Many automatic detection techniques have been conducted, but **no final solution**
  - Often **only classification** without new insights & **black-box** models
  - **Less German datasets** available

FILTER BUBBLES



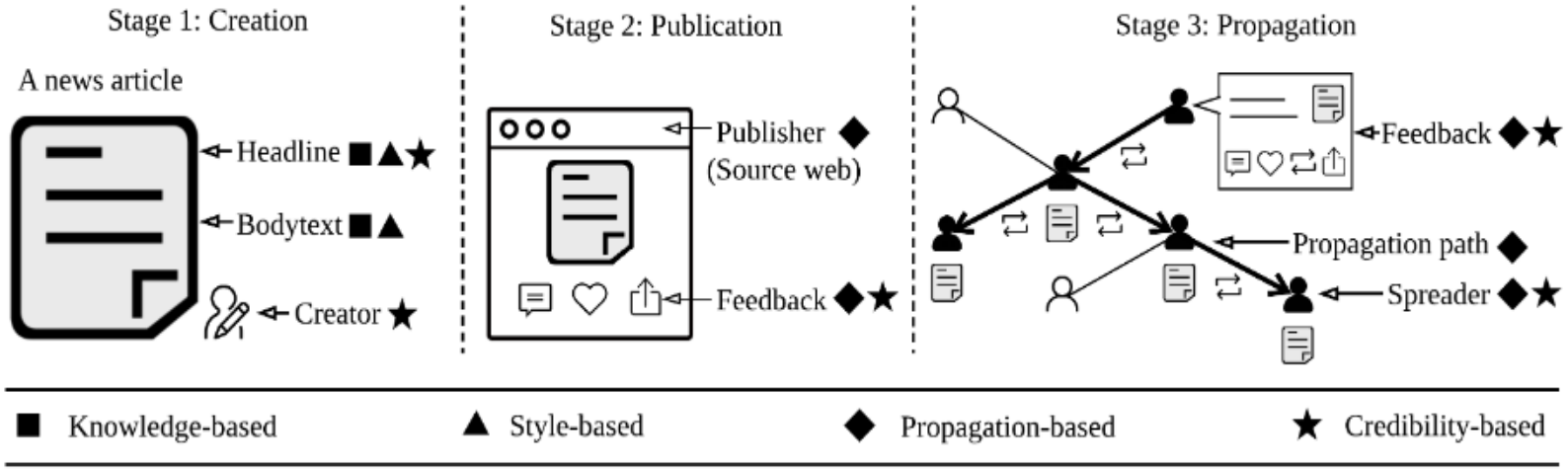
FAKE NEWS | DEFINITION

- **Defined by intention and factuality!**

- **Disinformation**
  - Intentional
  - Deceiving the reader
  - Propaganda, hoaxes, fabrications, rumors, clickbait

- **Misinformation**
  - Unintentional
  - For example: satire

DETECTION APPROACHES



RESEARCH QUESTIONS

To which extend can we obtain new linguistic knowledge about disinformation with natural language processing methods and visualization techniques?

- **Extraction of statements & narratives** (common context?)
- **Multi-lingual models** (predictions)
- **Explainable Artificial Intelligence** (XAI)
  - Visualizations
  - Breaking the black-box apart

- Generating a system **that helps users to identify misleading content**!

MAIN SCIENTIFIC CONTRIBUTION

- **Stand-alone** linguistic binary classification approaches **are not enough** to tackle the issue of fast spreading disinformation

- Diverse **dataset** in **German** (news articles & social media) & multi-lingual models

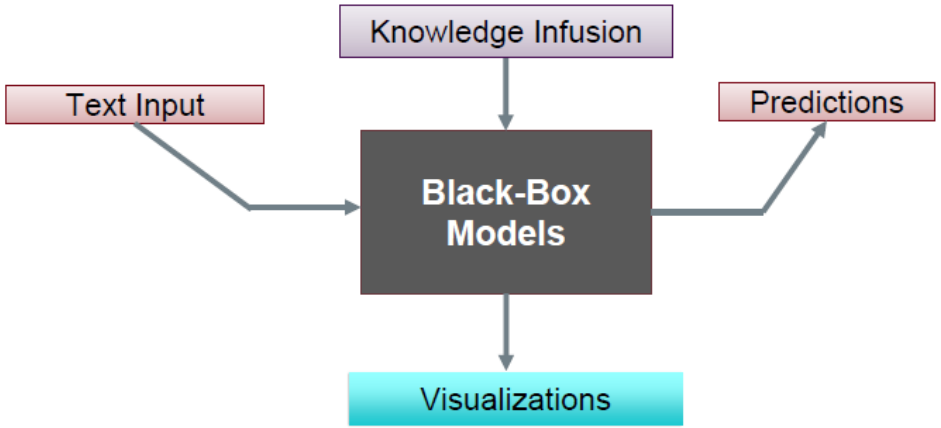
- **Combining** several **features to apply** approach for **other domains**
  - *Emotion, aggression, toxicity, sentiment*

- **Generalizability for multiple domains:**
  - False content, hate speech (sexism, racism..), conspiracy theories & propaganda

OUR APPROACH

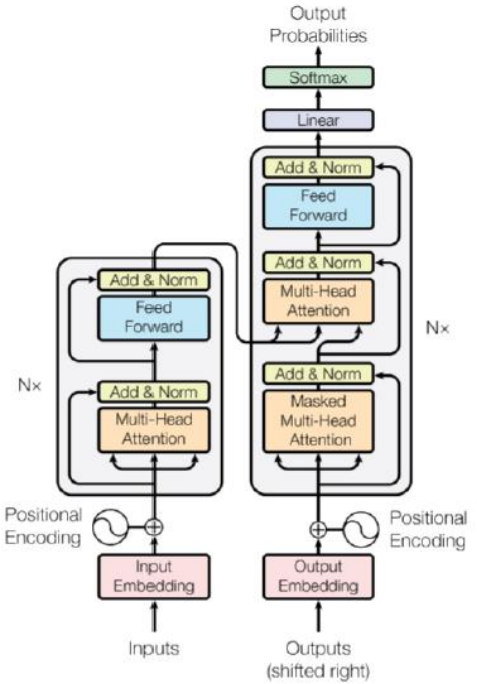
- **Explainable** AI for **understanding black-box** models & **interpreting** the **results**

- **Transfer Learning**
  - *claim* extraction & verification
  - *entity* extraction & relation
  - *event* detection



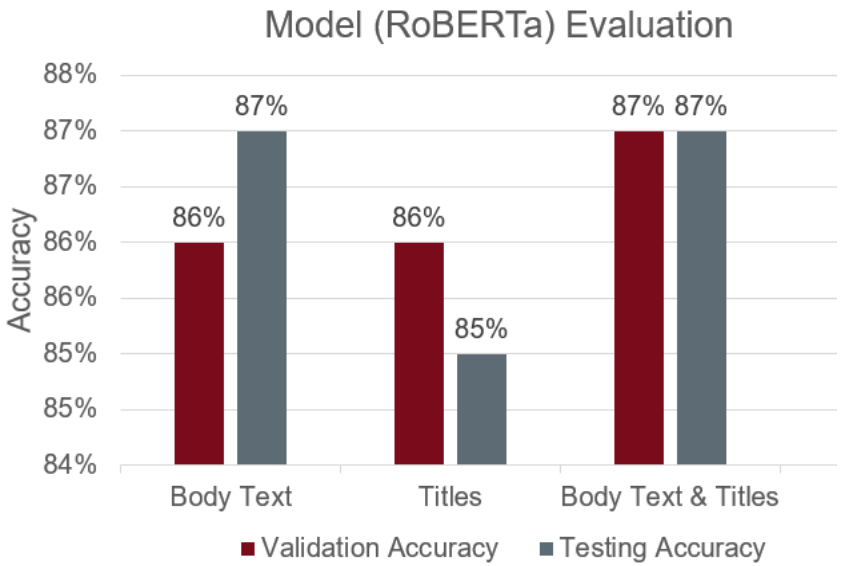
PRE-TRAINED LANGUAGE MODELS

- Latest innovation in language modeling: semi-supervised **Transformers**
- **Pre-trained** on **generic data**
- **Fine-tuned** with a dataset for the **specific task**
- **Sentence-** or **token-level-tasks**
- Using multi-head **attention**
- Architecture: **encoder** and **decoder**
- **Word embeddings** for tokenization



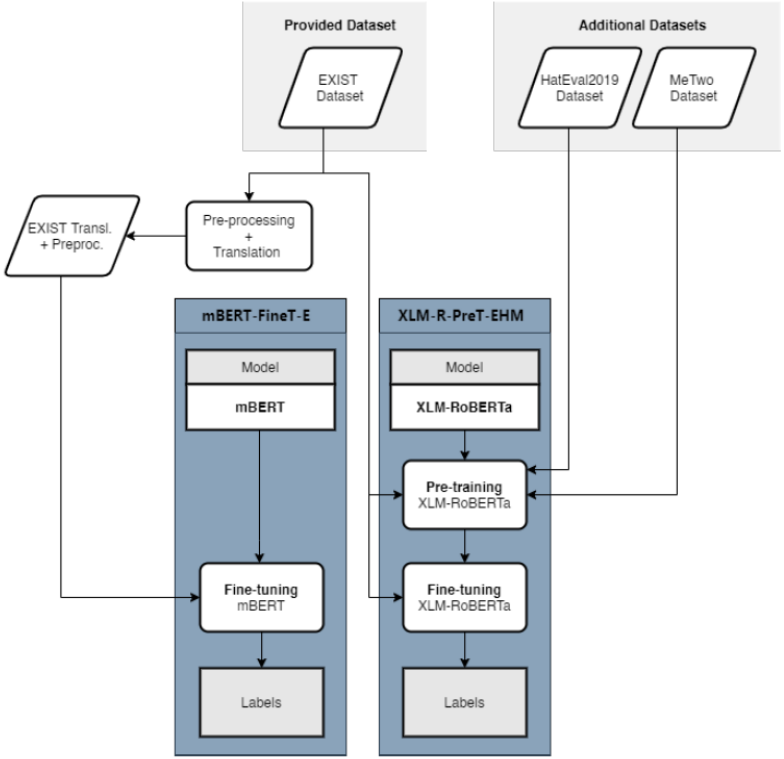
RESULTS SO FAR | MASTERTHESIS

- Binary automatic fake news classification
- 5 different Transformer
- Comparison on predictions between headlines and body text
- Evaluation of different preprocessing steps

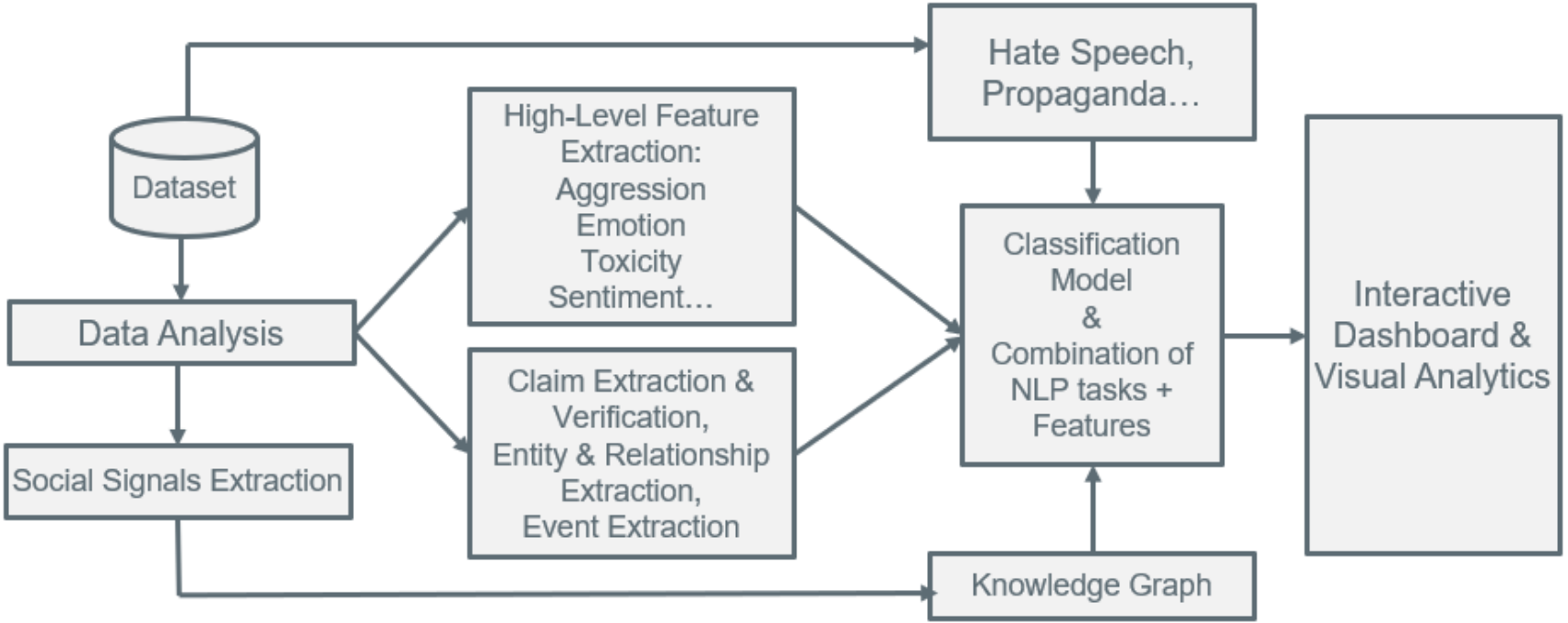


RESULTS SO FAR | EXIST2021

- Challenge on sexism detection
- Multi-class classification
- Different pre-training & fine-tuning strategies
- 3rd best team on task 1 (binary)
  - 77.52 macro-averaged F1-score

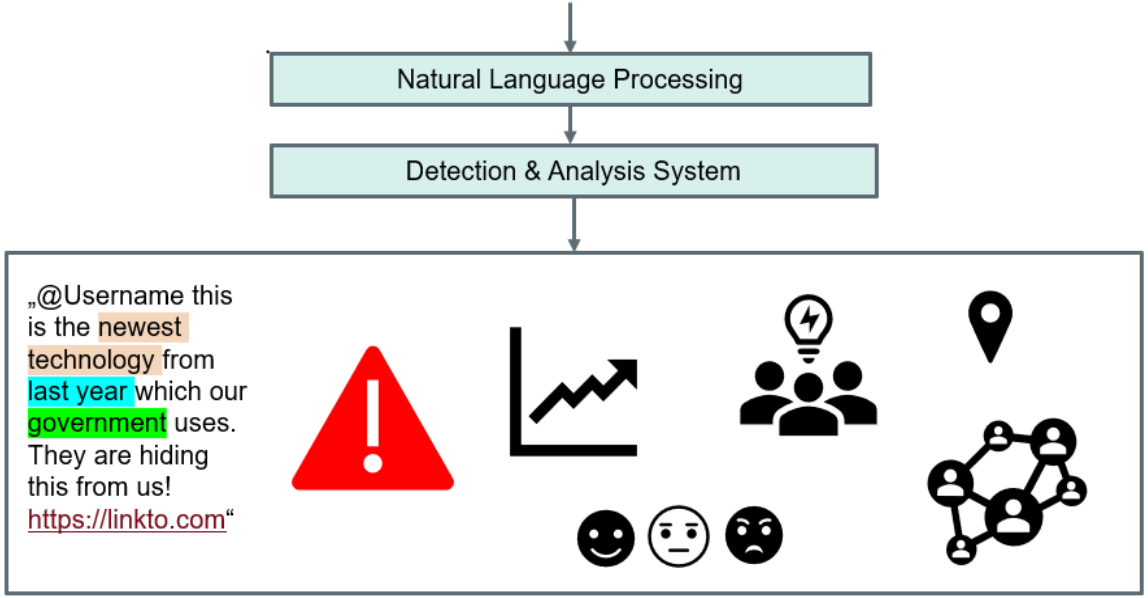


METHODOLOGY



OUTPUT

„@Username this is the newest technology from last year which our government uses. They are hiding this from us! https://linkto.com“



PROJECTS

University for Applied Sciences: DeTox (Hesse, Germany)

- „Detektion von Toxizität und Aggressionen in Postings und Kommentaren im Netz“
- Hate Speech, Toxicity, Aggression (Social Media content)
- Cybersecurity research funding of the Hessian Ministry of the Interior and Sports

Austrian Institute of Technology GmbH: Defalsif-AI (FFG KIRAS Austria)

- “Detection of false information via artificial intelligence”
- audio-visual media forensics, text analysis, and the fusion of these technologies with the help of AI

REFERENCES

[12] Mahid, Z.I., Manickam, S., Karuppayah, S.: Fake news on social media: Brief review on detection techniques. In: 2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA). pp. 1-5 (2018)

[17] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating fake news: A survey on identication and mitigation techniques. ACM Trans. Intell. Syst. Technol. 10(3) (Apr 2019)

[22] Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Comput. Surv. (2018)

[24] Zhou, X.; Zarafani, R. (2018). Fake News: A Survey of Research, Detection Meth-ods, and Opportunities. *ACM Comput. Surv. 1 (1)*, December 2018, 1 – 38. https://arxiv.org/abs/1812.00315 (Retrieved: 04.10.2019).

[11] Khan, S.A., Alkawaz, M.H., Zangana, H.M.: The use and abuse of social media for spreading fake news. In: 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS). pp. 145-148 (2019)

[8] Jr., E.C.T., Lim, Z.W., Ling, R.: Dening \fake news": A typology of scholarly definitions. Digital Journalism 6(2), 137-153 (2018)

[7] Istaiteh, O., Al-Omoush, R., Tedmori, S.: Racist and sexist hate speech detection: Literature review. In: 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA). pp. 9-99 (2020)

[15] Schütz, M., Boeck, J., Daria, L., Slijepcevic, D., Kirchknopf, A., Hecht, M., Bogensperger, J., Schlarb, S., Schindler, A., Zeppelzauer, M.: Automatic sexism detection with multilingual transformer models (2021). Arxive pre-print.

[16] Schütz, M., Schindler, A., Siegel, M., Nazemi, K.: Automatic fake news detection with pre-trained transformer models. In: Bimbo, D., et al (eds.) Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Sciences. vol. 12667. Springer, Cham (2021)

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. (2017)

[23] 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 Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171-4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019).

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

THANK YOU!

Do you have any questions?

Mina Schütz
Austrian Institute of Technology
Darmstadt University for Applied Sciences
Mina.schuetz@ait.ac.at