FZI Forschungszentrum Informatik

# German to English: Fake News Detection with Machine Translation

# Agenda

1. Introduction
2. Experiments
3. Conclusion & Outlook
4. Q&A

# 01

## –Introduction

# Introduction

- Fake News Definition: Fake news is a news article that is intentionally and verifiably false [1]
- Large part of existing datasets for Fake News Detection in English
  - BuzzFeed News [2], Fake News Challenge dataset [3], Fever [4], LIAR [5], etc.
- Pre-trained Language Models (PLMs) standard tool for NLP tasks
- Hugging Face Model Numbers [6]:
  - English **8153** VS German **650**

Given a news dataset in German, how well it works if we first translate the dataset to English and then verify the veracity of the news article with English PLMs?


# 02

## –Experiments

# Experiments

- Original Dataset
  - FANG-COVID (2021) [7], benchmark dataset for Fake News Detection in German
  - **41,242** news articles about COVID-19 pandemic
    - Online crawling
    - **28,056** news articles labeled as **Real**, from 3 reliable news agencies, Sueddeutsche Zeitung, Tagesspiegel, Zeit
    - **13,186** news articles labeled as **Fake**, from 10 unreliable news agencies, e.g., AnonymousNews, Contra-Magazin, etc.
  - Meta data
    - URL, date, source, Twitter history

# Experiments

- Dataset Translation (German to English)
  - Only news articles self are translated, meta data not included
  - **41,242** news articles, each article has **48** sentences on average
  - Neural Machine Translation Engine with **opus-mt-de-en** [8]
  - Each news article is separated into sentences with **spaCy** sentence detector [9]
  - Put each sentence into the translator to obtain corresponding English translation

# Experiments

- Methodology
  - Goal: Predicting labels of the news articles → Binary Classification (Real VS Fake)
  - Fine-tuning
    - Add classification head to pre-trained language models
    - Update the weights in the pre-trained language models for the classification task
  - Adapter
    - Add classification head to pre-trained language models
    - Add extra layers to the pre-trained language models
    - Freeze the weights in the pre-trained language models and only update the weights in the classification head and newly added layers [10]
    - Less parameters to update compared to fine-tuning and avoid catastrophic forgetting in fine-tuning [11]

# Experiments

- Implementation Details
  - Cross-entropy as loss function
  - Dataset split Train (**64%**) / Validation (**16%**) / Test (**20%**)
  - Three based models for fine-tuning and adapter methods
    - bert-base-german-cased [12] with original German input
    - bert-base-uncased [13], roberta-base [14] with translated English input
  - Fine-tuning **5** epochs with learning rate **5e-5**
  - Adapter **10** epochs with learning rate **1e-4**
  - Run each model **5** times with different seeds to avoid randomness
  - Model implemented with Hugging Face module and AdapterHub [15]

# Experiments

- Experiments results

Tab. 1: Performance of fine-tuning and adapter models

| Model | Input language | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| **Fine-tuning** | | | | | |
| bert-base-german-cased | German | 0.976 | 0.981 | 0.983 | 0.982 |
| bert-base-uncased | English | 0.971 | 0.979 | 0.979 | 0.979 |
| roberta-base | English | **0.980** | **0.983** | **0.988** | **0.985** |
| **Adapter** | | | | | |
| bert-base-german-cased | German | 0.976 | 0.978 | **0.988** | 0.983 |
| bert-base-uncased | English | 0.969 | 0.974 | 0.980 | 0.977 |
| roberta-base | English | **0.981** | **0.984** | **0.988** | **0.986** |

  - Across two groups, Fine-tuning and Adapter methods have similar performance

    → Confirm findings in [16], Adapter methods not necessarily show performance improvement against Fine-tuning on large datasets (FANG-COVID over **41k**)

  - Within each group, Roberta based model achieves the best performance among all three base models

    → Improvement of retrained roberta-base model over original bert-base model

# Experiments

- Prediction Error Analysis
  - News articles labeled as **Fake** higher probability of **4.5%** being misclassified compared to **1.5%** of news articles labeled as **Real**
  - Use Jaccard similarity $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ [17] to measure how similar the prediction errors are between two models
  - Pairwise comparison, **6** models **15** pairs
  - Similarity coefficients of prediction errors are low in general
    - Fine-tuning bert-base-uncased (English) and Adapter bert-base-uncased (English) , the highest similarity coefficient of **0.269**
    - Fine-tuning bert-base-german-cased (German) and Adapter bert-base-german-cased (German), similarity coefficient of **0.255**
    - Similarity coefficients of models with different input languages are mostly below **0.15**

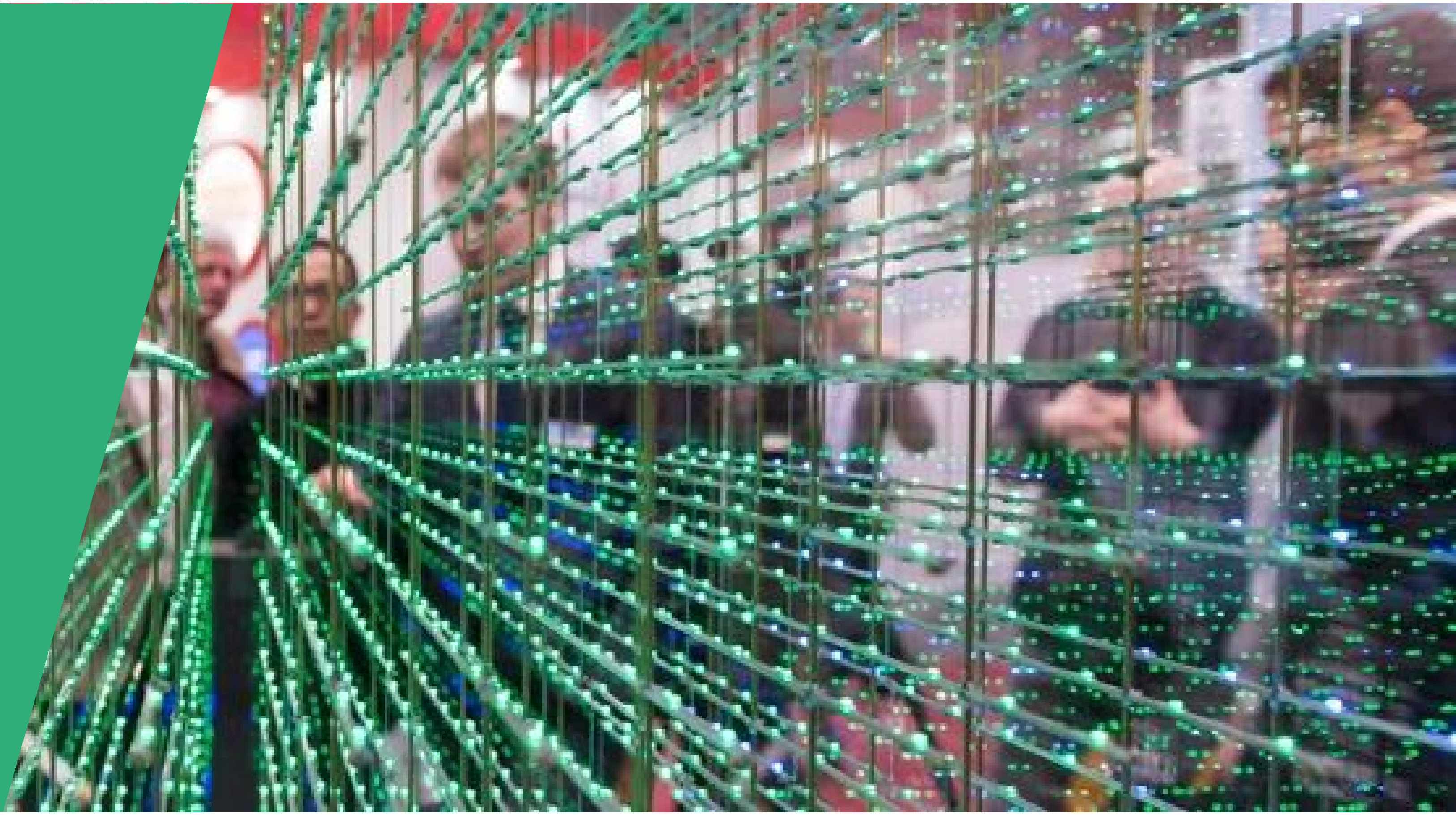
# 03

## –Conclusion & Outlook

# Conclusion & Outlook

- Conclusions
  - The errors resulting from Machine Translation from German to English can be compensated by the amount of available pre-trained language models in English.
  - Our preliminary experiments with FANG-COVID show that Machine Translation of the dataset from low-resource languages to English is a valid intermediate step.
- Outlook
  - Taxonomy with more fine-grained labels
  - Generate explanations for why a news articles is classified as fake

# 04 Q&A



## FZI Forschungszentrum Informatik

Jin Liu
jin.liu@fzi.de

Steffen Thoma

thoma@fzi.de

Haid-und-Neu-Str. 10-14
76131 Karlsruhe

**www.fzi.de**

# Reference


[1] Shu, K., Silva, A., Wang, S., Tang, J., Liu, H.: Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Exploration Newsletter 19(1), 22-36 (2017)

[2] BuzzFeedNews, https://github.com/BuzzFeedNews. Last accessed 06 September 2022

[3] FakeNewsChallenge, https://github.com/FakeNewsChallenge/fnc-1. Last accessed 06 September 2022

[4] Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp. 809-819, (2018)

[5] Wang, W.: "Liar, Lair Pants on Fire": A New Benchmark Dataset for Fake News Detection. In proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422-426, (2017)

[6] Hugging Face models, https://huggingface.co/models. Last accessed 15 September 2022

[7] Mattern, J., Qiao, Y., Kerz, E., Wiechmann, D., Strohmaier, M.: FANG-COVID: A New Large-Scale Benchmark Dataset for Fake News Detection in German. In Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), pp. 78-91. Association for Computational Linguistics, Dominican Republic (2021)

[8] Tiedmann, J., Thottingal, S.: OPUS-MT - Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pp. 479–480, European Association for Machine Translation, Lisboa, Portugal (2020)

[9] Honnibal, A., Montani, I., Van Landeghem, S., Boyd, A.: spyCy: Industrial-strength Natural Language Processing in Python. 2020

[10] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-Efficient Transfer Learning for NLP. In Proceedings of the 36th International Conference on Machine Learning, pp. 2790–2799, (2019)

[11] Lauscher, A., Majewska, O., Ribeiro, L., Gurevych, I., Rozanov, N., Glavaš, G.: Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers. In Proceedings of Deep Learning Inside Out, pp. 43–49, Online, (2020)

[12] deepset Homepage, https://www.deepset.ai/german-bert. Last accessed 30 May 2022

[13] Delvin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, (2018)

# Reference


[14] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, CoRR, (2019)

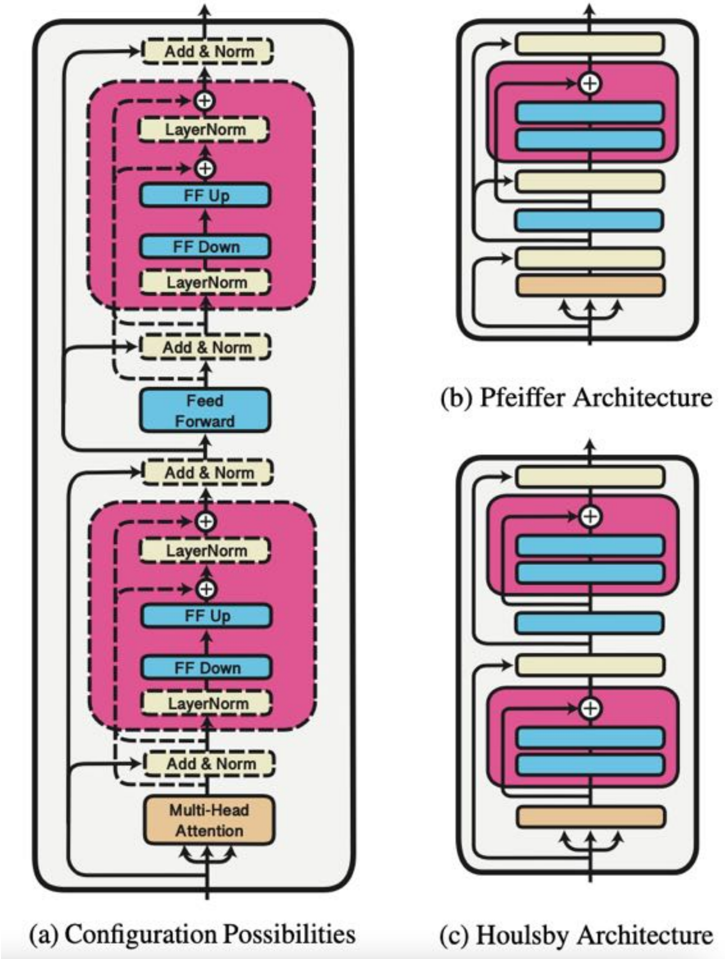
[15] Pfeiffer, J., Rückl., A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., Gurevych, I., AdapterHub: A Framework for Adapting Transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): System Demonstrations, pp. 46–54, Online, (2020)

[16] He, R., Liu, L., Ye, H., Tan, Q., Ding, B., Cheng, L., Low, J., Bing, L., Si, L.: On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation. In Proceedings of the 59 Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 2208–2222, Association for Computational Linguistics (2021)

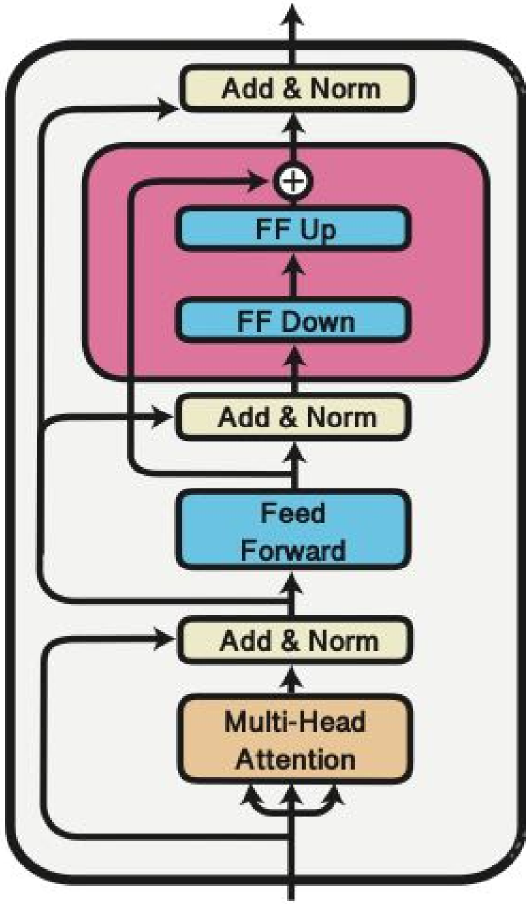
[17] Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S.: Using of Jaccard Coefficient for Keywords Similarity. In Proceedings of the International MultiConference of Engineers and Computer Scientist 2013 Vol I, Hong Kong (2013)

[18] Rücklé, A., Geigle, G., Glockner, M., Beck, T., Pfeiffer, J., Reimers, N., Gurevych, I.: AdapterDrop: On the Efficiency of Adapters in Transformers. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7930—7946, Online and Punta Cana, Dominican Republic, (2021)

# Backup




Configuration possibilities for Adapter models [15]




Pfeiffer configuration [18]