Czech
Technical University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

NLP Methods for Automated Fact-Checking

Dissertation Minimum Study of

Ing. Herbert Ullrich

FCHECK.FEL.CVUT.CZ

Supervisor: **Ing. Jan Drchal, Ph.D.**
Field of study: **Informatics**
Subfield: **Natural Language Processing**
August 2023

Chapter 1
Introduction

My dissertation, as well as my long-term research, centers around the field of *automated fact checking* through the means of Natural Language Processing and its modern methods. The work consists of the analysis of the whole fact-checking process, its subdivision and simplification into tasks that can be efficiently addressed using the current state-of-the-art NLP methods, collection of data appropriate to benchmark such tasks, delivery of example solutions and their validation against similar research in other languages and related tasks.

The main focus of our group are the fact-checking-related tasks in the West Slavic languages (Czech, Slovak and Polish) and secondarily in English. My contribution has so far been the collection and publication of novel datasets for the fact-checking task and its subroutines, models trained for the tasks and their debate, including the ongoing establishment of metrics that would rate the model success and error rates in terms close to the human notion of *facticity* (which proves to be a challenge on its own, requiring another round of novel research).

My doctoral aim is to cover every step on the path from gathering a factual claim – for example, extracting it from a political debate – to predicting its veracity verdict and justifying it rigorously with hard data. With the recent boom in NLP beginning with the advent of transformer networks and later the Large Language Models, prompting and few-shot learning, a significant part of the research is and has to be an appropriate and timely adoption of new ever-evolving sota NLP solutions, based on well-designed studies in our specific context.

Overall, my agenda is to follow up on our published research on fact checking in Czech with methods that reiterate on our results in other languages and evolving our previous methodology based on transformer *pre-training & fine-tuning* paradigm to a computationally feasible design based on LLMs. I want to establish the task of *claim generation* among the other commonly benchmarked NLP tasks within the scientific community, adjacent to that of *abstractive summarization.* I aim to give safeguards and explanations to the model decisions with human-understandable metrics, in particular revealing hallucinations – a common problem of modern day LLMs.

The goal of this study is to show the directions I am taking to address these challenges, reasoning behind them, my research questions and current results that motivated them.

1.1 Motivation

The spread of misinformation in the online space has a growing influence on the Czech public [STEM, 2021]. It has been shown to influence people's behaviour on the social networks [Lazer et al., 2018] as well as their decisions in elections [Allcott and Gentzkow, 2017], and real-world reasoning, which has shown increasingly harmful during the COVID-

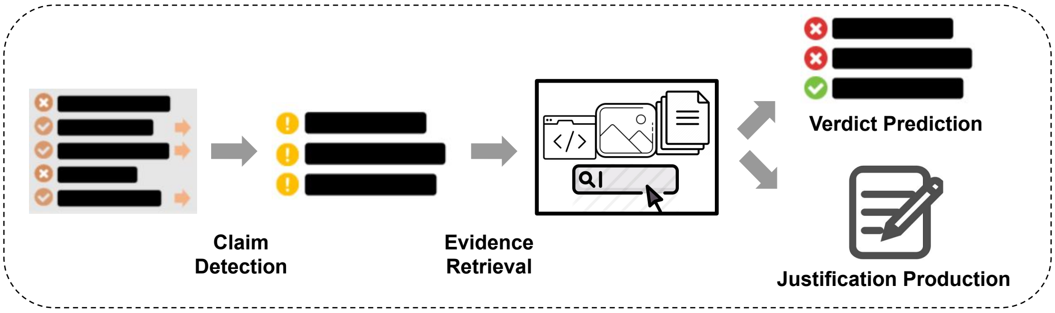


Figure 1.1: Automated fact-checking pipeline, reprinted from [Guo et al., 2022]

19 pandemic [Barua et al., 2020].

The recent advances in artificial intelligence and its related fields, in particular the recommendation algorithms, have contributed to the spread of misinformation on social media [Buchanan and Benson, 2019], as well as they hold a large potential for automation of the false content generation and extraction of sensational attention-drawing headlines – the "clickbait" generation [Shu et al., 2018].

Recent research has shown promising results [Thorne et al., 2019] in false claim detection for data in English, using a trusted knowledge base of true claims (for research purposes typically fixed to the corpus of Wikipedia articles), mimicking the *fact-checking* efforts in journalism.

Fact-checking is a rigorous process of matching every information within a *factic claim* to its *evidence* (or *disproof*) in trusted data sources to infer the claim veracity and verifiability. In exchange, if the trusted *knowledge base* contains a set of "ground truths" sufficient to fully infer the original claim or its negation, the claim is labelled as **supported** or **refuted**, respectively. If no such *evidence set* can be found, the claim is marked as **unverifiable**[1].

1.2 Challenges

Despite the existence of end-to-end fact-checking services, such as `politifact.org` or `demagog.cz`, the human-powered approach shows weaknesses in its scalability. By design, the process of finding an exhaustive set of evidence that decides the claim veracity is much slower than that of generating false or misguiding claims. Therefore, efforts have been made to move part of the load to a computer program that can run without supervision.

The common research goal is a fact verification tool that would, given a claim, semantically search provided knowledge base (stored for example as a *corpus* of some natural language), propose a set of evidence (e. g. k semantically nearest paragraphs of the corpus) and suggest the final verdict (Figure **??**). This would reduce the fact-checker's workload to mere adjustments of the proposed result and correction of mistakes on the computer side.

The goal of the ongoing efforts of FactCheck team at AIC CTU, as addressed in the works of [Rýpar, 2021,Dědková, 2021] and [Gažo, 2021] **TODO:** Good sources is to explore the state-of-the-art methods used for fact verification in other languages, and propose a strong baseline system for such a task in Czech.

[1] Hereinafter labelled as `NOT ENOUGH INFO`, in accordance to related research.

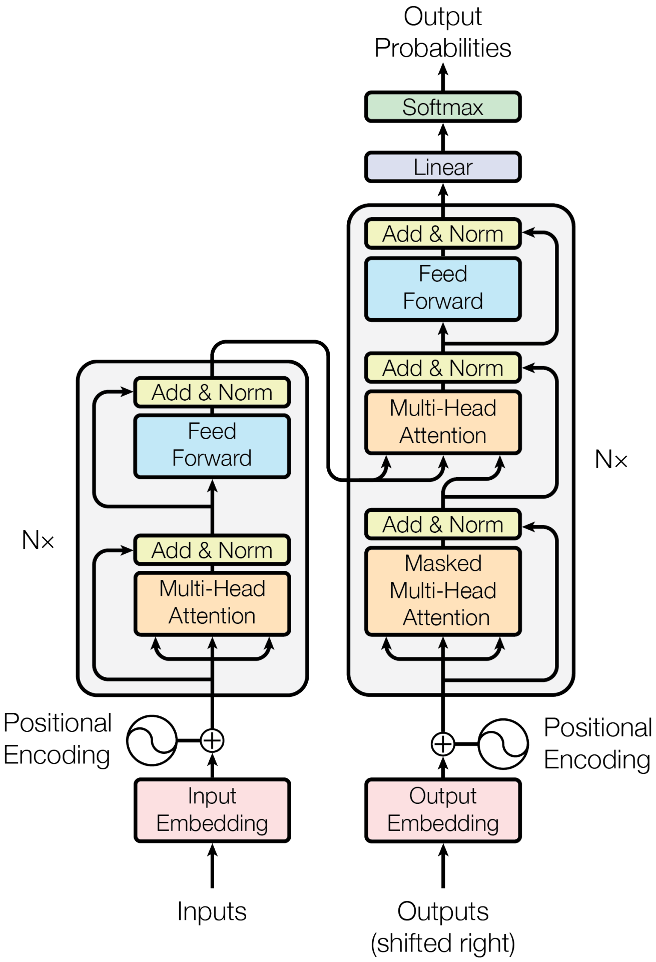


Figure 1.2: Transformer model architecture, reprinted from [Vaswani et al., 2017]

1.3 A word on the Transformers

For the past six years, the state-of-the-art solution for nearly every Natural Language Processing task is based on the concept of *transformer networks* or, simply, *Transformers*. This has been a major breakthrough in the field by [Vaswani et al., 2017], giving birth to the famous models such as Google's BERT [Devlin et al., 2019] and its descendants, or the OpenAI's GPT-3 [Brown et al., 2020].

In our proposed methods, we use Transformers in every step of the fact verification pipeline. Therefore, we would like to introduce this concept to our reader to begin with.

Transformer is a neural model for *sequence-to-sequence* tasks, which, similarly e.g. to the *LSTM-Networks* [Cheng et al., 2016], uses the Encoder–Decoder architecture. Its main point is that of using solely the *self-attention* mechanism to represent its input and output, instead of any sequence-aligned recurrence [Vaswani et al., 2017].

In essence, the *self-attention* (also known as the *intra-attention*) transforms every input vector to a weighted sum of the vectors in its neighbourhood, weighted by their *relatedness* to the input. One could illustrate this on the *euphony* in music, where every tone of a song relates to all of the precedent ones, to some more than to the others.

The full Transformer architecture is depicted in Figure 1.2.

1.4 Dissertation minimum study outline

- **Chapter 1** introduces the dissertation topic, motivates the research sets up our challenges for the future research
- **Chapter 2** examines the most relevant research in the field and tries to highlight the recent paradigm shift from models trained for a single task to a single large models that perform well in everything
- **Chapter 3** explains our current contributions to the field of automated fact-checking and NLP in Czech
- **Chapter 4** describes our plan for the dissertation and justifies the directions we are taking
- Finally, **Chapter 5** concludes the study with a wrapup of its findings

Chapter 2
State of the Art

This chapter will first describe the originally popular models for NLP such as BERT and its relatives and then the paradigm shift from pre-train+fine-tune frameworks to LLMs (possibly few-shot learned or adapted with LoRA).

TODO: Citations: Lora, daniil, Bert paper, GPT4 debilní report

2.1 Pre-train + Fine-tune

Surprisingly well performing paradigm for small data, vague tasks

2.1.1 BERT and derivatives

Simple training task, scalable architecture, potential to evolve into something grand

...Which currently arrives in form of:

2.2 Large Language Models

2.2.1 GPT-3.5 and GPT-4

Obscure, paid

2.2.2 LLaMA-2 and derivatives

Open-source, freely usable. Often poor czech coverage Quantizace, 4b, 8b, zlomek parametrů

2.3 Fact checking approaches

2.3.1 FEVER and followups

Yields interesting benchmark with statistically quantifiable model succes, oversimplificates the problem, as it uses Wikipedia for a trusted knowledge base and only reasons based on a data from a fixed period of time, focusing on “atomic” claims that do not match the complexity of real-world factoids.

2.3.2 Open-domain fact-checking

This paper with bing for example uses the whole internet, but is that really what we want? Like, every lie can be backed with an internet – at the end of the day you do need to draw the line of what to trust somewhere, which directly conflicts this design.

2.4 Claim generation

- Approches such as QACG exploit Question Answering
- The task of extreme summarization (XSum) focuses on summarizing a long body of text into a single sentence, focusing on its most relevant aspects and facts
- CLEF-CheckThat postulates the task of classifying *Checkworthiness* of different parts of a long texts, such as a political debate

2.5 NLP Generative task benchmarking

2.5.1 BERTScore

2.5.2 AlignScore

Chapter 3
Current Contribution

We have collected novel data, emulated and scraped inavailable datasets making them public or readying them for doing so, we have established numerous sota models and are currently working on establishing the topic of claim generation as a summarization-related NLP task. We are also readying metrics for fact-checking, experimenting with them and so on and soforth.

3.1 Datasets

3.1.1 CsFEVER

3.1.2 CTKFacts

3.1.3 Other NLP datasets in West Slavic languages

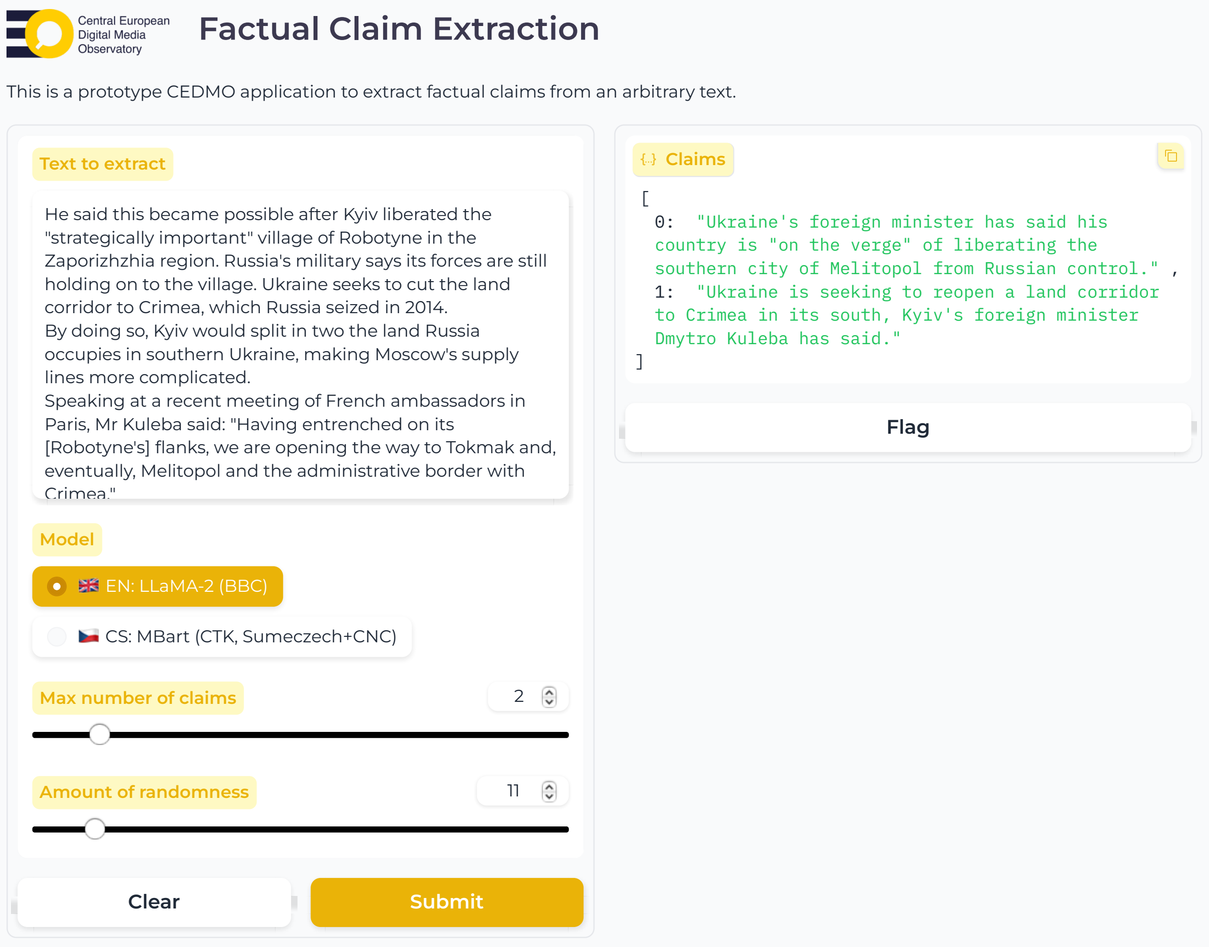
1. **Translated NLI datasets** – SNLI, ANLI, MultiNLI,
2. SmeSum, CTKSum, CsFEVERSum
3. Polish summarization data

3.2 Models

3.3 Publications

3.4 Applications

Here we will show off the demonstration tools, as well as our open-source platform `https://fcheck.fel.cvut.cz` and currently running claim extraction tools.



Factual Claim Extraction

This is a prototype CEDMO application to extract factual claims from an arbitrary text.

Text to extract

He said this became possible after Kyiv liberated the "strategically important" village of Robotyne in the Zaporizhzhia region. Russia's military says its forces are still holding on to the village. Ukraine seeks to cut the land corridor to Crimea, which Russia seized in 2014.
By doing so, Kyiv would split in two the land Russia occupies in southern Ukraine, making Moscow's supply lines more complicated.
Speaking at a recent meeting of French ambassadors in Paris, Mr Kuleba said: "Having entrenched on its [Robotyne's] flanks, we are opening the way to Tokmak and, eventually, Melitopol and the administrative border with Crimea."

Model

● 🇬🇧 EN: LLaMA-2 (BBC)

○ 🇨🇿 CS: MBart (CTK, Sumeczech+CNC)

Max number of claims 2

Amount of randomness 11

Claims

```
[
  0:  "Ukraine's foreign minister has said his
  country is "on the verge" of liberating the
  southern city of Melitopol from Russian control." ,
  1:  "Ukraine is seeking to reopen a land corridor
  to Crimea in its south, Kyiv's foreign minister
  Dmytro Kuleba has said."
]
```

Figure 3.1: Factual claim extraction application done for the CEDMO project

Chapter 4
Dissertation plan

4.1 Current research agenda

4.1.1 Automated claim generation

TODO:

4.1.2 Generative task facticity metrics

The common problem with generative tasks in NLP is that of explaining their reasoning in human-understandable manner and troubleshooting the prediction faults, such as the *model hallucination*.

For the task of claim generation, where we also face the challenge of the *relevance* of the information extracted by the model, we postulate the following metrics:

1. **Fluency** – LM-Critic [Yasunaga et al., 2021] perturbs the input to find local optima in output probability of its tokens, using a language model such as GPT-2 as its reference. GPTScore [Fu et al., 2023] uses prompting a LLM (such as GPT-3.5) to obtain a model-inferred score using zero-shot learning.

2. **Atomicity** – can be checked using the Relationship Extraction methods: we extract the fact triples and mark the claim as atomic if there is at most one of them (after removing symmetries)

3. **Faithfulness** – we proceed to use the score proposed within the FFCI evaluation framework: $AvgTopN(BERTScore(t_i, s_j))$. Similar metric was proposed in [Zha et al., 2023], looking for optimum alignment of output and parts of input, using an arbitrary

4. **Focus and Coverage** – we use the question-answering-based solutions of QAGS [Wang et al., 2020]

4.2 Data Collection

4.2.1 Human-in-the-loop grading of claim generators

4.2.2 Validation of the model outputs with human fact-checkers

4.2.3 Polish dataset scraping

Will be first of its kind for NLP purposes

4.3 The grand scope

1. **Claim extraction metrics proposal based on factuality of summarization**
2. **Claim extraction paradigm that benchmarks best in the newly given metrics**
3. Systems for NLI built on top of LoRA paradigm to score best in the task, as showed promising by Daniil

Chapter 5
Conclusion

Bibliography

[Allcott and Gentzkow, 2017] Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

[Barua et al., 2020] Barua, Z., Barua, S., Aktar, S., Kabir, N., and Li, M. (2020). Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, 8:100119.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.

[Buchanan and Benson, 2019] Buchanan, T. and Benson, V. (2019). Spreading disinformation on facebook: Do trust in message source, risk propensity, or personality affect the organic reach of “fake news”? *Social Media + Society*, 5(4):2056305119888654.

[Cheng et al., 2016] Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Dědková, 2021] Dědková, B. (2021). Multi-stage methods for document retrieval in the czech language.

[Fu et al., 2023] Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2023). Gptscore: Evaluate as you desire.

[Gažo, 2021] Gažo, A. (2021). Algorithms for document retrieval in czech language supporting long inputs.

[Guo et al., 2022] Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

[Lazer et al., 2018] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

[Rýpar, 2021] Rýpar, M. (2021). Methods of document retrieval for fact-checking. `https://www.overleaf.com/read/thbvcjvvvfjp`. [Online; accessed 21-May-2021].

[Shu et al., 2018] Shu, K., Wang, S., Le, T., Lee, D., and Liu, H. (2018). Deep headline generation for clickbait detection.

[STEM, 2021] STEM (2021). Mýtům a konspiracím o covid-19 věří více než třetina české internetové populace | stem.cz. `https://www.stem.cz/mytum-a-konspiracim-o-covid-19-veri-vice-nez-tretina-ceske-internetove-populace/`. Accessed: 2021-05-03.

[Thorne et al., 2019] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2019). The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Wang et al., 2020] Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

[Yasunaga et al., 2021] Yasunaga, M., Leskovec, J., and Liang, P. (2021). LM-critic: Language models for unsupervised grammatical error correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7763, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Zha et al., 2023] Zha, Y., Yang, Y., Li, R., and Hu, Z. (2023). Alignscore: Evaluating factual consistency with a unified alignment function.

Appendix A
Acronyms

BERT Bidirectional Encoder Representations from Transformers

GPT Generative Pre-trained Transformer

FEVER Fact Extraction and Verification – series of Shared tasks focused on fact-checking

CLI Command-Line Interface

NLI Natural Language Inference

ČTK Czech Press Agency