Language Technologies for fighting disinformation

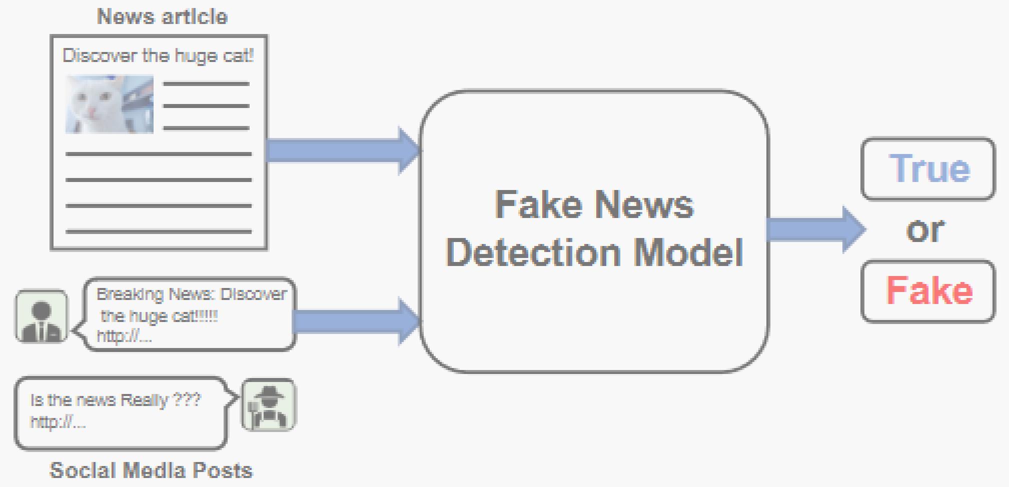
Maria Bielikova, Marian Simko
Kempelen Institute of Intelligent Technologies

Central European Digital Media Observatory

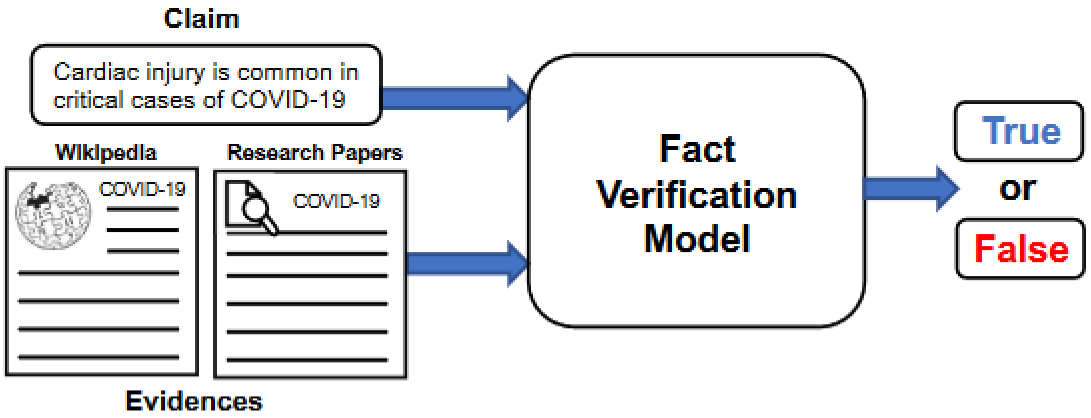
Co-financed by the Connecting Europe Facility of the European Union

KInIT

Two lines of research: Fake news detection vs. Fact verification (fact checking)*


News article
Discover the huge cat!
Breaking News: Discover the huge cat!!!!! http://...
Is the news Really ??? http://...
Social Media Posts
Fake News Detection Model
True
or
Fake

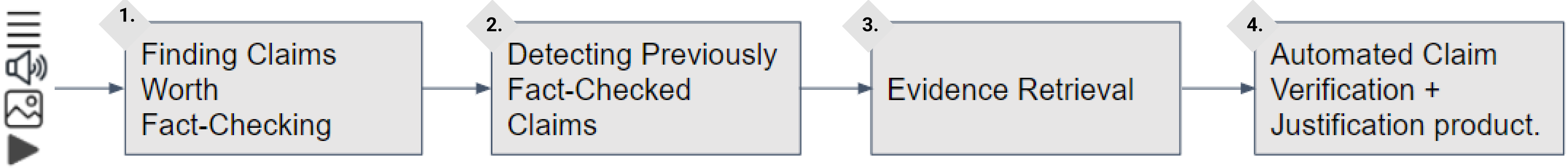
- Indirect features (content style, propagation patterns, author and source credibility, …)
- Low explainability
- Appropriate for early detection (if new disinformation cases share the characteristics with prior ones)



- Assessment of actual content veracity
- Better explainability
- Evidence needed

**Figures taken from Murayama (2021)*

Fact-checking tasks



1. Finding Claims Worth Fact-Checking

2. Detecting Previously Fact-Checked Claims

3. Evidence Retrieval

4. Automated Claim Verification + Justification product.

Ranking

Input: Document

Output: Ranked list of factual statements

Ranking

Input: Claim (or whole document) + database of previous fact-checks

Output: List of previously fact-checked claims ranked based on their **similarity** with the input claim or their **presence** in the document

Ranking

Input: Claim + database of evidence documents

Output: Ranked list of evidence documents

Classification

Input: Claim + evidence

Output: Veracity label (true/false, true/mostly-true/mixed/…)

Nakov et al., 2021; Guo et al., 2022; Zeng et al., 2021

Main underlying NLP tasks

- Claim matching 2. 3.
    - Identification of the semantically equivalent (or similar) occurrences of a given claim in a larger unit of text
- Stance detection 2. 4.
    - Detection of stance (position) of an author of an input piece of text towards a specified target
- Textual entailment (NLI) 4.
    - Verification whether the retrieved evidence (premise) supports or refutes the claim (hypothesis)

Existing LR and LT

- **Datasets of claims** from social media (mainly Twitter) or political debates; see Guo et al., 2022 for an overview

- **Factual verification datasets** from fact-checking sites (e.g., Politifact) or with artificial inputs (mainly generated from Wikipedia); see Guo et al., 2022 for an overview

- Datasets collected at CLEF CheckThat! Lab
  - For tasks 1. - 4.
  - In English and some additional languages (e.g., Arabic) differing from task to task

Challenges and LR needed

- Multilinguality and low resource languages
  - Dataset(s) of fact-checked claims in one language mapped to claims/documents in other languages
  - Typical use case: For global events, such as COVID-19 pandemic or war in Ukraine, disinformation crosses borders and languages, but is often fact-checked only in some of the languages
- Multimodality
  - Disinformation can combine several modalities (e.g., text and images)
  - Typical use case: viral memes or a social media post with a fabricated, manipulated or misinterpreted image/video
- Credibility – explainability and mitigation of biases

Steer clear of the phrase “automated fact checking” to avoid alienating potential users of automation technology: instead focus on **collaborating with fact checkers** and drawing on their expertise to identify which **repetitive tasks** can be done **reliably** by machines.

FullFact, 2020

Maria Bielikova

Expert researcher & Director KInIT
Strategy Board member CEDMO
maria.bielikova@kinit.sk

Kempelen Institute
of Intelligent Technologies

Nivy Tower
Mlynské Nivy II. 18890/5
811 09 Bratislava
Slovakia

www.kinit.sk

Central European
Digital Media
Observatory

Co-financed by the Connecting Europe
Facility of the European Union

List of References

1. FullFact. 2020. The challenges of online fact checking. Technical report. https://fullfact.org/media/uploads/coof-2020.pdf

2. Guo Z., Schlichtkrull, M., Vlachos, A. 2022. A Survey on Automated Fact-Checking. Transactions of the Association for Computational Linguistics. 10 178–206. https://doi.org/10.1162/tacl_a_00454

3. Küçük, D., Can, F. 2021. Stance Detection: A Survey. Comput. Surveys 53, 1 (Jan. 2021), 1–37. https://doi.org/10.1145/3369026

4. Murayama, T. 2020. Dataset of Fake News Detection and Fact Verification: A Survey. https://doi.org/10.48550/arXiv.2111.03299

5. Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., Da San Martino, G. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. https://doi.org/10.48550/arXiv.2103.07769

6. Zeng, X., Abumansour, A. S., & Zubiaga, A. 2021. Automated fact-checking: A survey. Language & Linguistics Compass, e12438. https://doi.org/10.1111/lnc3.12438