Media Futures

Media Content Analysis and Production:

Automated Fact-checking

Ghazaal Sheikhi

Postdoc Fellow

Research Centre for Responsible Media Technology & Innovation

Project number 309339

Disinformation, misinformation and fake news

- Disinformation: “dissemination of false information with the deliberate intent to deceive or mislead”
- Misinformation: “the unintentional dissemination of false information”
- Fake news: “originally U.S. news that conveys or incorporates false, fabricated, or deliberately misleading information, or that is characterized as or accused of doing so”
  - Fake news is a typical example of online disinformation.
  - Six types of fake news include satire, fabrication, parody, photo manipulation, advertising, and propaganda

Fact-Checking

"Seek truth and report it"

(The Society of Professional Journalists Code of Ethics)

Internal fact-checking

- Internal fact-checking (dated back to 1920s): the verification routines prior to publication to ensure factual accuracy.
  - Searching for common errors such as in numbers, statistics, names, dates, superlatives etc.
  - Checking the primary sources and verify the facts

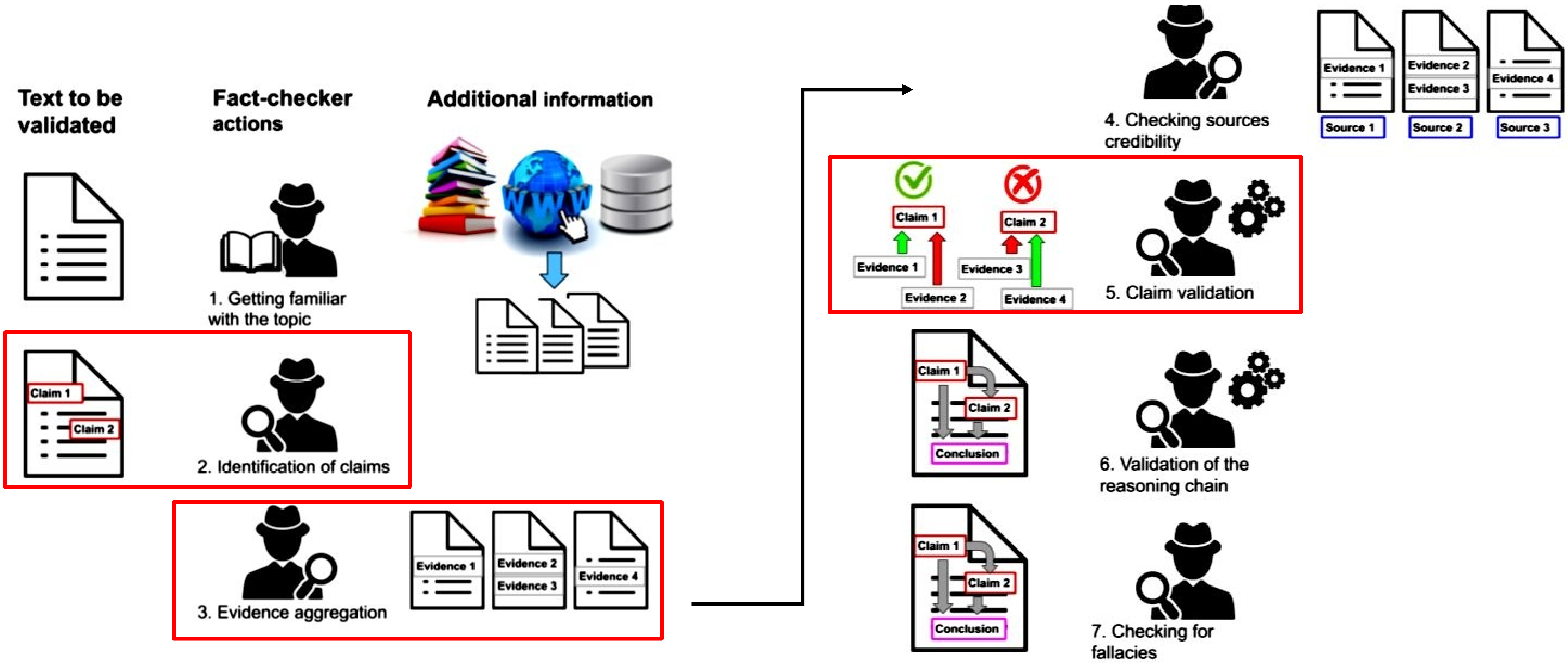
External fact-checking

- External fact-checking (emerged in 2000): the evidence-based analysis of the truthfulness of argumentative claims to publish systematic assessment articles.
  - Fact-checking of claims particularly in political debates, speeches and interviews
  - Precise investigation of assortments of exaggerations, false/misleading notes, and ambiguous factual statements
  - Has also given rise to dedicated fact-checking outlets such as PolitiFact and FactCheck.org

Fake News Detection

- Two primary categories of fake news detection methods:
  - Network-based: rely on social network behavior analysis, particularly on the network formed by interactions between people
  - Content-based: ground in text analysis such as linguistic features, content cues, deception modelling, clustering and classification
- The techniques in automated fact-checking and content-based fake news detection overlap to some extent.

Manual Fact-checking



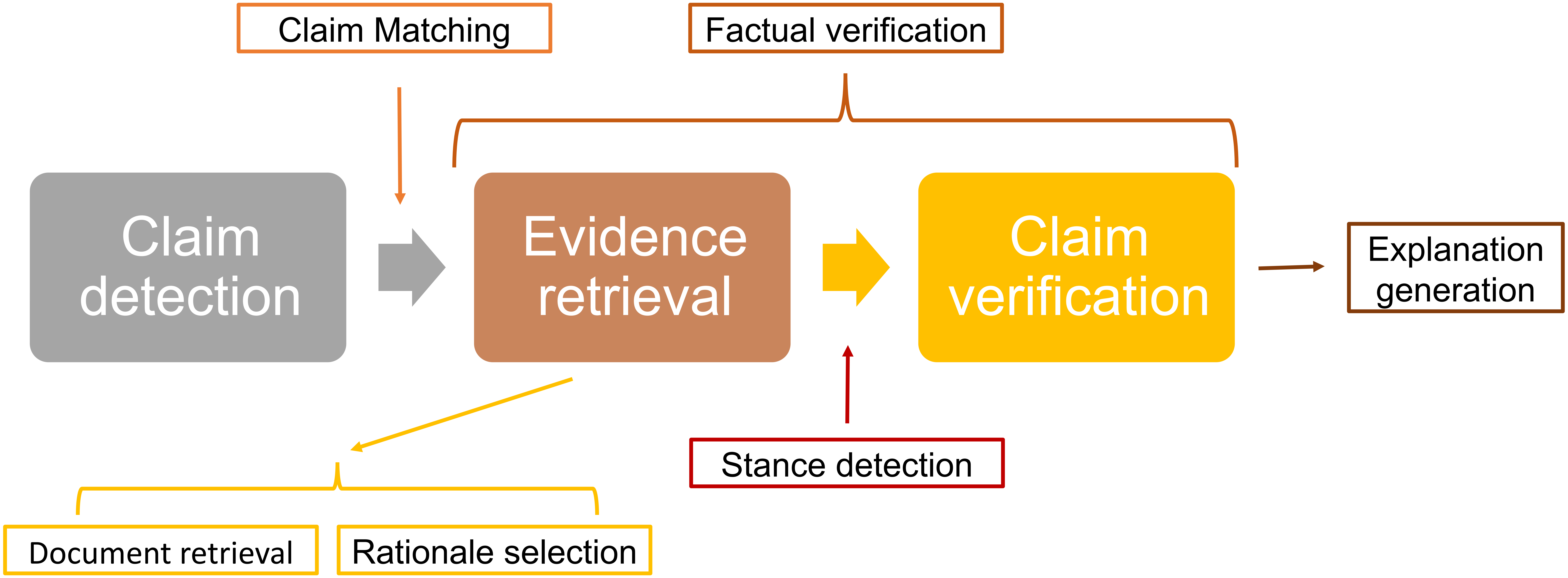
Areas of interest in news industry

- The augmented newsroom
  - ✓ New technology to help journalists work more efficiently
  - ✓ New methods for verification of text information and image/video authenticity
- Trustworthy, secure, transparent, explainable, and unbiased technologies
  - ✓ Technology as a transparent unbiased assistant, not as black boxes
  - ✓ Build trustworthy and secure tools for journalists
- New technology to improve business efficiency and sustainability
  - ✓ Discover new areas of use of AI, ML, semantics, and metadata

Areas of interest in NLP landscape

- Automated (assistance for) fact-checking
  - ✓ A pipeline of fully automated fact-checking
  - ✓ Automated fact-checking with human in the loop
  - ✓ Knowledge enhanced fact-checking
  - ✓ ...
- Fact-checking in NLG
  - ✓ Post-processing of artificially generated text such as in debaters and question answering
  - ✓ Factual error correction for extractive summarization

The Pipeline of Automated Fact-checking



NLP and machine learning methods

- NLP Features:
  - ✓ Name Entity Recognition
  - ✓ Part of Speech Tagging
  - ✓ Dependency Parsing
  - ✓ Word Embedding
  - ✓ Stance Detection
  - ✓ ...
- Neural Language Models:
  - ✓ BiLSTM
  - ✓ BERT and its variations
  - ✓ T5
  - ✓ …
- Traditional ML:
  - ✓ Feature Selection
  - ✓ Classification: SVM, DT, BC
- Knowledge graphs:
  - ✓ K-BERT
  - ✓ Knowledge linker
  - ✓ ClaimKG
- Information Retrieval:
  - ✓ BM25
  - ✓ LM
  - ✓ PL2
  - ✓ ...

Some Useful Python Tools

- **Beautifulsoup4**: a library to scrape information from web pages.
- **Urllib**: a package that collects several modules for working with URLs
- **googlesearch-python**: a library for searching Google using requests and BeautifulSoup4 to scrape Google.
- **nltk**: a suite of libraries and programs for symbolic and statistical NLP
- **SpaCy:** an open-source software python library used in advanced natural language processing and machine learning to build information extraction, natural language understanding systems, and to pre-process text for deep learning
- **Sklearn**: the most useful and robust library for machine learning in Python
- **PyTorch**: an open-source machine learning framework that accelerates the path from research prototyping to production deployment
- **TensorFlow**: a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top of TensorFlow

Claim Detection

- All other components need to rely on the output of this stage.
- It aims to relief the burden of identifying claims for fact-checkers.
- For instance:
  - ✓ "He voted against the first gulf war" can be deemed a claim that should be fact-checked.
  - ✓ "I think it's time to talk about the future" is not a claim that should be fact-checked.
- One can also distinguish between check-worthy vs non-check-worthy claims. For Example:
  - ✓ "the government invested more than 10 billion last year in education" is a claim that is worthy of fact-checking
  - ✓ "my friend had a coffee this morning for breakfast" may not be worthy of fact-checking.
- The problem is formulated as having a set of sentences as input (e.g. originating from a debate or conversation), and is tackled as
  - ✓ a classification task, where a binary decision is made on whether each input sentence constitutes a claim or not
  - ✓ or a ranking task, where input sentences are ranked by check-worthiness, prioritizing top claims on top positions of the list.

Claim Matching

- Claim matching consists in determining whether this is a claim that exists in the database and can be resolved by a previous fact-check.
- The task is formulated as:
  - ✓ given a check-worthy claim as input,
  - ✓ and a database of previously fact-checked claims,
  - ✓ determine if any of the claims in the database is related to the input; in this case, the new claim would not need fact-checking again, as it was fact-checked in the past.
  - ✓ It is normally framed as a ranking task, where claims in the database are ranked based on their similarity to the input claim.
- Two released datasets: one based on PolitiFact and the other based on Snopes.
- Initial explorations using BM25 and BERT-based models respectively.

A ranking function used by search engines to estimate the relevance of documents to a given search query

Evidence Retrieval

- Evidence retrieval is conventionally addressed in two steps:
  - ✓ document retrieval: the task of retrieving relevant documents that supports the prediction of a claim's veracity
  - ✓ rationale selection: the task of selecting directly relevant sentences out of the retrieved documents to get final supporting evidence for claim verification
- Two approaches
  - ✓ To limit evidence to only trusted resource such as Wikipedia, fact-checking websites, peer-reviewed academic papers, and government documents, achieving substantial coverage of information.
  - ✓ To verify the claim against existing knowledge bases, this faces bigger challenges in terms of coverage of reliable information: existing knowledge bases tend to be too small to cover sufficient information for claim validation purposes

Claim Verification

- Claim verification is commonly addressed as a text classification task by NLP researchers:
  - ✓ Given a claim under investigation and its retrieved evidence, models need to reach a verdict of the claim, which may be 'SUPPORT', 'CONTRADICTION' or 'NOT ENOUGH INFORMATION'.
  - ✓ Some other datasets include other labels such as 'mostly-true', 'half-true', 'pants-fire', 'most false', 'most true' and 'other', whose finer granularity is more difficult to tackle through automated means and are sometimes collapsed into fewer labels.
- Claim verification usually includes providing rationale sentences or evidence passages as explanation
  - ✓ A few efforts on generating justification

Examples from Previous Studies

ClaimBuster

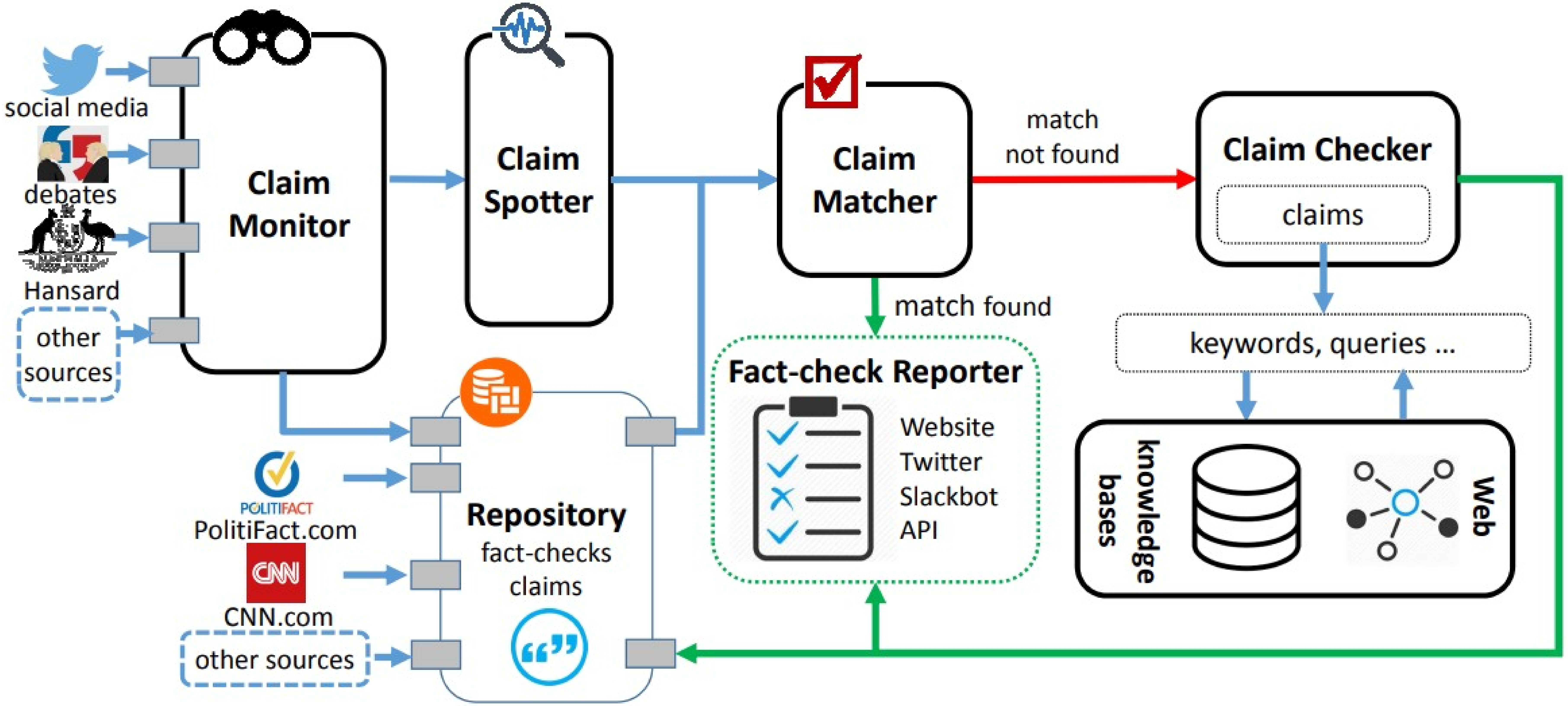
FEVER

CLEF CheckThat!

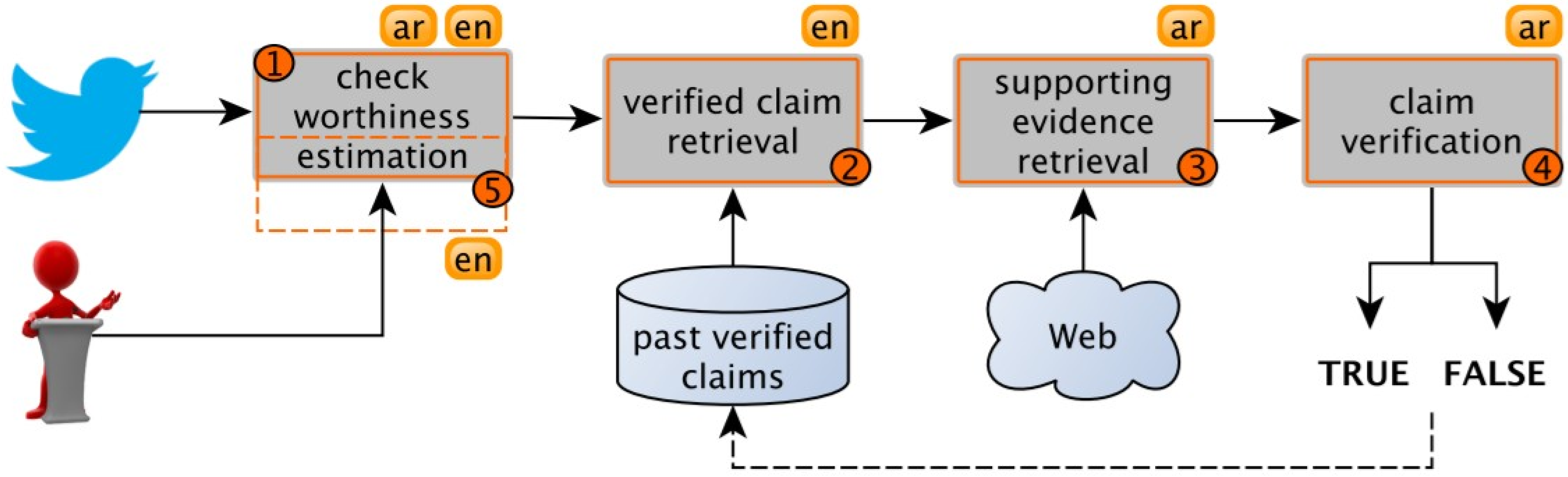
ClaimBuster

- 2015: A team at the University of Texas at Arlington developed the ClaimBuster algorithm to automate the process of finding factual claims in political transcripts.
  - The data was derived from transcripts of U.S. presidential debates from 1960 to 2012.
  - Sentence categorized into three categories: NFS, UFS, and CFS.
  - Proposed system: a set of **lexical, syntactic, and semantic features** --> **feature selection** **-->** **traditional classifiers** (NB, SVM and RF)
- 2016: Tested in real-time during the live coverage of all primary and general debates throughout the 2016 U.S. election.
  - Post-hoc analysis of the claims checked by professional fact-checkers at CNN, PolitiFact.com, and FactCheck.org reveals a highly positive correlation in deciding which claims to check.

2017: ClaimBuster with Expanded Features

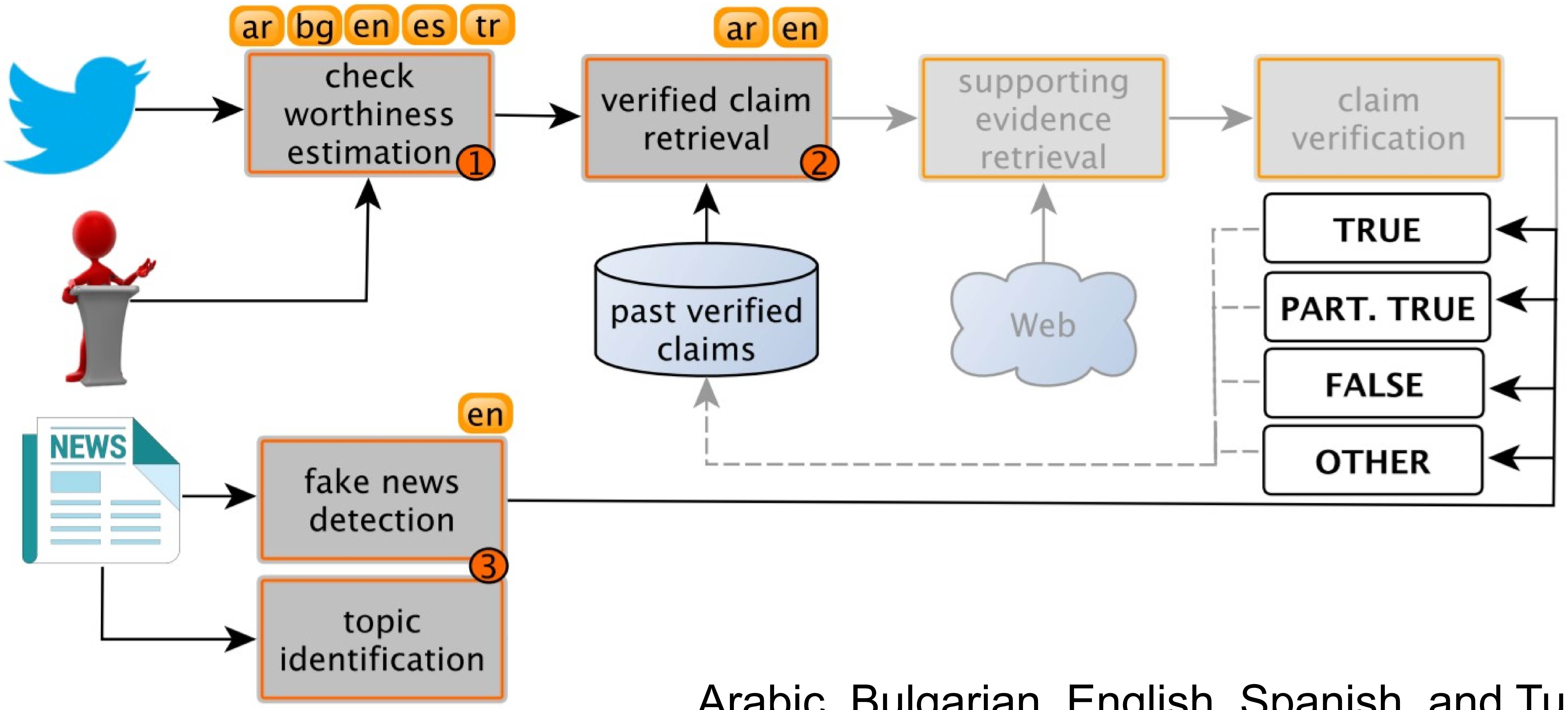


CLEF-2020 CheckThat! Lab



Task 5 complements the lab. It is as Task 1, but on political debates ad speeches rather than on tweets: given a debate segmented into sentences, together with speaker information, prioritize sentences for fact-checking.

CLEF-2021 CheckThat! Lab



Arabic, Bulgarian, English, Spanish, and Turkish

CLEF-2022 CheckThat! Lab

The CheckThat! lab aims at fighting misinformation and disinformation in social media, in political debates and in the news, with focus on three tasks (in seven languages: Arabic, Bulgarian, Dutch, English, German, Spanish, and Turkish).

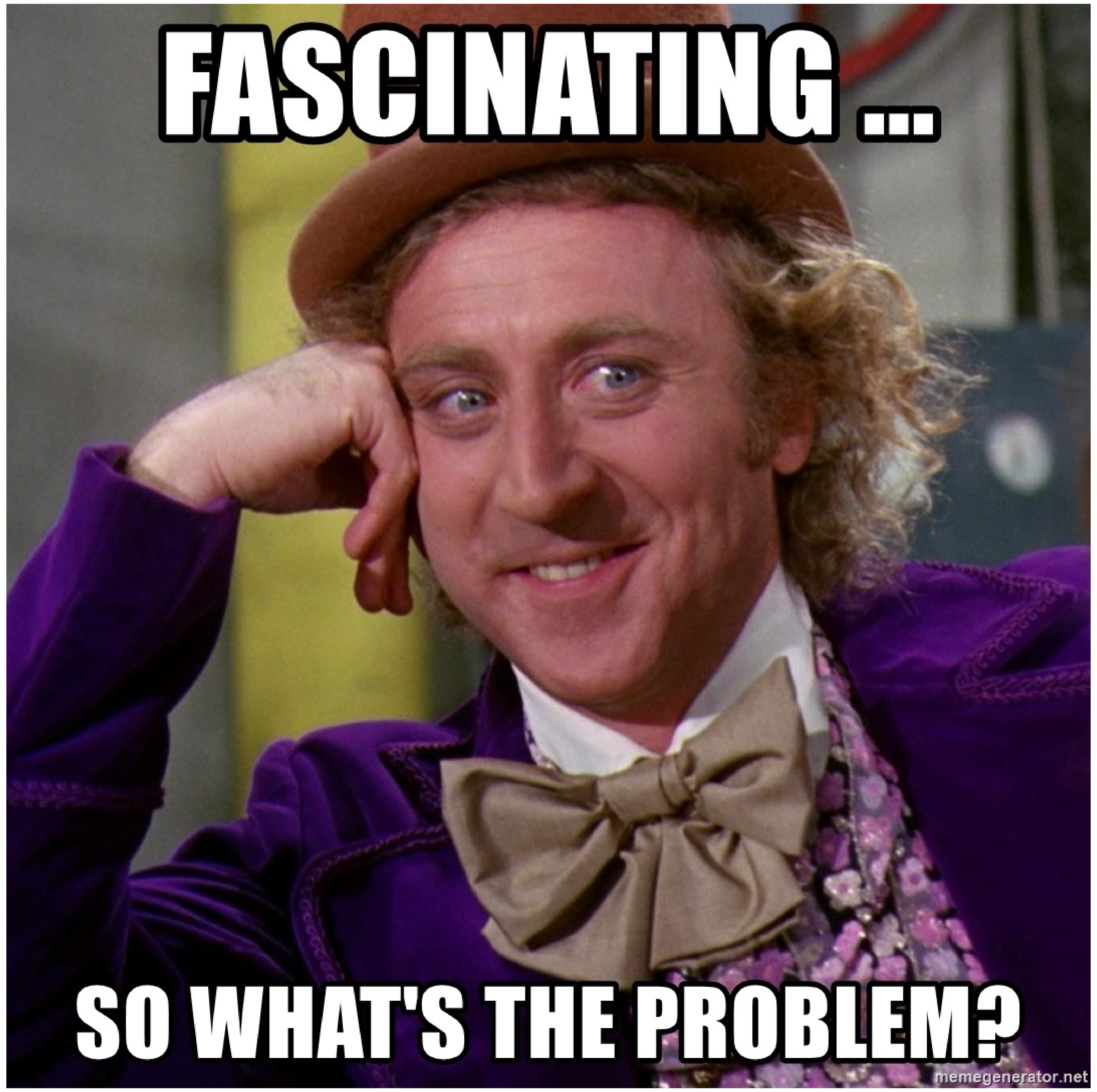
- **Task 1:** Fighting the COVID-19 infodemic
- **Task 2:** Detecting previously fact-checked claims
- **Task 3:** Fake news detection

https://sites.google.com/view/clef2022-checkthat

2018: FEVER

- Contains 185,445 human-generated claims labeled as SUPPORTED, REFUTED or NOTENOUGHINFO.
- Generated by paraphrasing facts from Wikipedia and mutating them in a variety of ways.
- For each claim annotators selected evidence in the form of sentences from Wikipedia.

- FEVER shared task: <u>label claims with the correct class and return the sentence(s) forming the necessary evidence for the assigned label</u>.



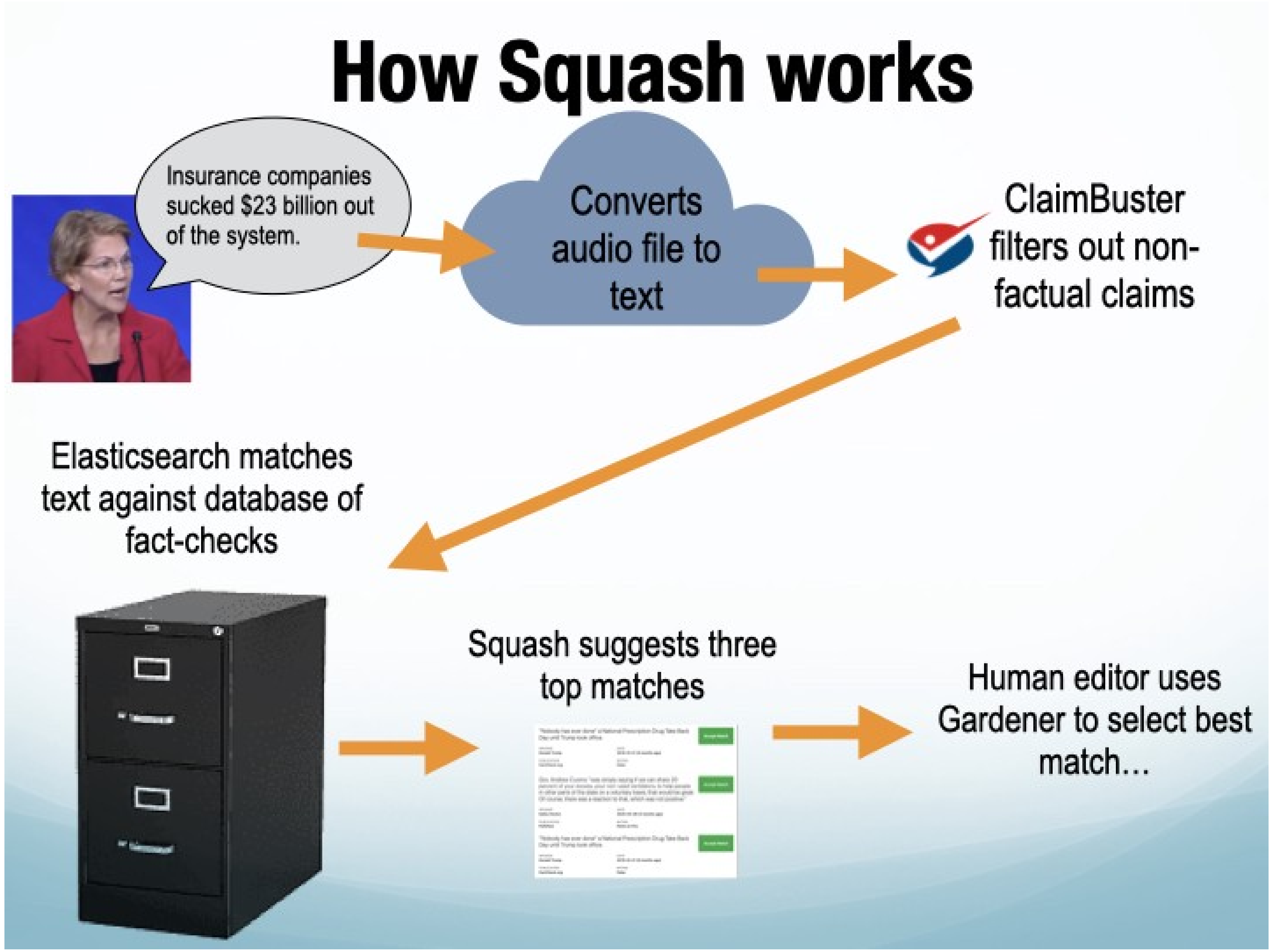
Performance

CLEF-2021 CheckThat!

Task	MAP*
Check-Worthiness of Tweets	0.224
Check-Worthiness of Debates/Speeches	0.402
Detecting Previously Fact-Checked Claims in Tweets	0.883
Detecting Previously Fact-Checked Claims in Political Debates and Speeches	0.346
Multi-Class Fake News Detection of News Articles	0.853**
Topical Domain Identification of News Articles	0.905**

*Mean Average Precision
**Accuracy

Duke’s automated **fact-checking platform**



https://www.poynter.org/fact-checking/2021/the-lessons-of-squash-the-first-automated-fact-checking-platform/

Challenges Ahead

1. Data:
   - Most data sets are in English
   - High quality annotated data of naturally occurring claims is scarce
   - Data sets are usually biased
2. Claim difficulty
   - Claims have vague and diverse conceptualization.
   - Ambiguity is a natural obstacle.
   - Given the appropriate evidence, natural language inference could be difficult .
   - Some claims require a multi-hop reasoning chain which is difficult to be automated
3. Evidence
   - Previously checked claims are not always the solution
   - Retrieving evidence in the wild is difficult (an understudied task)
   - Trustworthiness: High quality sources such as established news outlets, accredited journalism, scientific research articles are unavailable for many claims
4. Explainability
5. Keeping human in the loop

Media Futures

Thank you

for your attention

Contact information:

ghazaal.sheikhi@uib.no

Research Centre for Responsible Media Technology & Innovation
Project number 309339