A Survey on Automated Fact-Checking

Zhijiang Guo*, **Michael Schlichtkrull***, Andreas Vlachos

Introduction

- Fact-checking is the task of assessing and arguing for the **factuality** of **claims** made in written or spoken language.

Introduction

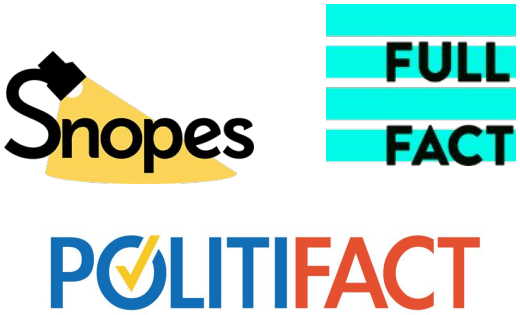
- Fact-checking is the task of assessing and arguing for the **factuality** of **claims** made in written or spoken language.
- Covers two essential tasks in journalism:
  - **External fact-checking**
  - **Internal fact-checking**

Introduction

- Fact-checking is the task of assessing and arguing for the **factuality** of **claims** made in written or spoken language.
- Covers two essential tasks in journalism:
  - **External fact-checking**
  - **Internal fact-checking**

Introduction

- Fact-checking is the task of assessing and arguing for the **factuality** of **claims** made in written or spoken language.
- Covers two essential tasks in journalism:
  - **External fact-checking**
  - **Internal fact-checking**



Introduction

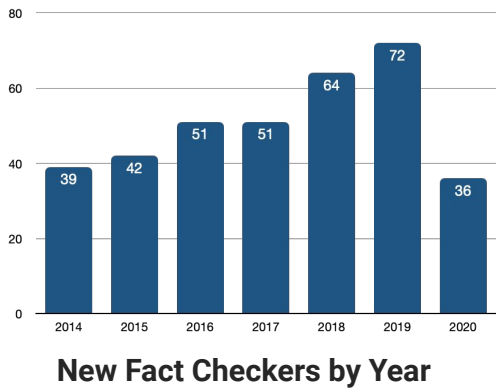
- In the last few years, we have seen **misinformation** play a large role in major elections, as well as in ongoing crises.

Introduction

- In the last few years, we have seen **misinformation** play a large role in major elections, as well as in ongoing crises.
- The amount of articles subject to internal fact-checking has **decreased**, because of stricter deadlines.

Introduction

- In the last few years, we have seen **misinformation** play a large role in major elections, as well as in ongoing crises.
- The amount of articles subject to internal fact-checking has **decreased**, because of stricter deadlines.
- The amount and quality of external fact-checkers has **increased**.



Introduction

- NLP can play a major role in addressing the challenge:
  - **Searching** large collections of documents for evidence, making the task easier for journalists and accessible for lay people.

Introduction

- NLP can play a major role in addressing the challenge:
  - **Searching** large collections of documents for evidence, making the task easier for journalists and accessible for lay people.
  - **Triaging** claims to identify the highest priority targets for professionals to debunk.

Introduction

- NLP can play a major role in addressing the challenge:
  - **Searching** large collections of documents for evidence, making the task easier for journalists and accessible for lay people.
  - **Triaging** claims to identify the highest priority targets for professionals to debunk.
  - **Spotting connections** between distant pieces of evidence.

Introduction

- NLP can play a major role in addressing the challenge:
  - **Searching** large collections of documents for evidence, making the task easier for journalists and accessible for lay people.
  - **Triaging** claims to identify the highest priority targets for professionals to debunk.
  - **Spotting connections** between distant pieces of evidence.
  - Identifying previously fact-checked claims to **intervene** where necessary, and to **prevent duplicate work**.

Introduction

- We wanted to create a resource that…
  - Is **up-to-date** on recent developments, including the production of justifications.

Introduction

- We wanted to create a resource that…
  - Is **up-to-date** on recent developments, including the production of justifications.
  - Presents a **unified framework** for the topic.

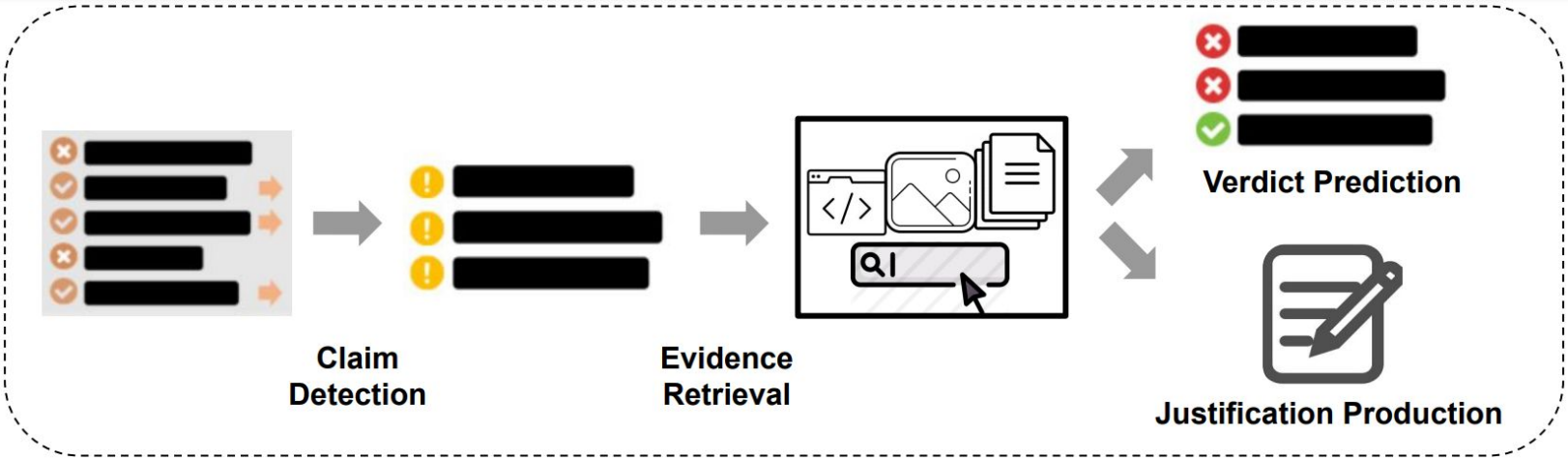
Introduction

- We wanted to create a resource that…
  - Is **up-to-date** on recent developments, including the production of justifications.
  - Presents a **unified framework** for the topic.
  - Documents and compares existing **datasets and models** across different approaches

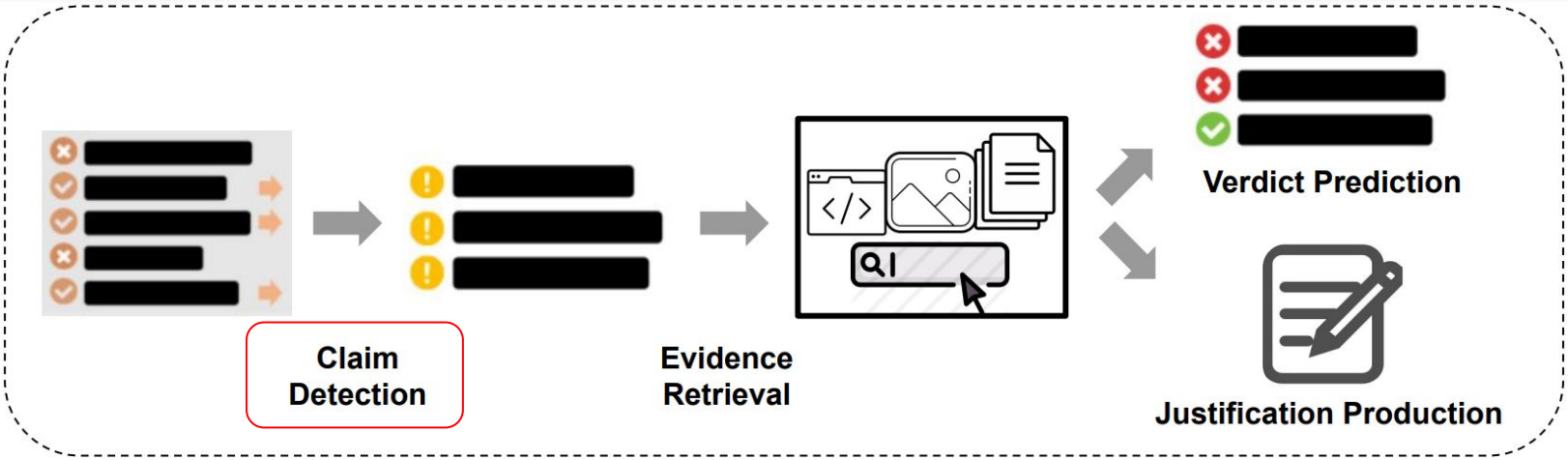
Introduction

- We wanted to create a resource that…
  - Is **up-to-date** on recent developments, including the production of justifications.
  - Presents a **unified framework** for the topic.
  - Documents and compares existing **datasets and models** across different approaches
  - Anticipates **future challenges**.

Task Definition



Claim Detection



Claim Detection

- Claim Detection aims to ***select claims for verification***. Commonly, detection relies on the concept of ***check-worthiness.***
- Check-worthy claims: those for which the ***general public would be interested in knowing the truth***. (Hassan et al. 2015)

"Over six million Americans had COVID-19 in January."

"Water is wet."

Hassan et al. 2015, Detecting Check-worthy Factual Claims in Presidential Debates. *CIKM 2021*

Claim Detection

- Check-worthiness is ***subjective.***
- Check-worthiness ***varies over time***, as countering misinformation related to current events is understood to be ***more important than countering older misinformation.***

Claim Detection

- Check-worthiness is ***subjective.***
- Check-worthiness ***varies over time***, as countering misinformation related to current events is understood to be ***more important than countering older misinformation.***
- Journalists have been shown to assign greater trust and therefore ***lower need for verification*** to stories produced by ***male sources*** (Barnoy and Reich, 2019).

Barnoy and Reich 2019, The When, Why, How and So-What of Verifications. *Journalism Studies 2019*

Claim Detection

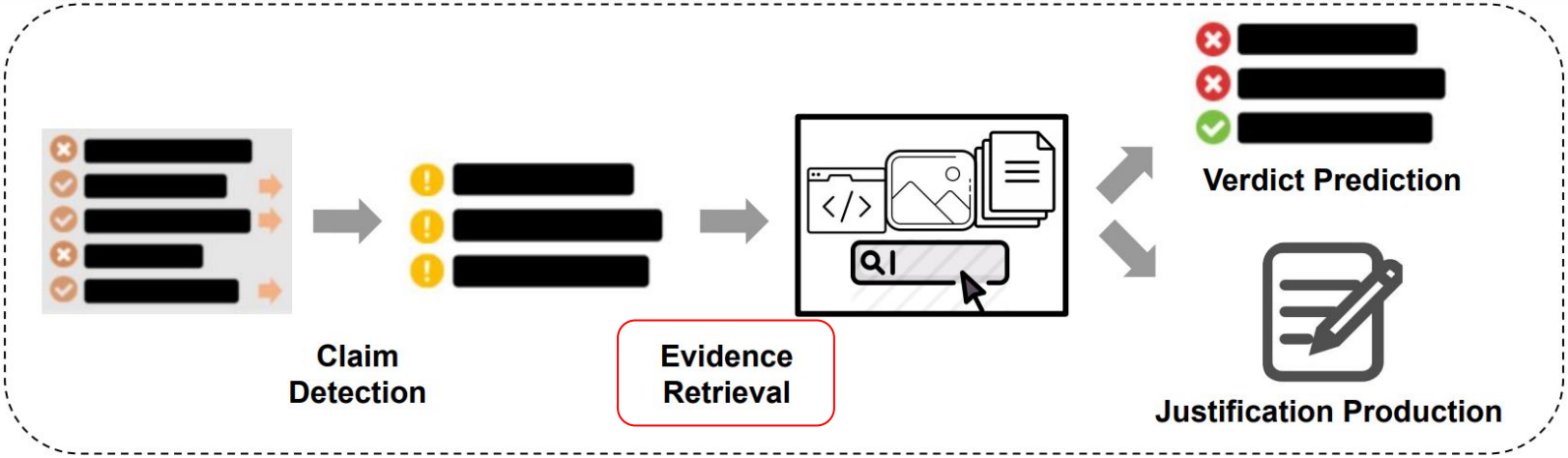
- Konstantinovskiy et al. (2021) framed claim detection as whether a claim makes an assertion about the world that is ***checkable***, i.e. ***whether it is verifiable with readily available evidence.***

Trump suggests "injecting disinfectant as treatment".

"I woke up at 7 am today"

Konstantinovskiy et al., 2021. Toward automated fact checking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice 2021*

Evidence Retrieval



Evidence Retrieval

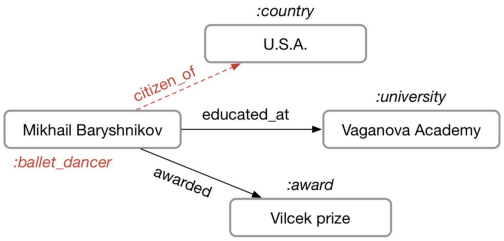
Evidence retrieval aims to find information ***beyond the claim*** to ***indicate veracity***:

- Text
- Tables
- Knowledge bases
- Images
- Videos
- Relevant metadata

Evidence Retrieval

Evidence retrieval aims to find information ***beyond the claim*** to ***indicate veracity***:

- Text
- Tables
- Knowledge bases
- Images
- Videos
- Relevant metadata



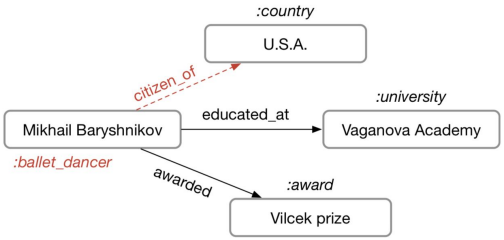
United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Evidence Retrieval

Evidence retrieval aims to find information ***beyond the claim*** to ***indicate veracity***:

- Text
- Tables
- Knowledge bases
- Images
- Videos
- Relevant metadata



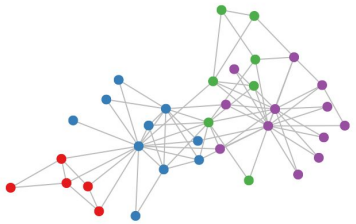
United States House of Representatives Elections, 1972

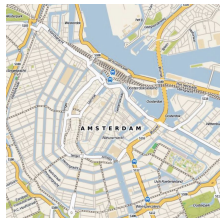
District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

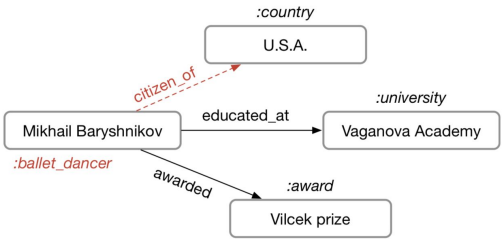
Evidence Retrieval

Evidence retrieval aims to find information ***beyond the claim*** to ***indicate veracity***:

- Text
- Tables
- Knowledge bases
- Images
- Videos
- Relevant metadata



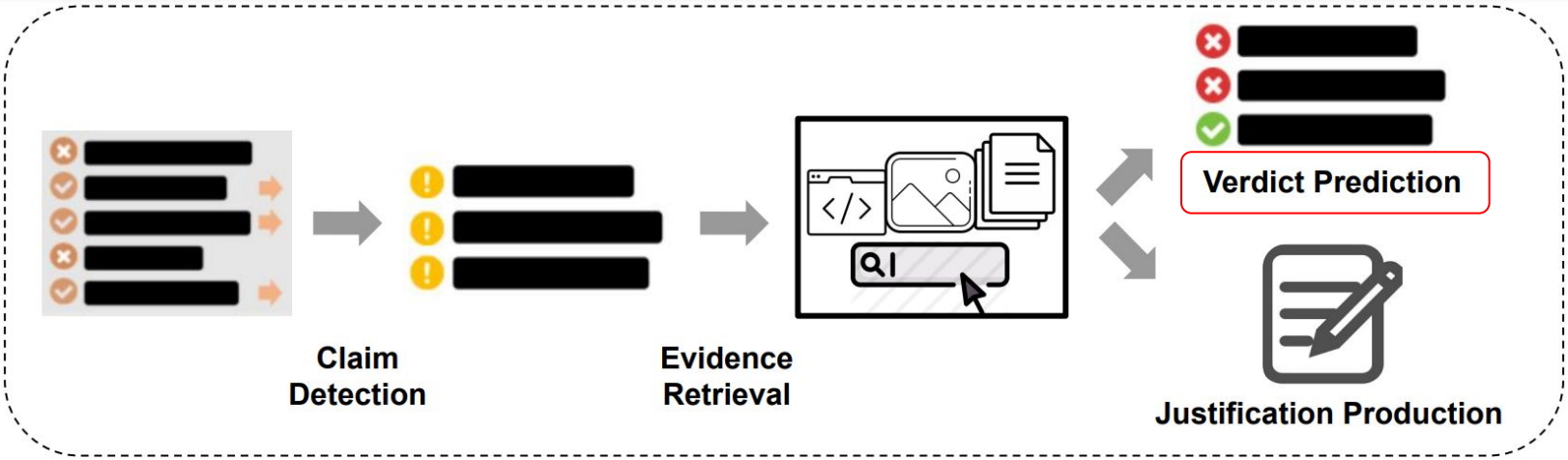




United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Verdict Prediction

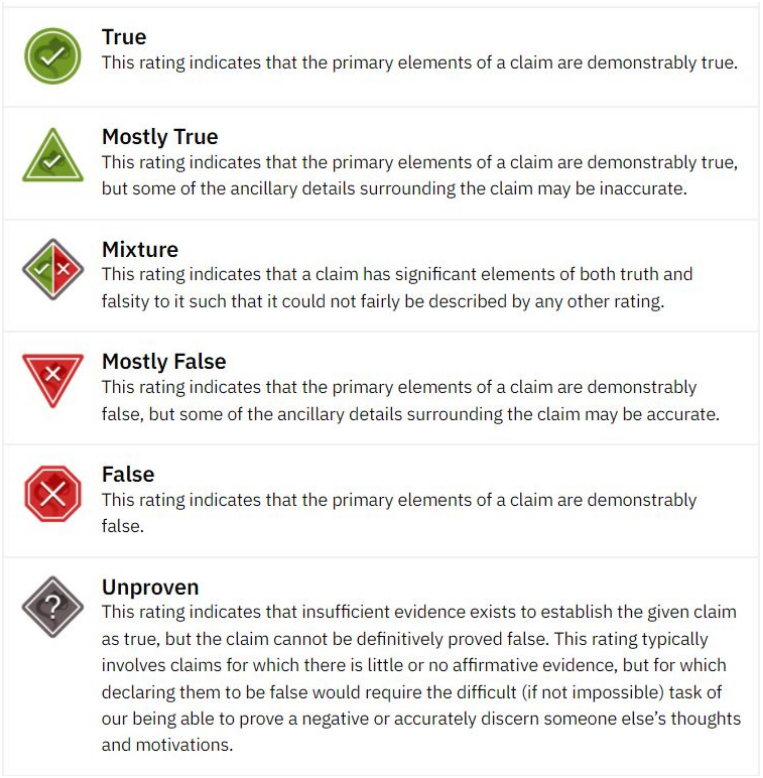


Verdict Prediction

Given an ***identified claim*** and the ***pieces of evidence*** retrieved for it, verdict prediction attempts to determine the ***degree of support*** for the claim expressed in the evidence.

- Binary classification (Supported/Refuted)
- Ternary classification (Supported/Refuted/Not Enough Information)
- Finer-grained classification

Verdict Prediction



True
This rating indicates that the primary elements of a claim are demonstrably true.

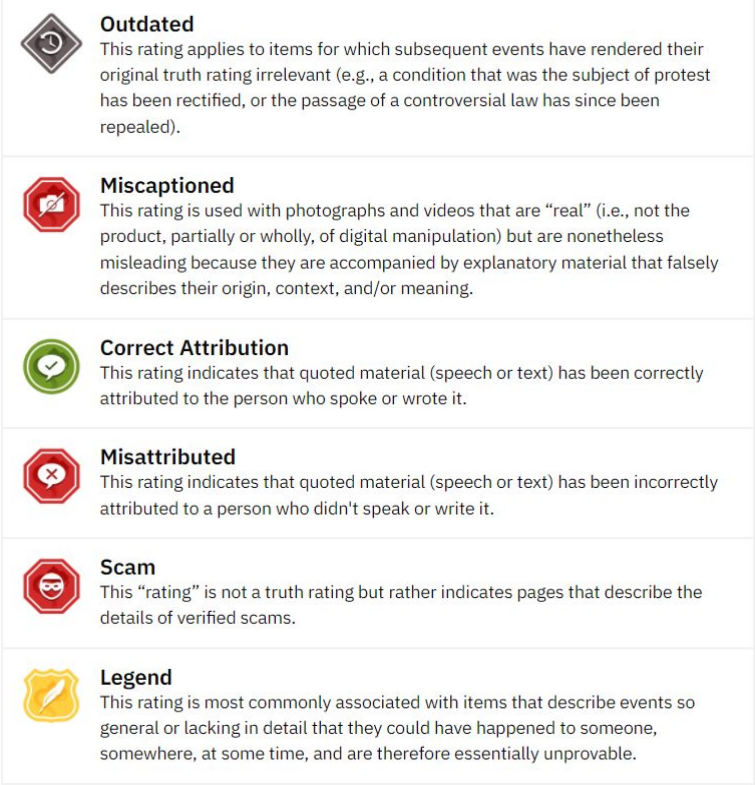
Mostly True
This rating indicates that the primary elements of a claim are demonstrably true, but some of the ancillary details surrounding the claim may be inaccurate.

Mixture
This rating indicates that a claim has significant elements of both truth and falsity to it such that it could not fairly be described by any other rating.

Mostly False
This rating indicates that the primary elements of a claim are demonstrably false, but some of the ancillary details surrounding the claim may be accurate.

False
This rating indicates that the primary elements of a claim are demonstrably false.

Unproven
This rating indicates that insufficient evidence exists to establish the given claim as true, but the claim cannot be definitively proved false. This rating typically involves claims for which there is little or no affirmative evidence, but for which declaring them to be false would require the difficult (if not impossible) task of our being able to prove a negative or accurately discern someone else's thoughts and motivations.



Outdated
This rating applies to items for which subsequent events have rendered their original truth rating irrelevant (e.g., a condition that was the subject of protest has been rectified, or the passage of a controversial law has since been repealed).

Miscaptioned
This rating is used with photographs and videos that are "real" (i.e., not the product, partially or wholly, of digital manipulation) but are nonetheless misleading because they are accompanied by explanatory material that falsely describes their origin, context, and/or meaning.

Correct Attribution
This rating indicates that quoted material (speech or text) has been correctly attributed to the person who spoke or wrote it.

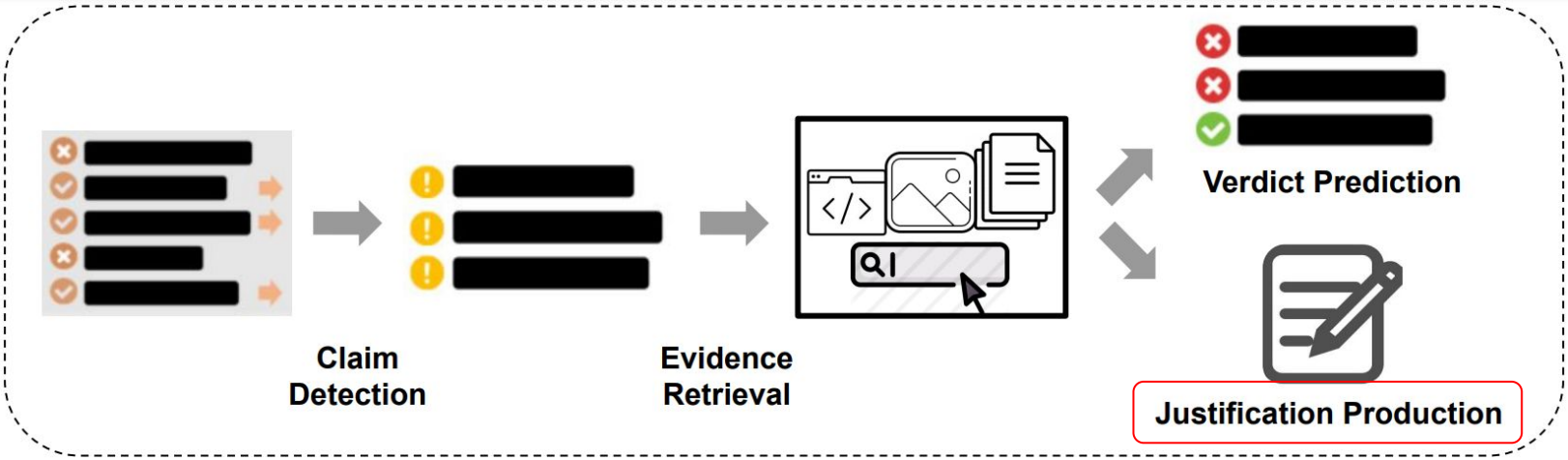
Misattributed
This rating indicates that quoted material (speech or text) has been incorrectly attributed to a person who didn't speak or write it.

Scam
This "rating" is not a truth rating but rather indicates pages that describe the details of verified scams.

Legend
This rating is most commonly associated with items that describe events so general or lacking in detail that they could have happened to someone, somewhere, at some time, and are therefore essentially unprovable.

https://www.snopes.com/fact-check-ratings/

Justification Production



Justification Production

- “The SS Cotopaxi could not have drifted whole out of the Bermuda Triangle in 2015, because the wreck is still lying off the coast of Florida.”

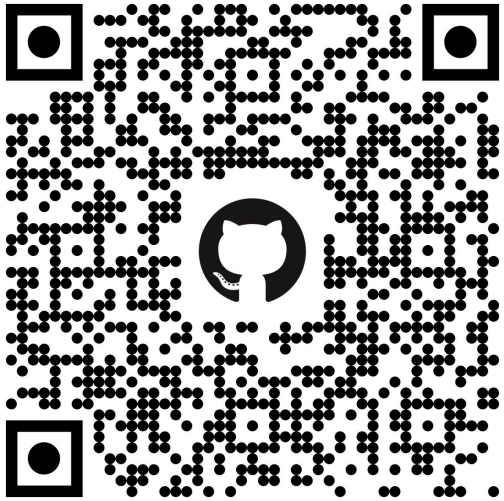
Justification Production

- "The SS Cotopaxi could not have drifted whole out of the Bermuda Triangle in 2015, because the wreck is still lying off the coast of Florida."

- Criteria for justification production:
  - ***Readability*** (how accessible an explanation is to humans)
  - ***Plausibility*** (how convincing an explanation is)
  - ***Faithfulness*** (how accurately an explanation reflects internal reasoning)

Datasets & Models

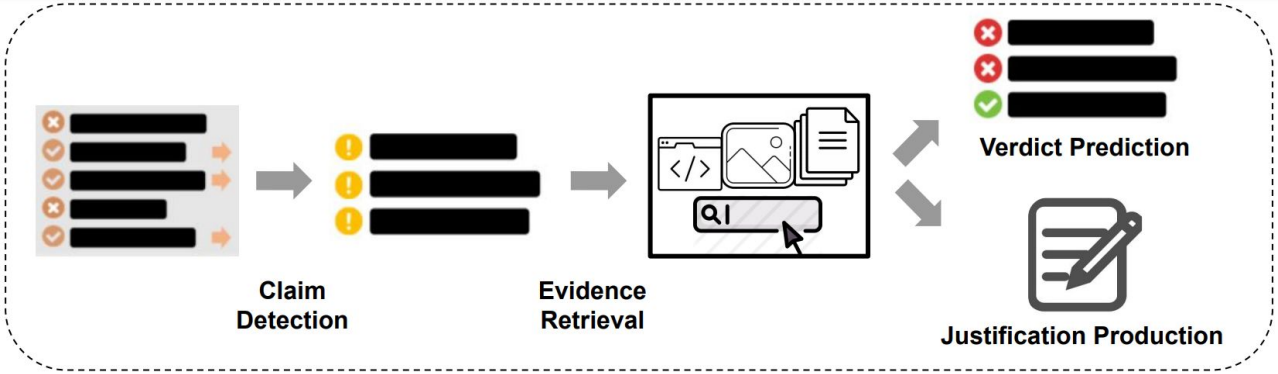
- We analyse and compare datasets and models for all four components of our framework.
- We include work that targets the entire pipeline, and work that targets just one part of it.



https://github.com/Cartus/Automated-Fact-Checking-Resources

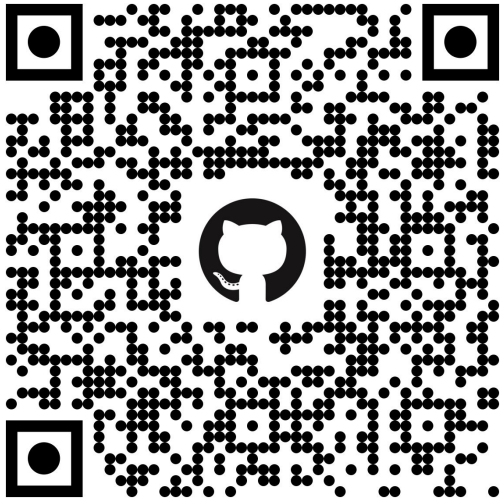
Research Challenges

- Epistemology (binary labelling simplistic, untrustworthy or contradictory sources an inevitability)
- Practice (multilinguality, multimodality, faithfulness)
- Application (Prebunking instead of debunking)



Conclusion

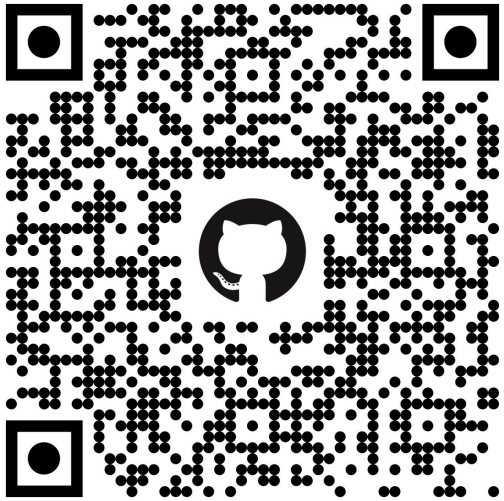
- We analyse fact-checking as a four-step framework: select claims, gather evidence, give verdicts, and produce justifications.
- Check out our repository if you are looking for resources, or send us a pull request if your work is missing!



https://github.com/Cartus/Automated-Fact-Checking-Resources

Conclusion

- We analyse fact-checking as a four-step framework: select claims, gather evidence, give verdicts, and produce justifications.
- Check out our repository if you are looking for resources, or send us a pull request if your work is missing!



https://github.com/Cartus/Automated-Fact-Checking-Resources

FEVER