Accountable and Robust Automatic Fact Checking

Isabelle Augenstein
augenstein@di.ku.dk
@IAugenstein
http://isabelleaugenstein.github.io/

ECIR

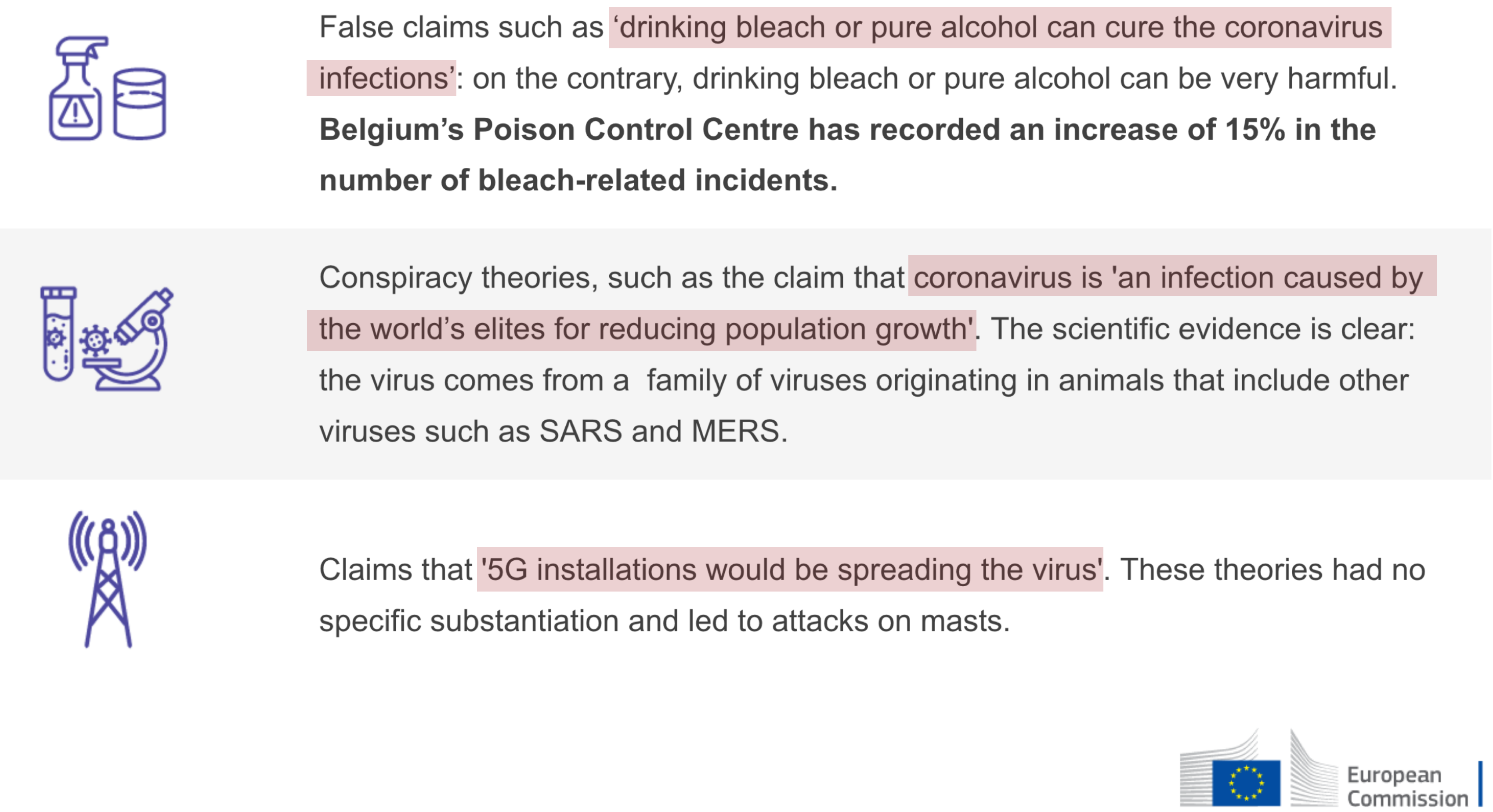
11 April 2022

COPENLU
NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP

UNIVERSITY OF COPENHAGEN

Fact Checking

False Information Online



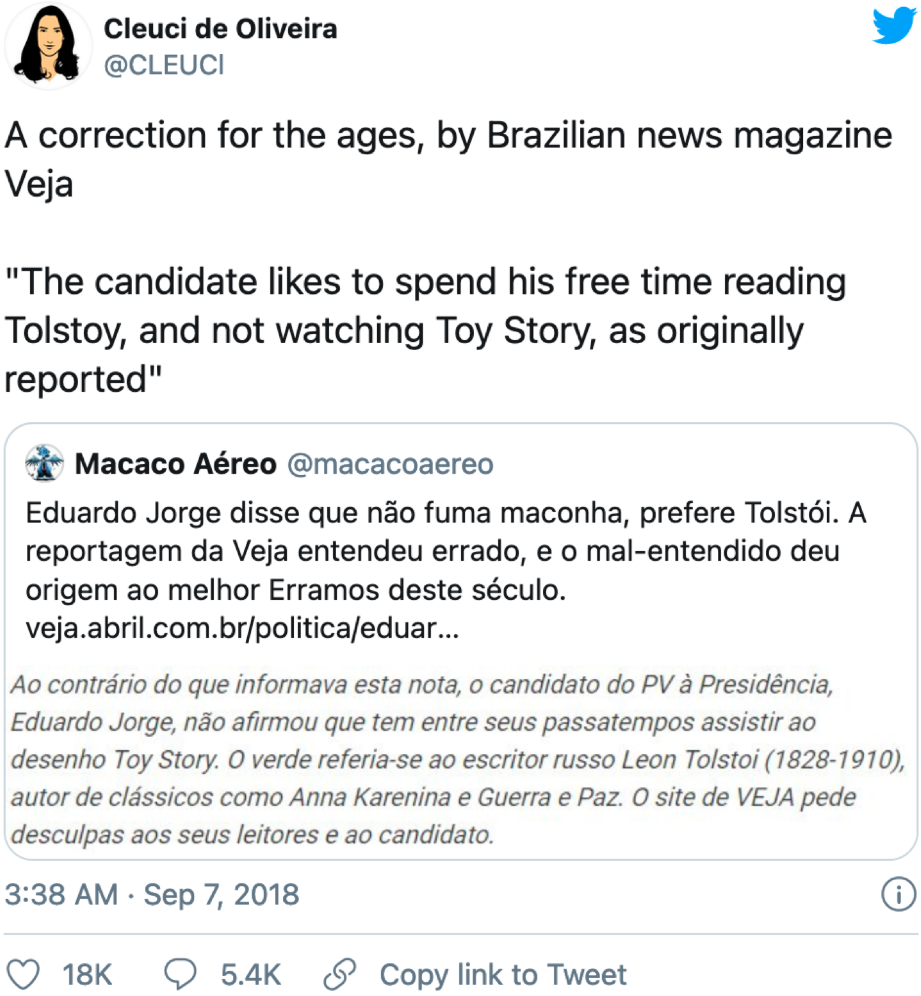
False claims such as 'drinking bleach or pure alcohol can cure the coronavirus infections': on the contrary, drinking bleach or pure alcohol can be very harmful. **Belgium's Poison Control Centre has recorded an increase of 15% in the number of bleach-related incidents.**

Conspiracy theories, such as the claim that coronavirus is 'an infection caused by the world's elites for reducing population growth'. The scientific evidence is clear: the virus comes from a family of viruses originating in animals that include other viruses such as SARS and MERS.

Claims that '5G installations would be spreading the virus'. These theories had no specific substantiation and led to attacks on masts.

https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation/tackling-coronavirus-disinformation_en

Types of False Information



Cleuci de Oliveira
@CLEUCI

A correction for the ages, by Brazilian news magazine Veja

"The candidate likes to spend his free time reading Tolstoy, and not watching Toy Story, as originally reported"

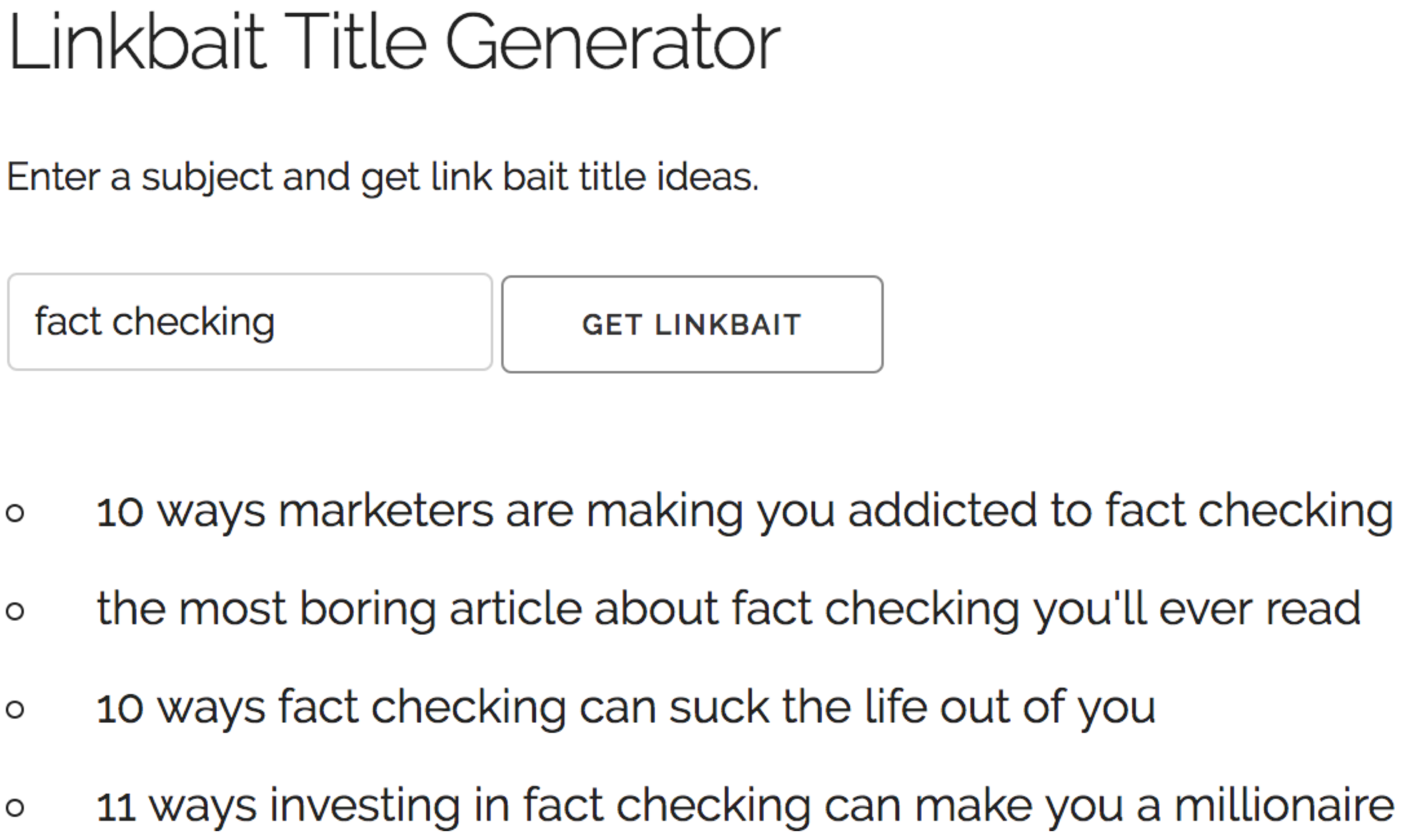
Macaco Aéreo @macacoaereo

Eduardo Jorge disse que não fuma maconha, prefere Tolstói. A reportagem da Veja entendeu errado, e o mal-entendido deu origem ao melhor Erramos deste século. veja.abril.com.br/politica/eduar...

Ao contrário do que informava esta nota, o candidato do PV à Presidência, Eduardo Jorge, não afirmou que tem entre seus passatempos assistir ao desenho Toy Story. O verde referia-se ao escritor russo Leon Tolstoi (1828-1910), autor de clássicos como Anna Karenina e Guerra e Paz. O site de VEJA pede desculpas aos seus leitores e ao candidato.

3:38 AM · Sep 7, 2018

Types of False Information



Linkbait Title Generator

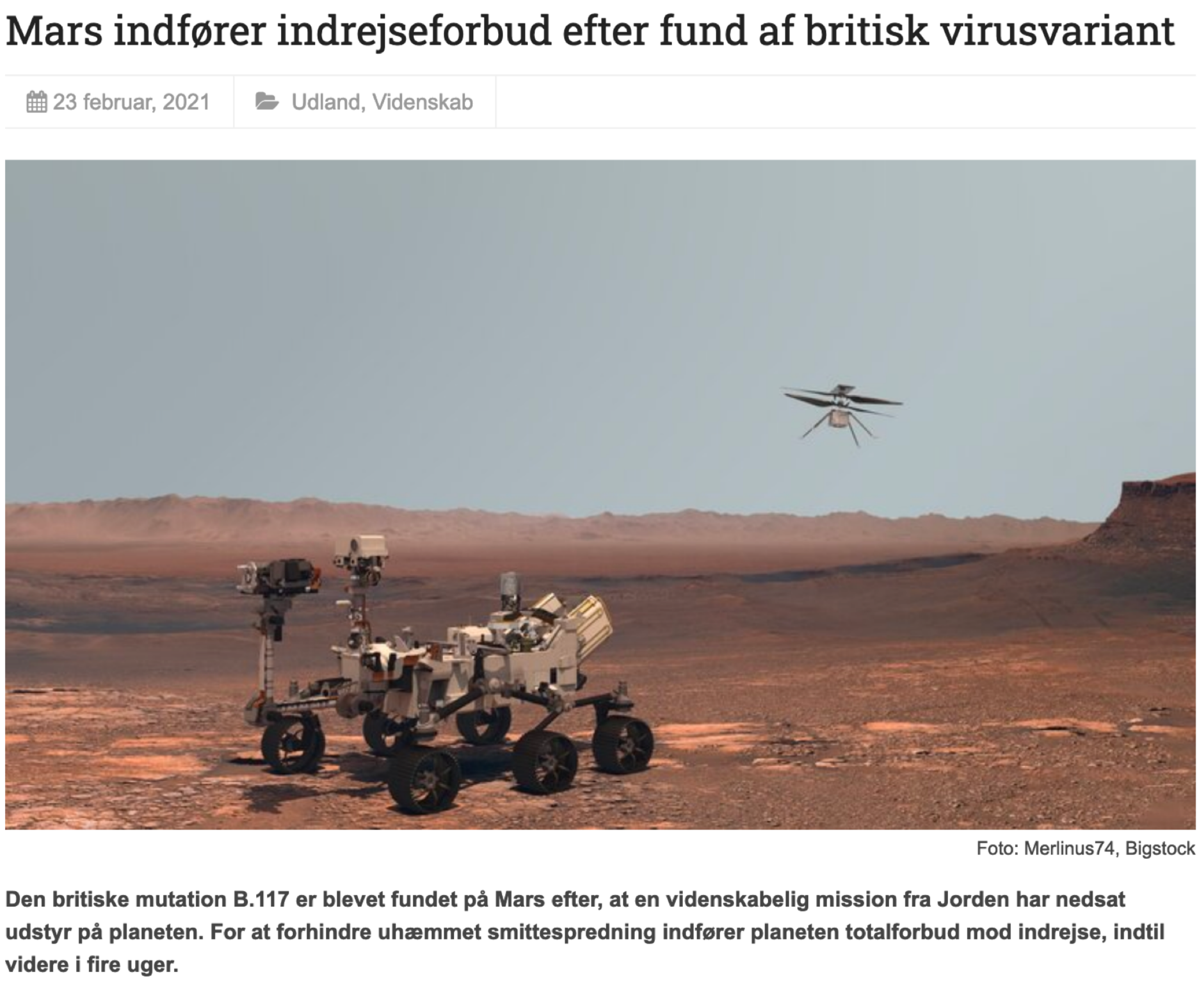
Enter a subject and get link bait title ideas.

fact checking

- 10 ways marketers are making you addicted to fact checking
- the most boring article about fact checking you'll ever read
- 10 ways fact checking can suck the life out of you
- 11 ways investing in fact checking can make you a millionaire

http://www.contentrow.com/tools/link-bait-title-generator

Types of False Information



Mars indfører indrejseforbud efter fund af britisk virusvariant

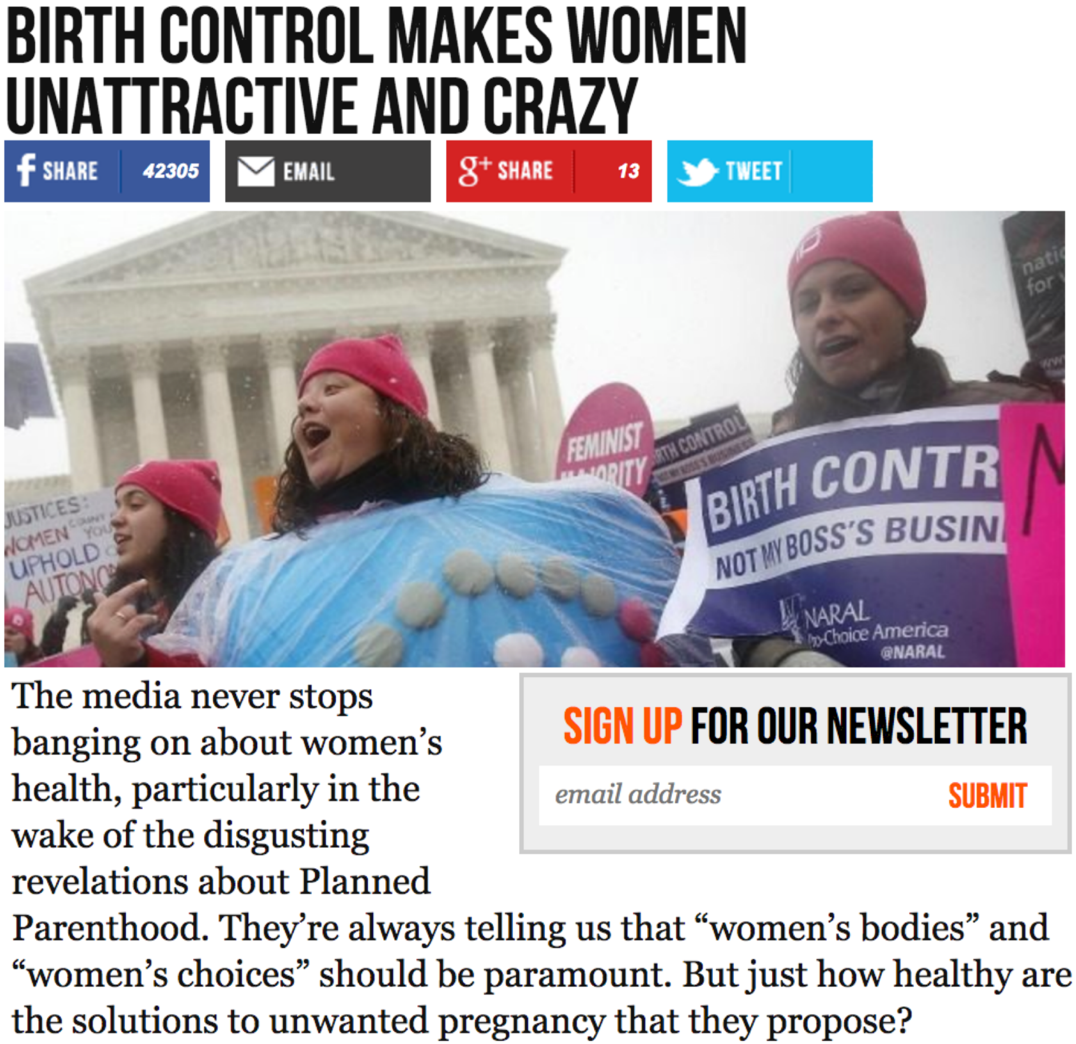
23 februar, 2021 | Udland, Videnskab

Foto: Merlinus74, Bigstock

Den britiske mutation B.117 er blevet fundet på Mars efter, at en videnskabelig mission fra Jorden har nedsat udstyr på planeten. For at forhindre uhæmmet smittespredning indfører planeten totalforbud mod indrejse, indtil videre i fire uger.

https://rokokoposten.dk/2021/02/23/mars-indfoerer-indrejseforbud-efter-fund-af-britisk-virusvariant/

Types of False Information



BIRTH CONTROL MAKES WOMEN UNATTRACTIVE AND CRAZY

SHARE 42305
EMAIL
SHARE 13
TWEET
JUSTICES: WOMEN CAN'T ... YOU UPHOLD ... AUTONO...
FEMINIST MAJORITY
BIRTH CONTROL NOT MY BOSS'S BUSINESS
BIRTH CONTR
NOT MY BOSS'S BUSIN
NARAL Pro-Choice America @NARAL
SIGN UP FOR OUR NEWSLETTER
email address
SUBMIT

The media never stops banging on about women's health, particularly in the wake of the disgusting revelations about Planned Parenthood. They're always telling us that "women's bodies" and "women's choices" should be paramount. But just how healthy are the solutions to unwanted pregnancy that they propose?

BREITBART

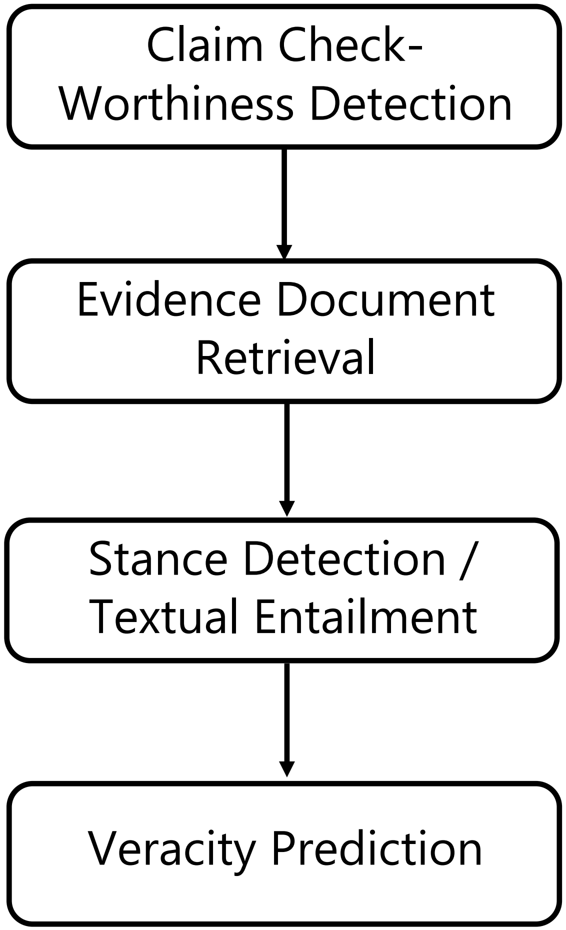
Types of False Information

- Disinformation:
  - Intentionally false, spread deliberately
- Misinformation:
  - Unintentionally false information
- Clickbait:
  - Exaggerating information and under-delivering it
- Satire:
  - Intentionally false for humorous purposes
- Biased Reporting:
  - Reporting only some of the facts to serve an agenda

Types of False Information

- **Disinformation:**
  - **Intentionally false, spread deliberately**
- **Misinformation:**
  - **Unintentionally false information**
- Clickbait:
  - Exaggerating information and under-delivering it
- Satire:
  - Intentionally false for humorous purposes
- Biased Reporting:
  - Reporting only some of the facts to serve an agenda

Full Fact Checking Pipeline



Claim Check-Worthiness Detection

Evidence Document Retrieval

Stance Detection / Textual Entailment

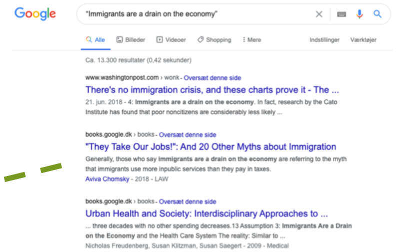
Veracity Prediction

“Immigrants are a drain on the economy”

not check-worthy

check-worthy

“Immigrants are a drain on the economy”



Ca. 13.300 resultater (0,42 sekunder)

www.washingtonpost.com › wonk - Oversæt denne side

There's no immigration crisis, and these charts prove it - The ...

21. jun. 2018 - 4: **Immigrants are a drain on the economy**. In fact, research by the Cato Institute has found that poor noncitizens are considerably less likely ...

books.google.dk › books - Oversæt denne side

"They Take Our Jobs!": And 20 Other Myths about Immigration

Generally, those who say **Immigrants are a drain on the economy** are referring to the myth that immigrants use more inpublic services than they pay in taxes.

Aviva Chomsky - 2018 - LAW

books.google.dk › books - Oversæt denne side

Urban Health and Society: Interdisciplinary Approaches to ...

... three decades with no other spending decreases.13 Assumption 3: **Immigrants Are a Drain on the Economy** and the Health Care System The reality: Similar to ...

Nicholas Freudenberg, Susan Klitzman, Susan Saegert - 2009 - Medical

“Immigrants are a drain on the economy”, “EU immigrants have a positive impact on public finances”

positive

neutral

negative

“Immigrants are a drain on the economy”

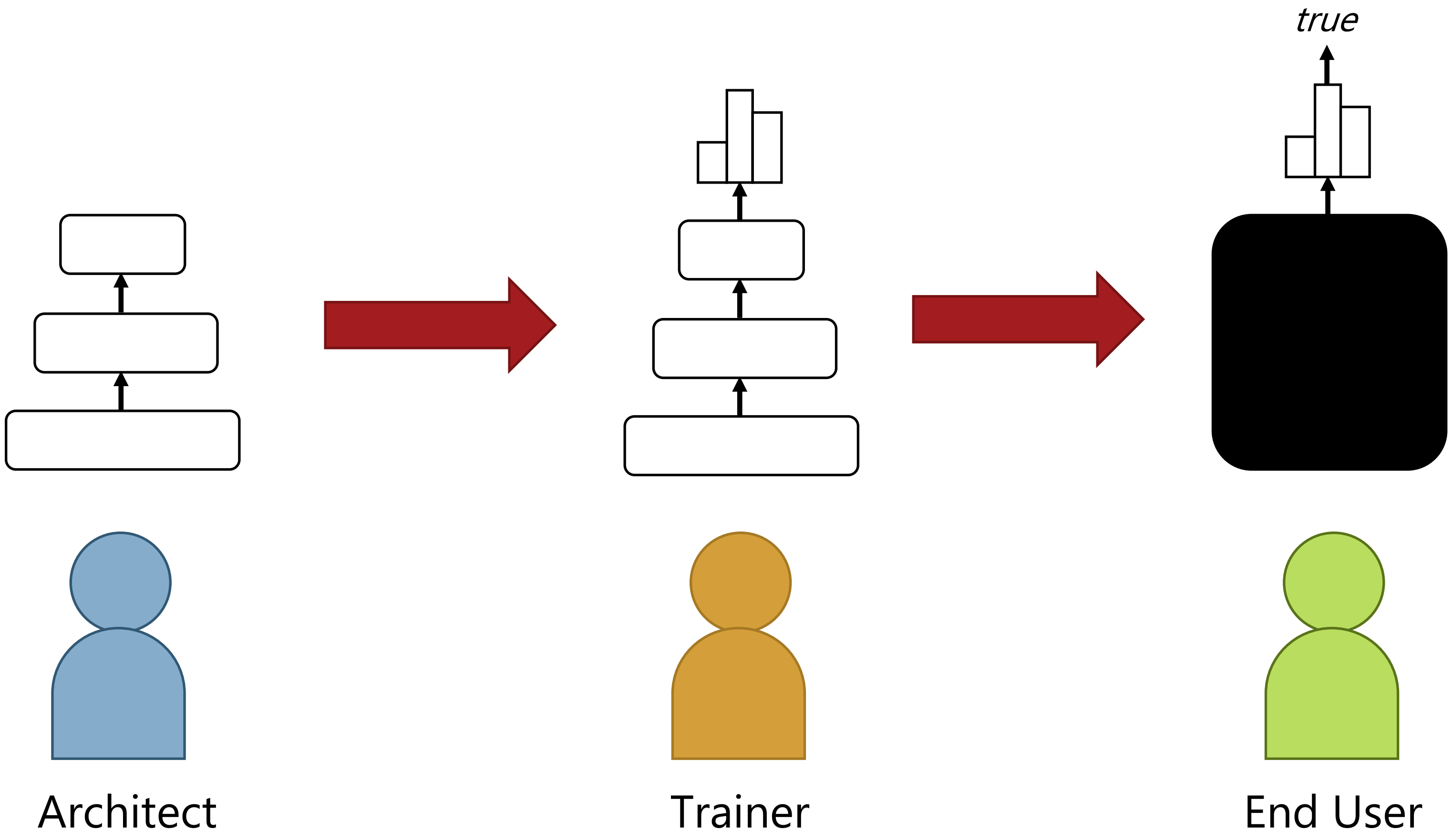
true

not enough info

false

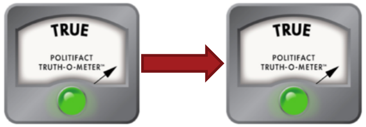
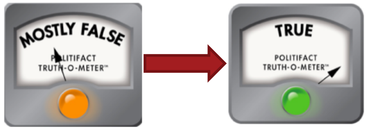
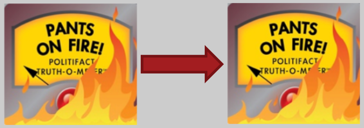
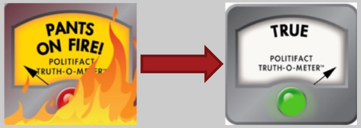
Explainability

Explainability – what is it and why do we need it?

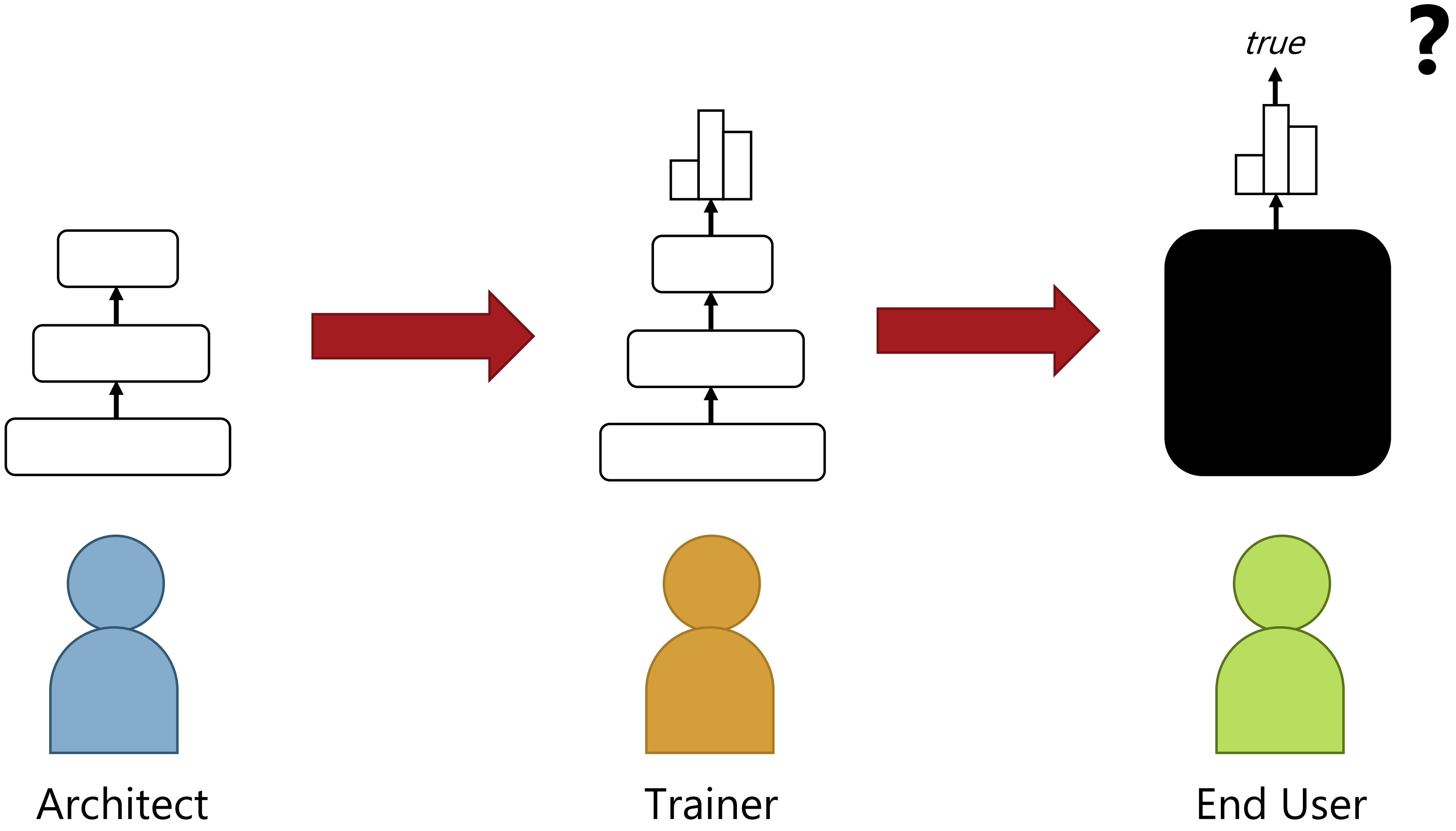


Terminology borrowed from Strobelt et al. (2018), "LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE transactions on visualization and computer graphics."

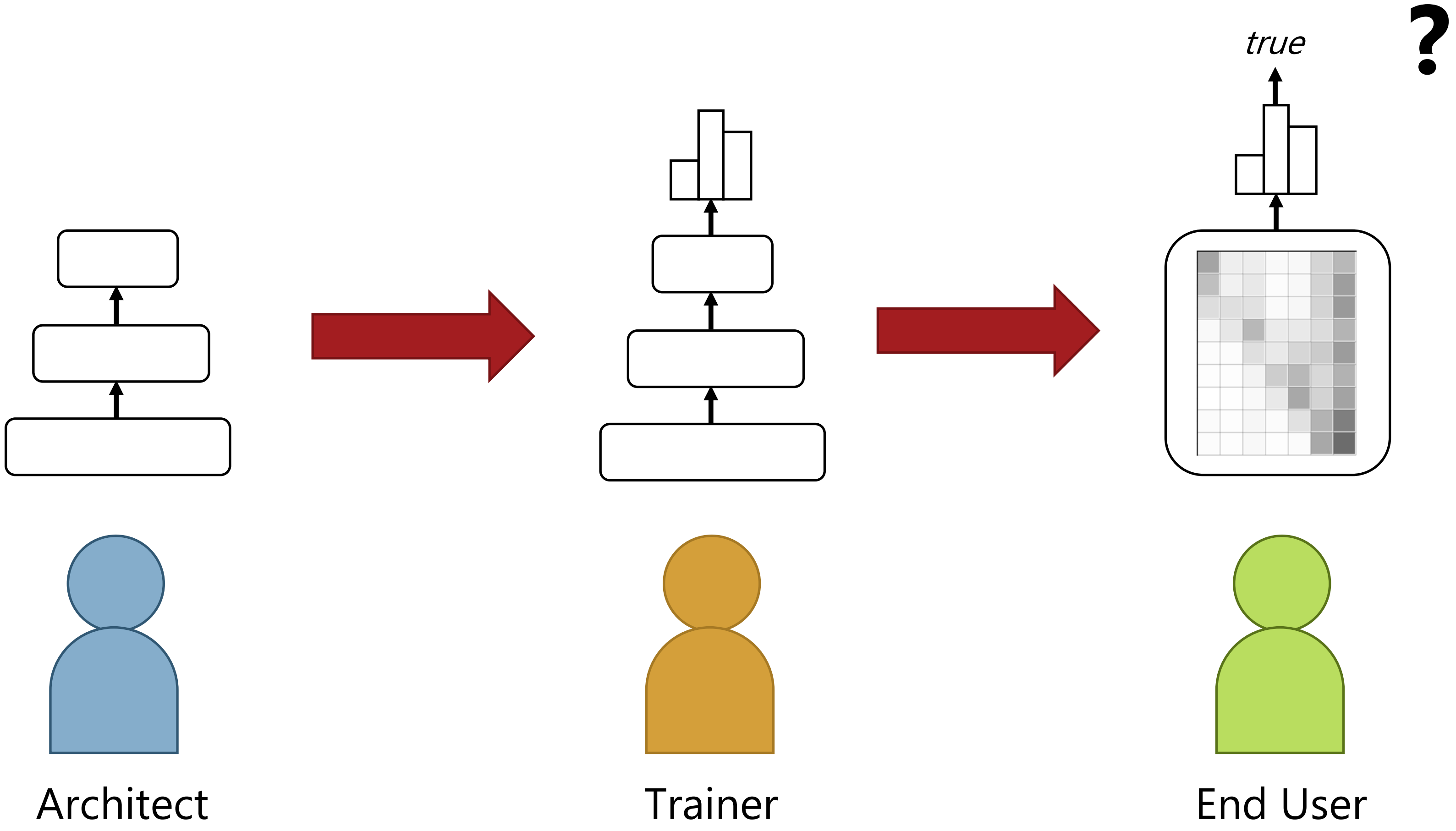
Explainability – what is it and why do we need it?

	Right prediction	Wrong prediction
Right reasons	**Claim:** “In the COVID-19 crisis, ‘only 20% of African Americans had jobs where they could work from home.’” **Evidence:** “20% of black workers said they could work from home in their primary job, compared to 30% of white workers.”  TRUE POLITIFACT TRUTH-O-METER TRUE POLITIFACT TRUTH-O-METER	**Claim:** “Children don’t seem to be getting this virus.” **Evidence**: “There have been no reported incidents of infection in children.”  MOSTLY FALSE POLITIFACT TRUTH-O-METER TRUE POLITIFACT TRUTH-O-METER
Wrong reasons	**Claim:** “Taylor Swift had a fatal car accident.” **Reason**: overfitting to spurious patterns (celebrity death hoaxes are common)  PANTS ON FIRE! POLITIFACT TRUTH-O-METER PANTS ON FIRE! POLITIFACT TRUTH-O-METER	**Claim**: “Michael Jackson is still alive, appears in daughter’s selfie.” **Reason**: overfitting to spurious patterns (celebrity death hoaxes are common)  PANTS ON FIRE! POLITIFACT TRUTH-O-METER TRUE POLITIFACT TRUTH-O-METER

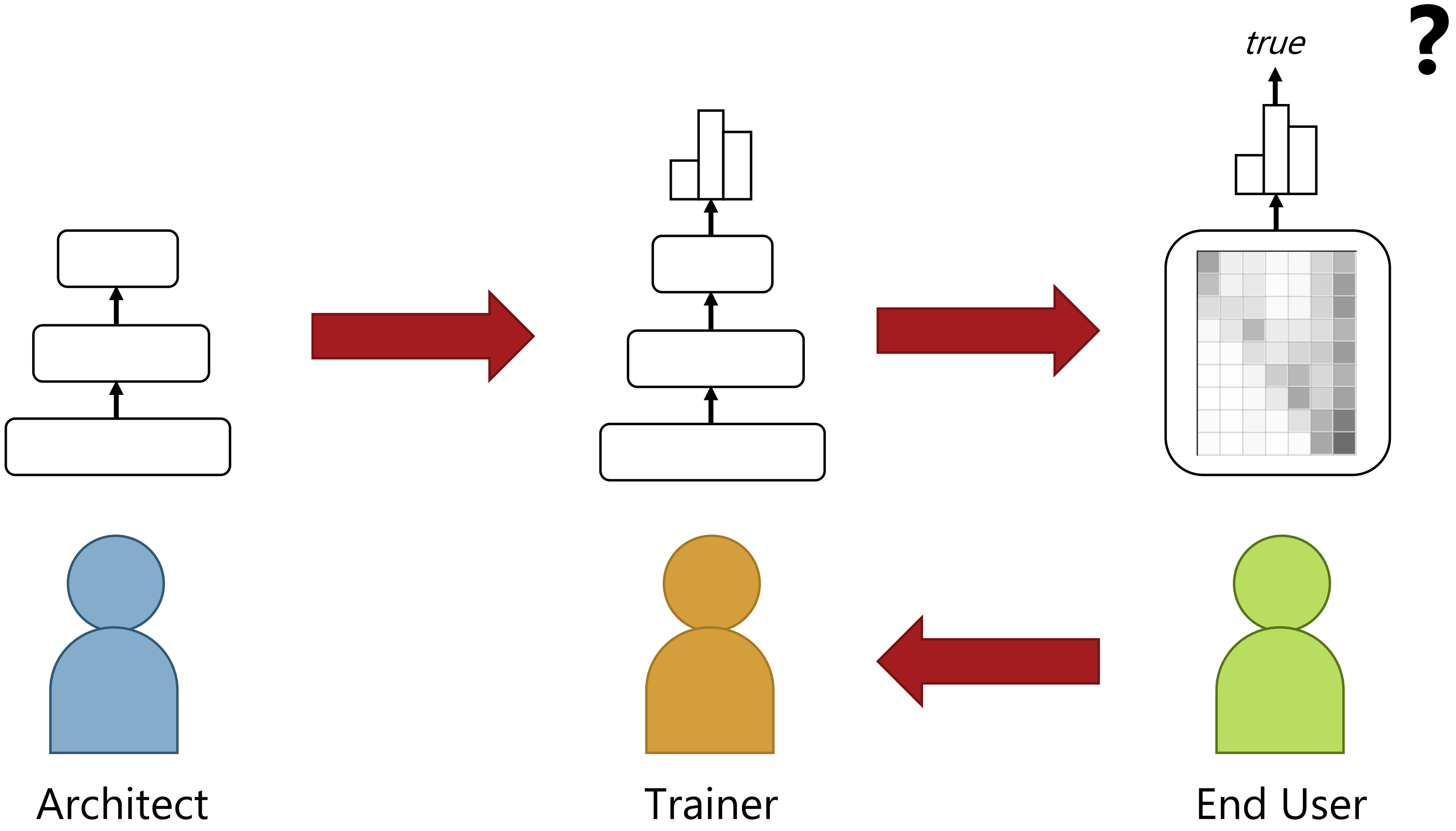
Explainability – what is it and why do we need it?



Explainability – what is it and why do we need it?



Explainability – what is it and why do we need it?



Overview of Today’s Talk

- **Introduction**
  - Fact checking – detecting false information online
  - Explainability – making the right prediction for the right reasons
- **Diagnostic Properties**
  - Evaluating explanations using different diagnostic properties
  - Diagnostics-guided explanation generation
- **Generating Free-Text Explanations for Fact Checking**
  - Supervised generation of explanations
  - Unsupervised post-editing of explanations
- **Detecting Vulnerabilities of Fact Checking Models**
  - Prediction without sufficient evidence
- **Wrap-up**

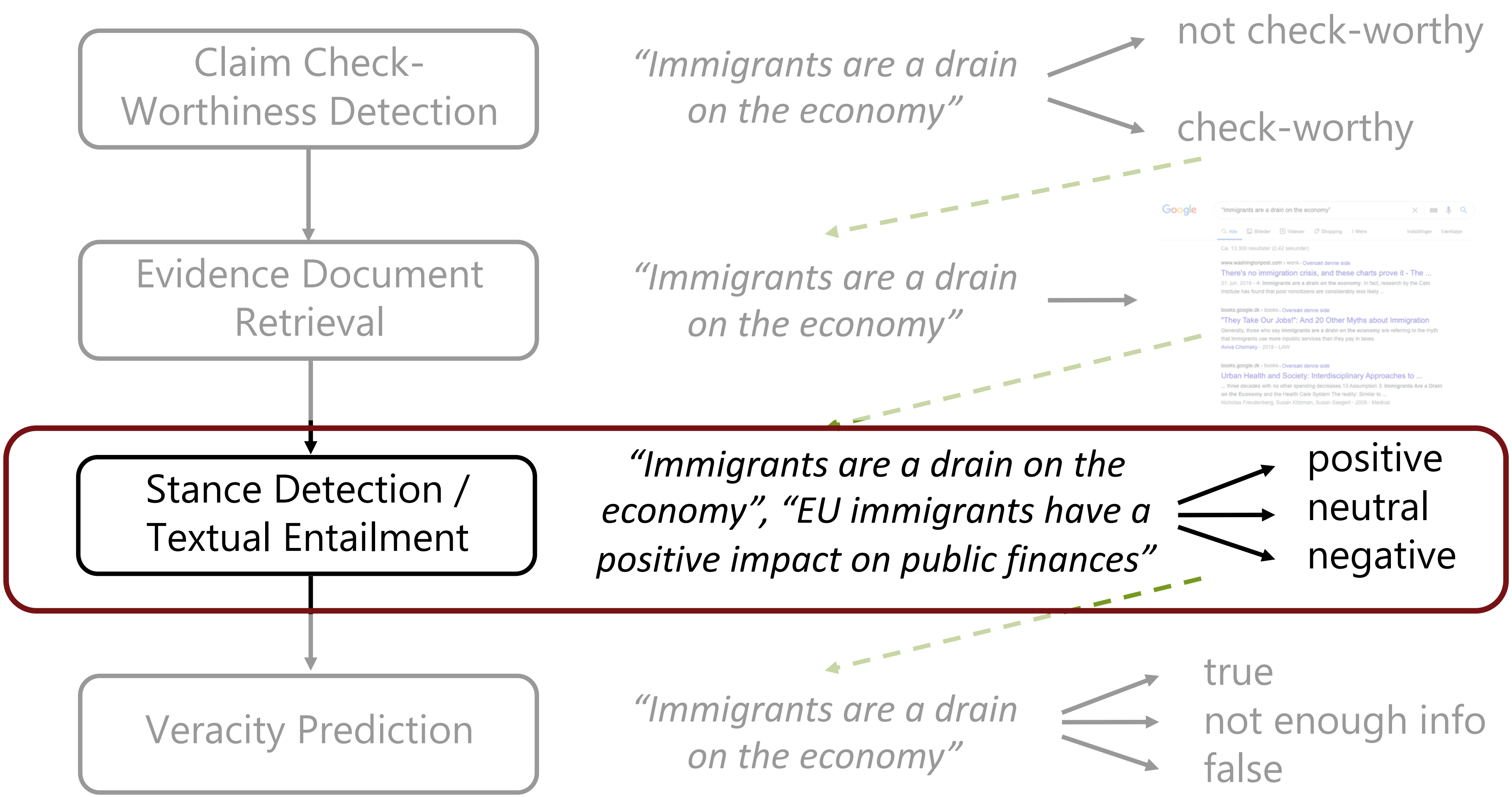
Diagnostic Properties

A Diagnostic Study of Explainability Techniques for Text Classification

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein

EMNLP 2020

A Typical Fact Checking Pipeline



Claim Check-Worthiness Detection
“Immigrants are a drain on the economy”
not check-worthy
check-worthy
Evidence Document Retrieval
“Immigrants are a drain on the economy”
Stance Detection / Textual Entailment
“Immigrants are a drain on the economy”, “EU immigrants have a positive impact on public finances”
positive
neutral
negative
Veracity Prediction
“Immigrants are a drain on the economy”
true
not enough info
false
Google
Alle
Billeder
Videoer
Shopping
Mere
Indstillinger
Værktøjer

"Immigrants are a drain on the economy"

Ca. 13.300 resultater (0,42 sekunder)

www.washingtonpost.com › wonk - Oversæt denne side
There's no immigration crisis, and these charts prove it - The ...
21. jun. 2018 - 4: **Immigrants are a drain on the economy**. In fact, research by the Cato Institute has found that poor noncitizens are considerably less likely ...

books.google.dk › books - Oversæt denne side
"They Take Our Jobs!": And 20 Other Myths about Immigration
Generally, those who say **Immigrants are a drain on the economy** are referring to the myth that immigrants use more inpublic services than they pay in taxes.
Aviva Chomsky - 2018 - LAW

books.google.dk › books - Oversæt denne side
Urban Health and Society: Interdisciplinary Approaches to ...
... three decades with no other spending decreases.13 Assumption 3: **Immigrants Are a Drain on the Economy** and the Health Care System The reality: Similar to ...
Nicholas Freudenberg, Susan Klitzman, Susan Saegert - 2009 - Medical

Explanation via Rationale Selection -- Sentences

Claim: The Faroe Islands are no longer part of a kingdom.

Evidence document: The Faroe Islands, also called the Faeroes, is an archipelago between the Norwegian Sea and the North Atlantic, about halfway between Norway and Iceland, 200 mi north-northwest of Scotland.
The Islands are an autonomous country within the Kingdom of Denmark.
Its area is about 1,400 km2 with a population of 50,030 in April 2017.
(···)

Label: refutes

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal (2018). FEVER: a large-scale dataset for Fact Extraction and VERification. In Proc. of NAACL.
https://arxiv.org/abs/1803.05355

Explanation via Rationale Selection -- Sentences

Claim: The Faroe Islands are no longer part of a kingdom.

Evidence document: The Faroe Islands, also called the Faeroes, is an archipelago between the Norwegian Sea and the North Atlantic, about halfway between Norway and Iceland, 200 mi north-northwest of Scotland.
The Islands are an autonomous country within the Kingdom of Denmark.
Its area is about 1,400 km2 with a population of 50,030 in April 2017.
(···)

Label: refutes

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal (2018). FEVER: a large-scale dataset for Fact Extraction and VERification. In Proc. of NAACL.
https://arxiv.org/abs/1803.05355

Explanation via Rationale Selection -- Words

Hypothesis: An adult dressed in black holds a stick.

Premise: An adult is walking away empty-handedly.

Label: contradiction

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, Phil Blunsom (2018). e-SNLI: Natural Language Inference with Natural Language Explanations. In Proc. of NeurIPS. https://proceedings.neurips.cc/paper/2018/hash/4c7a167bb329bd92580a99ce422d6fa6-Abstract.html

Explanation via Rationale Selection -- Words

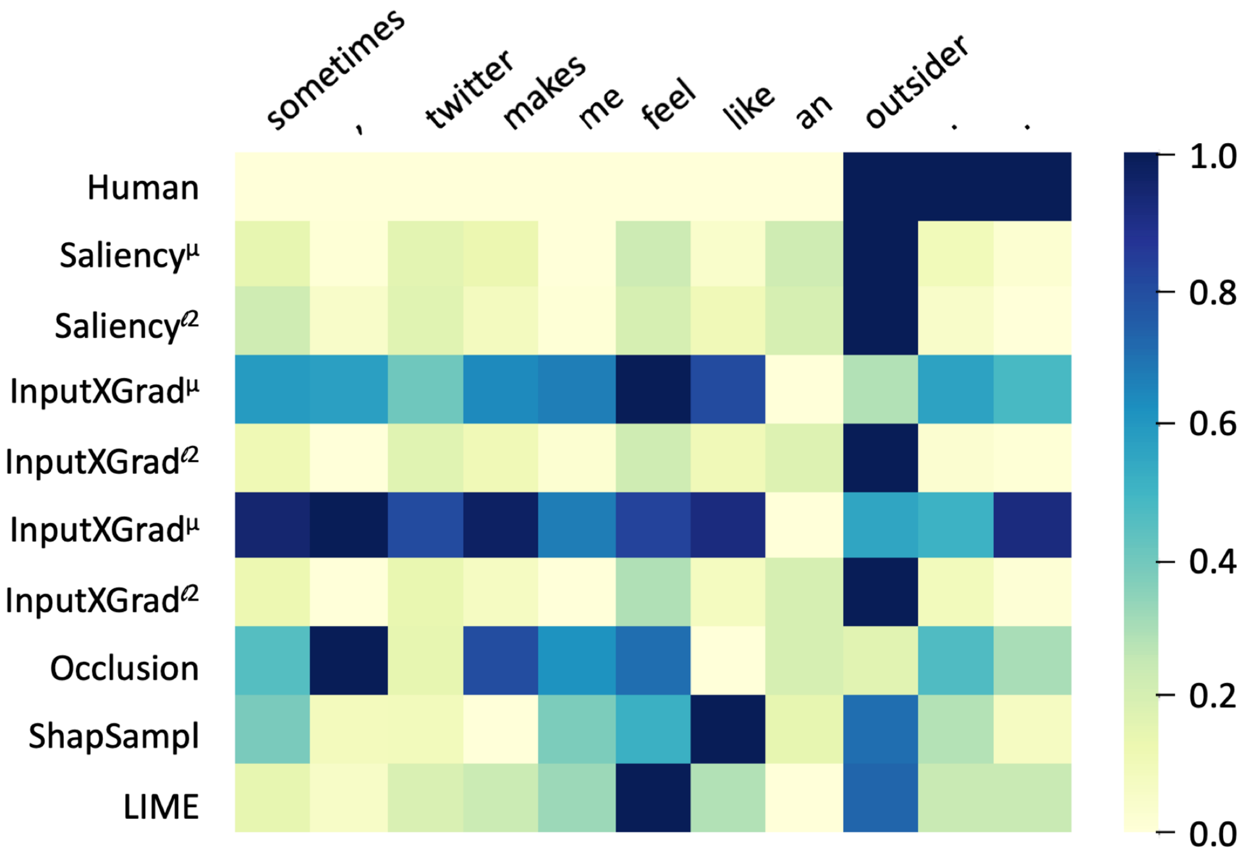
Hypothesis: An adult dressed in black holds a stick.

Premise: An adult is walking away empty-handedly.

Label: contradiction

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, Phil Blunsom (2018). e-SNLI: Natural Language Inference with Natural Language Explanations. In Proc. of NeurIPS. https://proceedings.neurips.cc/paper/2018/hash/4c7a167bb329bd92580a99ce422d6fa6-Abstract.html

Post-Hoc Explainability Methods via Rationale Selection



Example from Twitter Sentiment Extraction (TSE) dataset

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein (2020). A Diagnostic Study of Explainability Techniques for Text Classification. In EMNLP. https://www.aclweb.org/anthology/2020.emnlp-main.263/

Post-Hoc Explainability Methods via Rationale Selection

Gradient-Based Approaches

- Compute gradient of input w.r.t. output
- Different aggregation methods used to produce one score per input token (mean average, L2 norm aggregation)
- *Saliency, InputX-Gradient, Guided Backpropagation*

Perturbation-Based Approaches

- Replace tokens in input with other tokens to compute their relative contributions
- *Occlusion, Shapley Value Sampling*

Simplification-Based Approaches

- Train local linear models to approximate local decision boundaries
- *LIME*

Post-Hoc Explainability via Rationale Selection: Research Questions

- How can explainability methods be evaluated?
  - Proposal: set of **diagnostic properties**

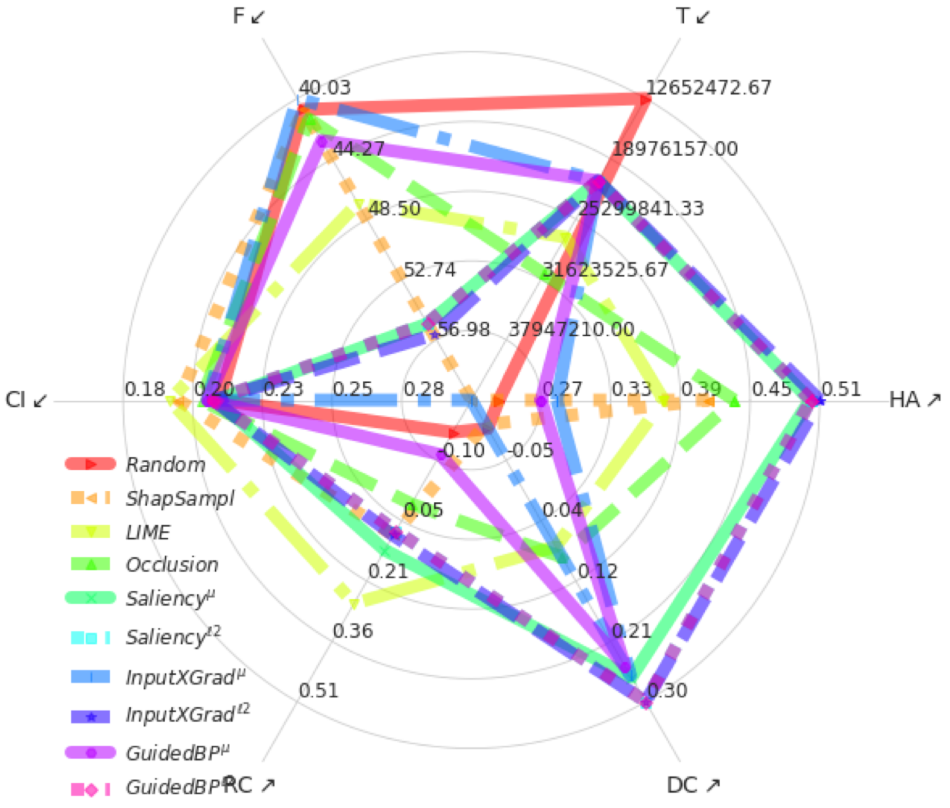
- What are characteristics of different explainability methods?
- How do explanations for models with different architectures differ?
- How do automatically and manually generated explanations differ?

Atanasova, et al., 2020 "A Diagnostic Study of Explainability Techniques for Text Classification"

Post-Hoc Explainability Methods via Rationale Selection: Diagnostic Properties

- **Agreement with Human Rationales (HA)**
  - Degree of overlap between human and automatic saliency scores
- **Faithfulness (F)**
  - Mask most salient tokens, measure drop in performance
- **Rationale Consistency (RC)**
  - Difference between explanations for models trained with different random seeds, with model with random weights
- **Dataset Consistency (DC)**
  - Difference between explanations for similar instances
- **Confidence Indication (CI)**
  - Predictive power of produced explanations for model's confidence

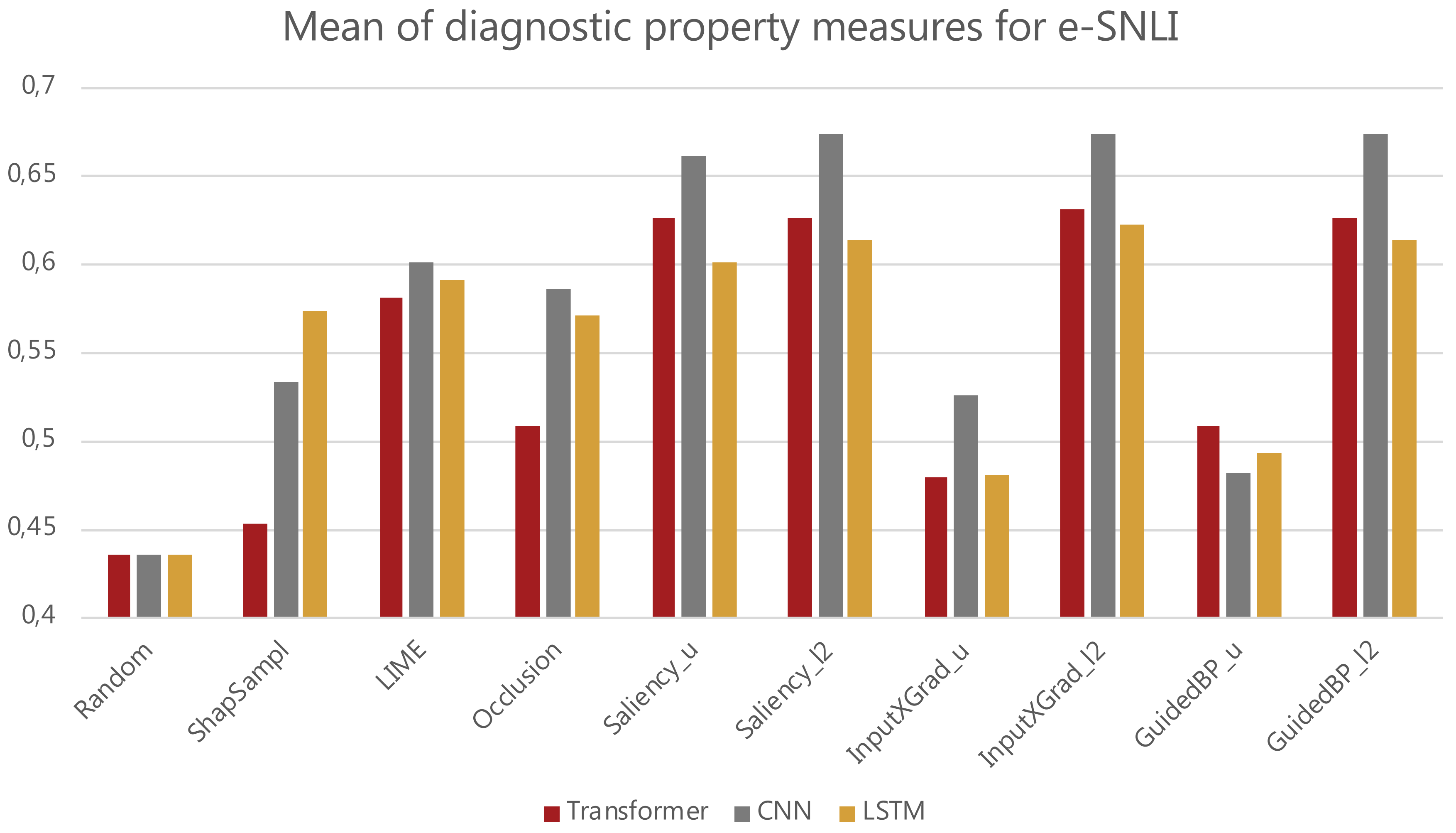
Post-Hoc Explainability Methods via Rationale Selection: Selected Results



HA: Agreement with human rationales
F: Faithfulness
RC: Rationale Consistency
DC: Dataset Consistency
CI: Confidence indication
T: Computing Time

Spider chart for Transformer model on e-SNLI

Post-Hoc Explainability Methods via Rationale Selection: Aggregated Results



Summary

- Different explainability techniques
    - Produce different explanations
    - More than one explanation can be correct
    - Different aspects to what makes a good explanation
- Diagnostic properties allow one to assess different aspects of explainability techniques

Diagnostics-Guided Explanation Generation

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein

AAAI 2022

A Typical Fact Checking Pipeline

Claim Check-Worthiness Detection

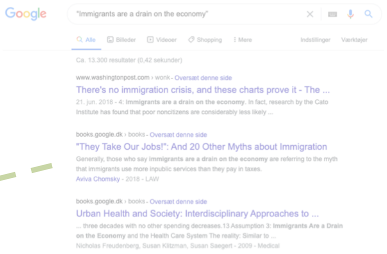
"Immigrants are a drain on the economy"

not check-worthy

check-worthy

Evidence Document Retrieval

"Immigrants are a drain on the economy"



Ca. 13.300 resultater (0,42 sekunder)

www.washingtonpost.com › wonk - Oversæt denne side
There's no immigration crisis, and these charts prove it - The ...
21. jun. 2018 - 4: **Immigrants are a drain on the economy**. In fact, research by the Cato Institute has found that poor noncitizens are considerably less likely ...

books.google.dk › books - Oversæt denne side
"They Take Our Jobs!": And 20 Other Myths about Immigration
Generally, those who say **Immigrants are a drain on the economy** are referring to the myth that immigrants use more inpublic services than they pay in taxes.
Aviva Chomsky - 2018 - LAW

books.google.dk › books - Oversæt denne side
Urban Health and Society: Interdisciplinary Approaches to ...
... three decades with no other spending decreases.13 Assumption 3: **Immigrants Are a Drain on the Economy** and the Health Care System The reality: Similar to ...
Nicholas Freudenberg, Susan Klitzman, Susan Saegert - 2009 - Medical

Stance Detection / Textual Entailment

"Immigrants are a drain on the economy", "EU immigrants have a positive impact on public finances"

positive
neutral
negative

Veracity Prediction

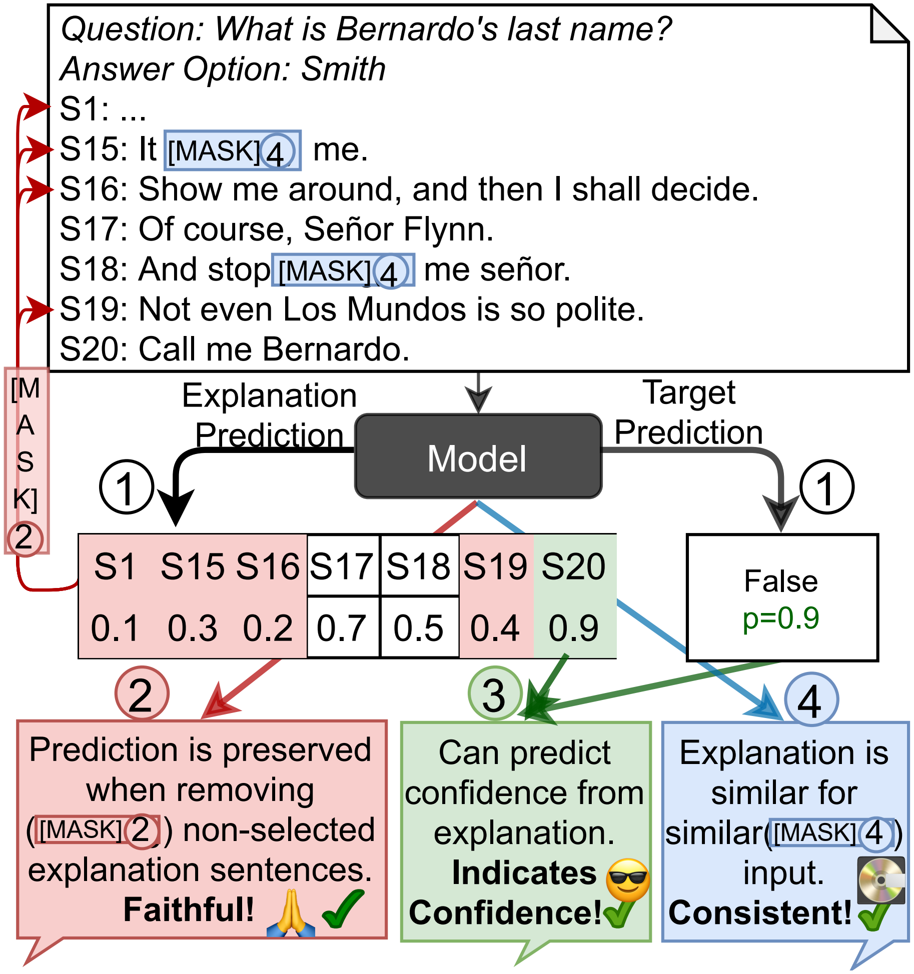
"Immigrants are a drain on the economy"

true
not enough info
false

Diagnostic-Guided Explanations

- Explanations currently suffer from:
  - 🙏 lack of faithfulness to the underlying model
  - 📀 lack of consistency for similar data points
  - 😎 lack of confidence in explanation predictions
- We present the first method to learn those diagnostic properties in an unsupervised way
- We directly optimise for them to improve the quality of generated explanations

Diagnostic-Guided Explanations



Question: What is Bernardo's last name?
Answer Option: Smith
S1: ...
S15: It [MASK] 4 me.
S16: Show me around, and then I shall decide.
S17: Of course, Señor Flynn.
S18: And stop [MASK] 4 me señor.
S19: Not even Los Mundos is so polite.
S20: Call me Bernardo.

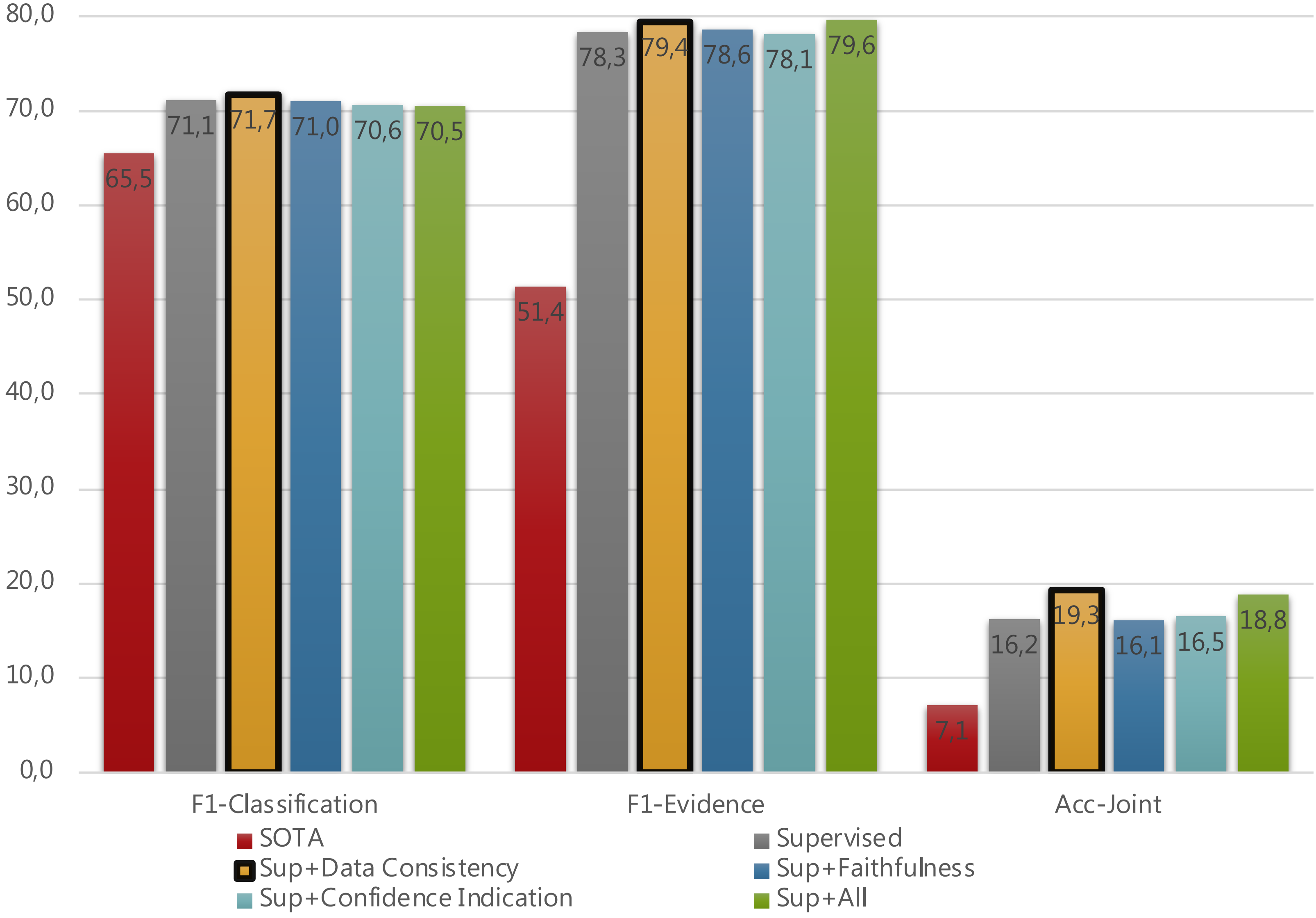
S1	S15	S16	S17	S18	S19	S20
0.1	0.3	0.2	0.7	0.5	0.4	0.9

Prediction is preserved when removing ([MASK] 2) non-selected explanation sentences. **Faithfull!** 🙏 ✓

Can predict confidence from explanation. **Indicates** 😎 **Confidence!** ✓

Explanation is similar for similar([MASK] 4) input. 💿 **Consistent!** ✓

Selected Results (MultiRC)



Effects of Diagnostic Property Objectives

Data Consistency

- Improves explanations by removing sentences unrelated to the target prediction.

Question: What colors are definitely used in the picture Lucy drew?

Answer: Yellow and purple **Label**: True

Supervised

Pred: True, p=.98

E: She makes sure to draw her mom named Martha wearing a purple dress, because that is her favorite. She draws many yellow feathers for her pet bird named Andy. She draws a picture of her family.

Remove unrelated

Supervised + DC

Pred: True, p=.99

E: She makes sure to draw her mom named Martha wearing a purple dress, because that is her favorite. She draws many yellow feathers for her pet bird named Andy.

Keep related to target prediction

Effects of Diagnostic Property Objectives

🙏 Faithfulness

- Improves the explanations to reflect the rationale used to predict the target for instances classified correctly by the supervised model.

Claim: Zoey Deutch did not portray Rosemarie Hathaway in Vampire Academy. **Label**: REFUTE

Supervised

Pred: Refute, p=.99
E: Zoey Francis Thompson Deutch (born November 10, 1994) is an American actress.

Does not lead to correct prediction

Supervised + F

Pred: Refute, p=.99
E: She is known for portraying Rosemarie ``Rose'' Hathaway in Vampire Academy(2014), Beverly in the Richard Link later film Everybody Wants Some!!

Effects of Diagnostic Property Objectives

😎 Confidence Indication

- Re-calibrates the prediction probabilities of generated explanations and predicted target tasks.
- Does not change many target predictions.

Label: Positive

E (both Sup-Sup-CI): For me, they calibrated my creativity as a child; they are masterful, original works of art that mix moving stories with what were astonishing special effects at the time (and they still hold up pretty well).

Predicted Sup: negative, p=.99

Predicted Sup+CI: positive, p=.99

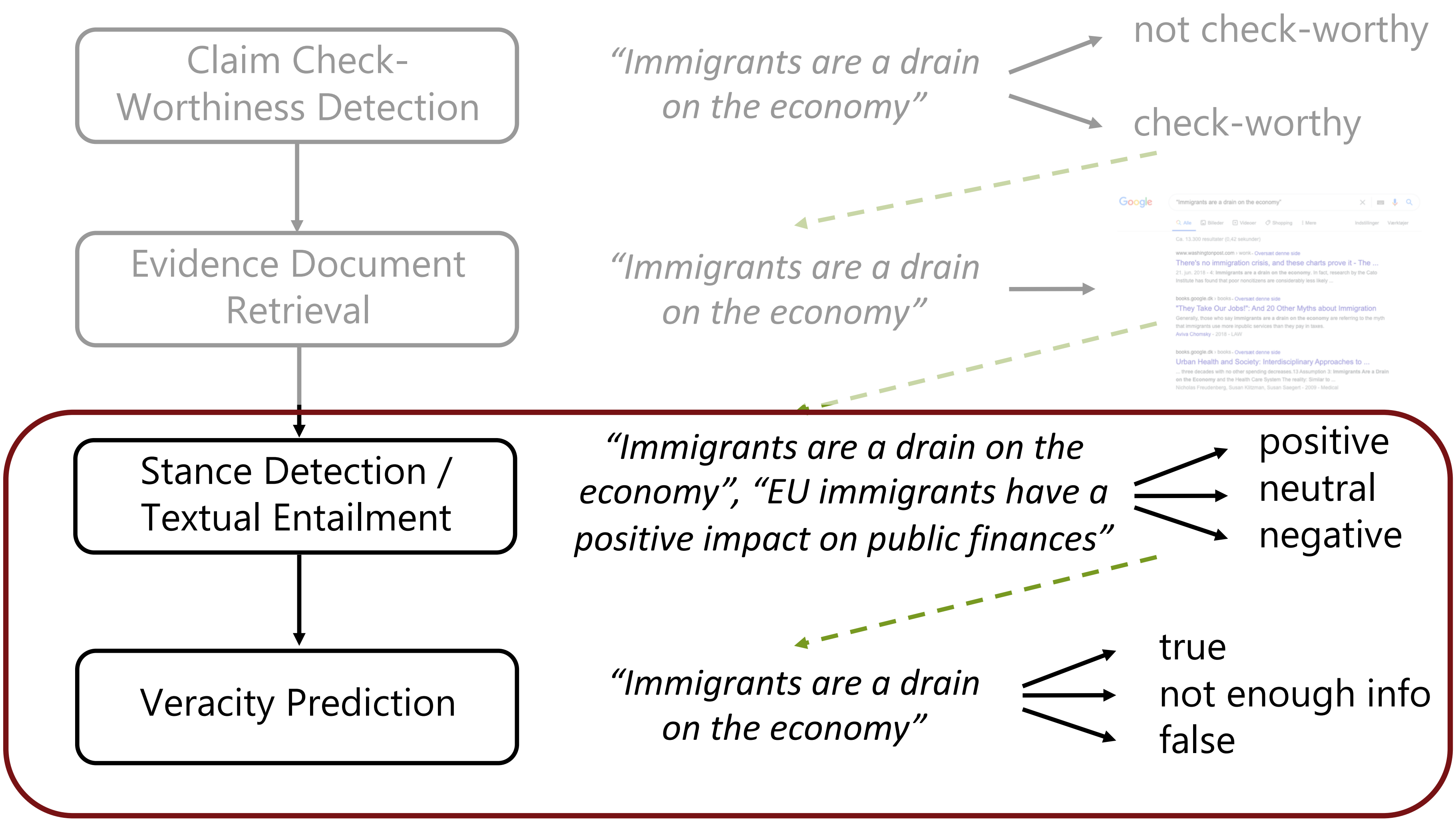
Not aligned with the high confidence of the explanation for the positive class.

Generating Fact Checking Explanations

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein

ACL 2020

A Typical Fact Checking Pipeline



Claim Check-Worthiness Detection

"Immigrants are a drain on the economy"

not check-worthy

check-worthy

Evidence Document Retrieval

"Immigrants are a drain on the economy"

Google

"Immigrants are a drain on the economy"

Alle · Billeder · Videoer · Shopping · Mere · Indstillinger · Værktøjer

Ca. 13.300 resultater (0,42 sekunder)

www.washingtonpost.com › wonk - Oversæt denne side

There's no immigration crisis, and these charts prove it - The ...

21. jun. 2018 - 4: **Immigrants are a drain on the economy**. In fact, research by the Cato Institute has found that poor noncitizens are considerably less likely ...

books.google.dk › books - Oversæt denne side

"They Take Our Jobs!": And 20 Other Myths about Immigration

Generally, those who say **Immigrants are a drain on the economy** are referring to the myth that immigrants use more inpublic services than they pay in taxes.

Aviva Chomsky - 2018 - LAW

books.google.dk › books - Oversæt denne side

Urban Health and Society: Interdisciplinary Approaches to ...

... three decades with no other spending decreases.13 Assumption 3: **Immigrants Are a Drain on the Economy** and the Health Care System The reality: Similar to ...

Nicholas Freudenberg, Susan Klitzman, Susan Saegert - 2009 - Medical

Stance Detection / Textual Entailment

"Immigrants are a drain on the economy", "EU immigrants have a positive impact on public finances"

positive

neutral

negative

Veracity Prediction

"Immigrants are a drain on the economy"

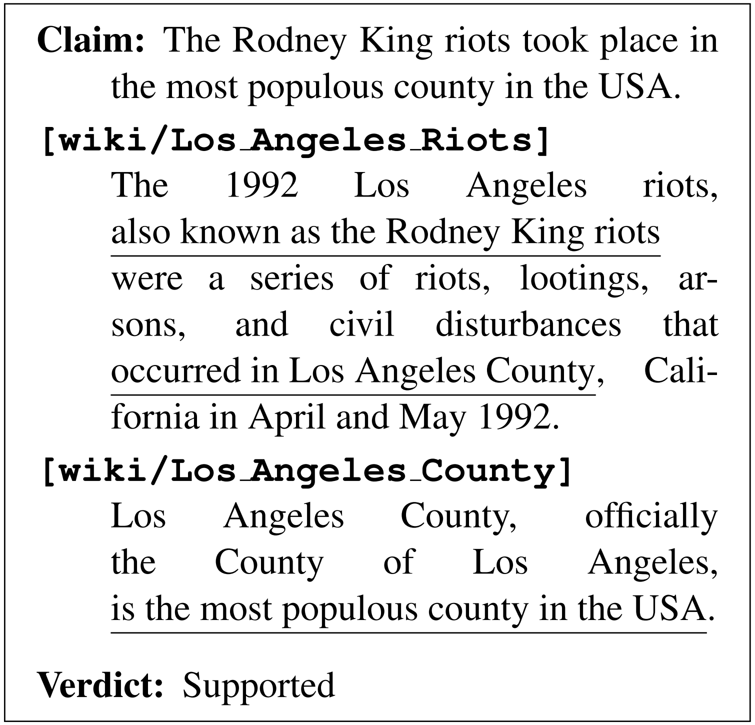
true

not enough info

false

Extracted Wikipedia Sentences as Explanations

- FEVER dataset (Thorne et al., 2018)
  - Claims are re-written sentences from Wikipedia.
  - Explanation consists of evidence sentences extracted from Wikipedia pages.
  - Biggest dataset for fact checking
  - Brings small or no improvements for veracity prediction with real-world claims (Wadden et al., 2020, Ostrowski et al., 2020)



Claim: The Rodney King riots took place in the most populous county in the USA.

`[wiki/Los_Angeles_Riots]`

The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

`[wiki/Los_Angeles_County]`

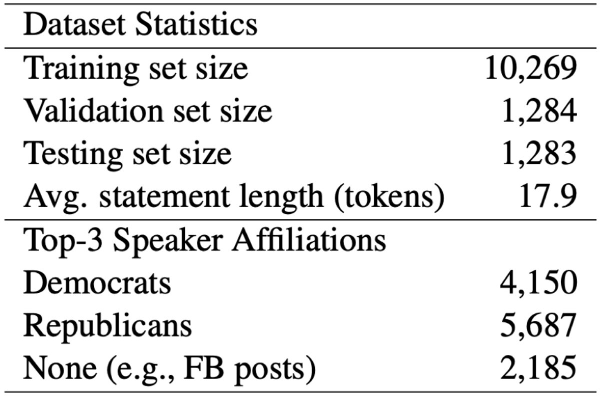
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

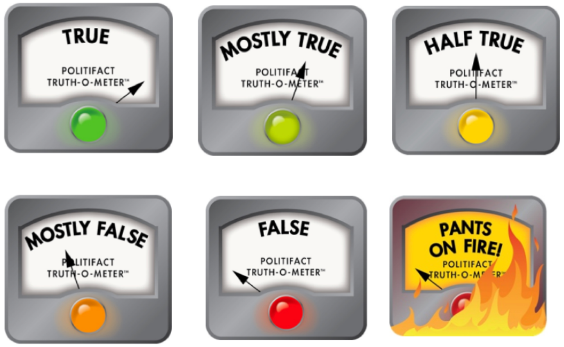
Thorne et al., 2018 “FEVER: a large-scale dataset for Fact Extraction and VERification”
Thorne et al., 2021 “Evidence-based Verification for Real World Information Needs”
Ostrowski et al., 2021 “Multi-Hop Fact Checking of Political Claims”
Wadden et al., 2020 “Fact or Fiction: Verifying Scientific Claims”

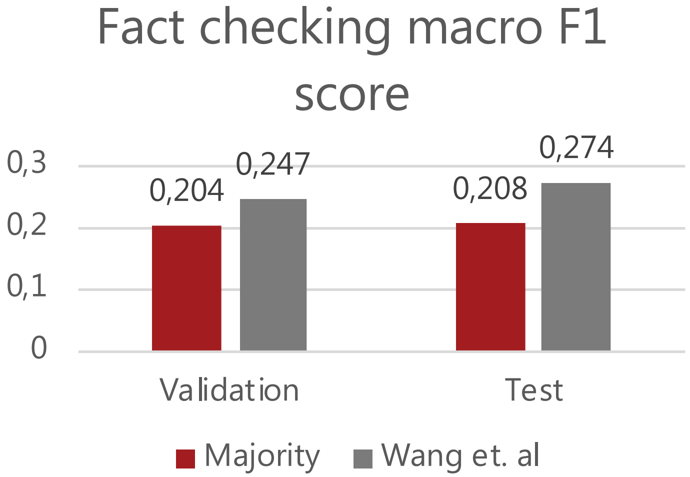
Real World Fact Checking

Statement: “The last quarter, it was just announced, our gross domestic product was below zero. Who ever heard of this? Its never below zero.”
Speaker: Donald Trump
Context: presidential announcement speech
Label: Pants on Fire



Dataset Statistics	
Training set size	10,269
Validation set size	1,284
Testing set size	1,283
Avg. statement length (tokens)	17.9
Top-3 Speaker Affiliations	
Democrats	4,150
Republicans	5,687
None (e.g., FB posts)	2,185

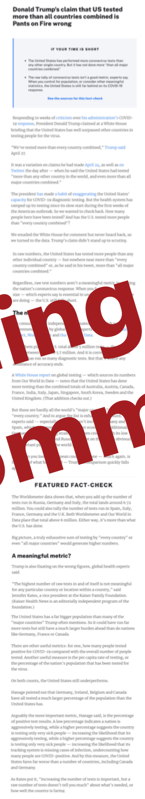

TRUE
POLITIFACT TRUTH-O-METER™
MOSTLY TRUE
POLITIFACT TRUTH-O-METER™
HALF TRUE
POLITIFACT TRUTH-O-METER™
MOSTLY FALSE
POLITIFACT TRUTH-O-METER™
FALSE
POLITIFACT TRUTH-O-METER™
PANTS ON FIRE!
POLITIFACT TRUTH-O-METER™



Wang, William Yang. "“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection." Proceedings ACL’2017.

Real-World Fact Checking Explanations

POLITIFACT
The Poynter Institute



Donald Trump's claim that US tested more than all countries combined is Pants on Fire wrong

IF YOUR TIME IS SHORT

- The United States has performed more coronavirus tests than any other single country. But it has not done more "than all major countries combined."
- The raw tally of coronavirus tests isn't a good metric, experts say. When you control for population, or consider other meaningful statistics, the United States is still far behind on its COVID-19 response.

See the sources for this fact-check

Responding to weeks of criticism over his administration's COVID-19 response, President Donald Trump claimed at a White House briefing that the United States has well surpassed other countries in testing people for the virus.

"We've tested more than every country combined," Trump said April 27.

It was a variation on claims he had made April 24, as well as on Twitter the day after — when he said the United States had tested "more than any other country in the world, and even more than all major countries combined."

The president has made a habit of exaggerating the United States' capacity for COVID-19 diagnostic testing. But the health system has ramped up its testing since its slow start during the first weeks of the American outbreak. So we wanted to check back. How many people here have been tested? And has the U.S. tested more people than "every country combined"?

We emailed the White House for comment but never heard back, so we turned to the data. Trump's claim didn't stand up to scrutiny.

In raw numbers, the United States has tested more people than any other individual country — but nowhere near more than "every country combined" or, as he said in his tweet, more than "all major countries combined."

Regardless, raw test numbers aren't a meaningful metric [illegible] the nation's coronavirus response. When you [illegible] size — which experts say is essential in un[illegible] are doing — the U.S. still falls short.

The n[illegible]

[illegible] consu[illegible] independ[illegible]mates [illegible] [illegible]mmend[illegible] by global [illegible] experts [illegible] [illegible]ect, Wo[illegible] and Our [illegible] Data.

[illegible] them pl[illegible].S. total above 5 million tests — [illegible] [illegible]etween [illegible] 5.7 million. And it is corre[illegible] [illegible] has run so many diagnostic tests. But that's where any semblance of accuracy ends.

A White House report on global testing — which sources its numbers from Our World in Data — notes that the United States has done more testing than the combined totals of Australia, Austria, Canada, France, India, Italy, Japan, Singapore, South Korea, Sweden and the United Kingdom. (That addition checks out.)

But those are hardly all the world's "major co[illegible]" [illegible] "every country." And to argue the list is exh[illegible]surd, experts said -- especiall[illegible]n't incl[illegible]any and Spain, which [illegible]t econ[illegible] additionall[illegible] [illegible] strat[illegible]en cre[illegible] its low [illegible]nd Russi[illegible]t on th[illegible] obvious[illegible] [illegible]rtant p[illegible]he world [illegible]

[illegible]n you loo[illegible]pean cou[illegible]ne — which again, is [illegible] of what h[illegible] — Trum[illegible] comparison quickly falls ap[illegible]

FEATURED FACT-CHECK

The Worldometer data shows that, when you add up the number of tests run in Russia, Germany and Italy, the total lands around 6.72 million. You could also tally the number of tests run in Spain, Italy, France, Germany and the U.K. Both Worldometer and Our World in Data place that total above 6 million. Either way, it's more than what the U.S. has done.

Big picture, a truly exhaustive sum of testing by "every country" or even "all major countries" would generate higher numbers.

A meaningful metric?

Trump is also fixating on the wrong figures, global health experts said.

"The highest number of raw tests in and of itself is not meaningful for any particular country or location within a country," said Jennifer Kates, a vice president at the Kaiser Family Foundation. (Kaiser Health News is an editorially independent program of the foundation.)

The United States has a far bigger population than many of the "major countries" Trump often mentions. So it could have run far more tests but still have a much larger burden ahead than do nations like Germany, France or Canada.

There are other useful metrics: for one, how many people tested positive for COVID-19 compared with the overall number of people tested. Another useful measure is the per capita rate of testing, or the percentage of the nation's population that has been tested for the virus.

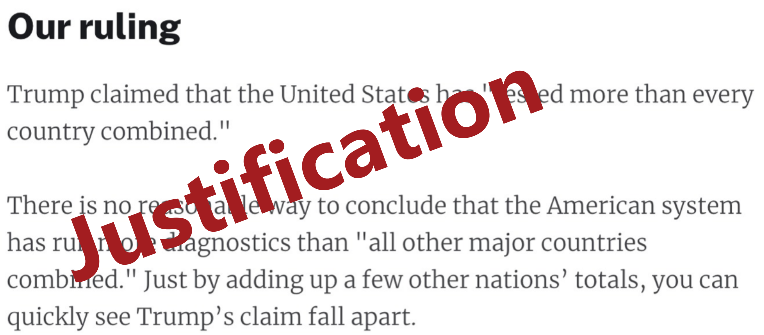
On both counts, the United States still underperforms.

Hanage pointed out that Germany, Ireland, Belgium and Canada have all tested a much larger percentage of the population than the United States has.

Arguably the more important metric, Hanage said, is the percentage of positive test results. A low percentage indicates a nation is aggressively testing, while a higher percentage suggests the country is testing only very sick people — increasing the likelihood that its aggressively testing, while a higher percentage suggests the country is testing only very sick people — increasing the likelihood that its tracking system is missing cases of infection, undercounting how many people are COVID-positive. And by this measure, the United States fares far worse than a number of countries, including Canada and Germany.

As Kates put it, "Increasing the number of tests is important, but a raw number of tests doesn't tell you much" about what's needed, or how well the country is faring.

Ruling Comments



Our ruling

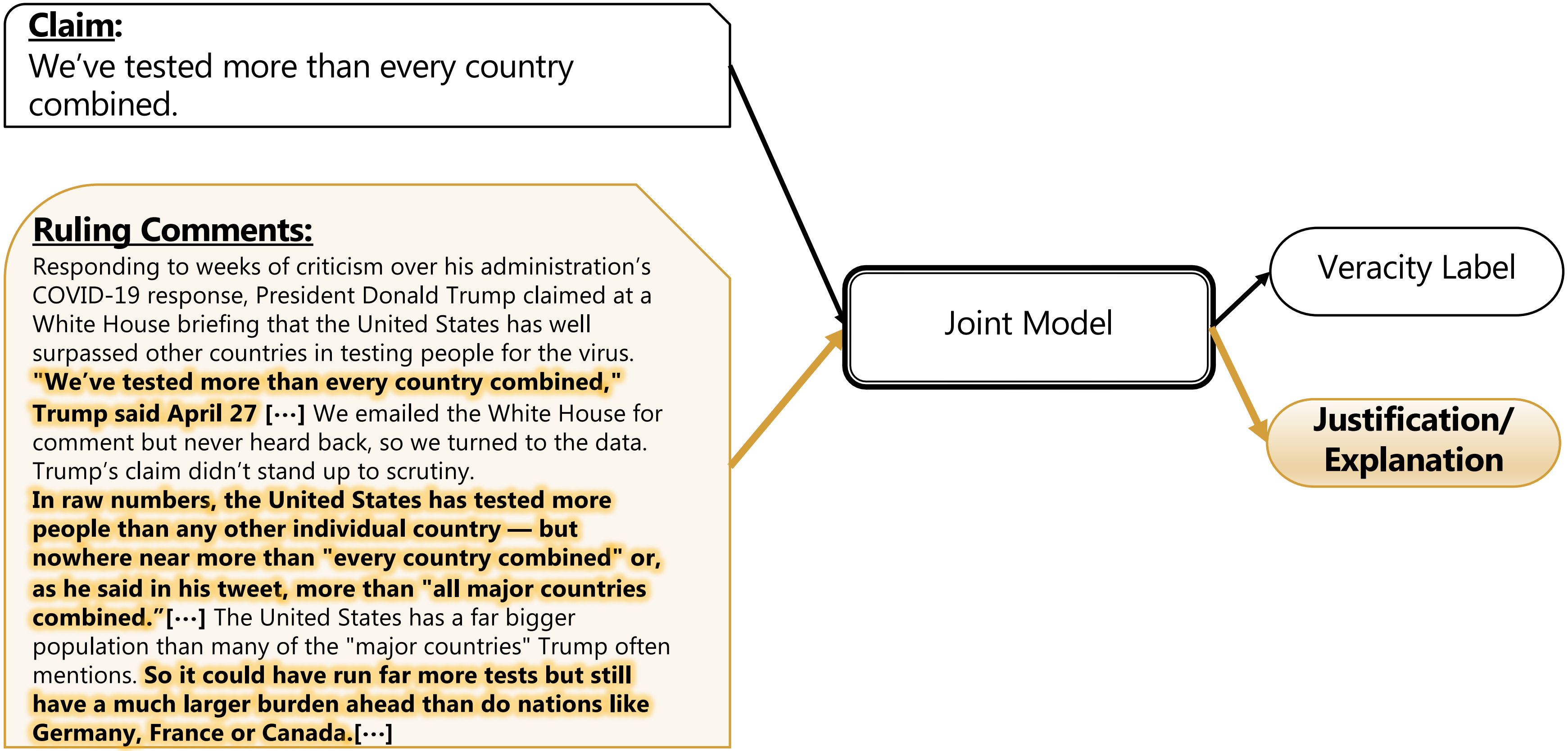
Trump claimed that the United States has "tested more than every country combined."

There is no reasonable way to conclude that the American system has run more diagnostics than "all other major countries combined." Just by adding up a few other nations' totals, you can quickly see Trump's claim fall apart.

Justification

https://www.politifact.com/factchecks/2020/apr/30/donald-trump/trumps-boasts-testing-spring-cherry-picking-data/

Generating Explanations from Ruling Comments



Claim:
We’ve tested more than every country combined.

Ruling Comments:
Responding to weeks of criticism over his administration’s COVID-19 response, President Donald Trump claimed at a White House briefing that the United States has well surpassed other countries in testing people for the virus. **"We’ve tested more than every country combined," Trump said April 27** **[⋯]** We emailed the White House for comment but never heard back, so we turned to the data. Trump’s claim didn’t stand up to scrutiny. **In raw numbers, the United States has tested more people than any other individual country — but nowhere near more than "every country combined" or, as he said in his tweet, more than "all major countries combined.”[⋯]** The United States has a far bigger population than many of the "major countries" Trump often mentions. **So it could have run far more tests but still have a much larger burden ahead than do nations like Germany, France or Canada.[⋯]**

Atanasova, et al., 2020 "Generating Fact Checking Explanations."

Related Studies on Generating Explanations

- *Camburu et. al; Rajani et. al* generate abstractive explanations
  - Short input text and explanations;
  - Large amount of annotated data.
- Real world fact checking datasets are of limited size and the input consists of long documents
- We take advantage of the LIAR-PLUS dataset:
  - Use the summary of the ruling comments as a gold explanation;
  - Formulate the problem as extractive summarization.

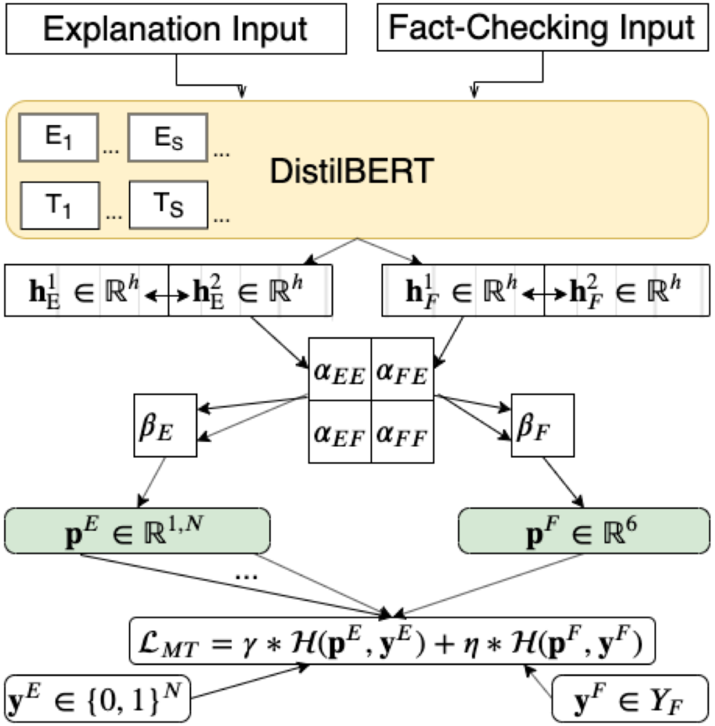
- *Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. "e-SNLI: Natural language inference with natural language explanations." In Advances in Neural Information Processing Systems,. 2018.*
- *Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4932-4942. 2019.*

Joint Explanation and Veracity Prediction

Which sentences should be selected for the explanation?

Cross-stitch layer

Multi-task objective


Explanation Input
Fact-Checking Input
E_1 ... E_S ...
T_1 ... T_S ...
DistilBERT
$\mathbf{h}_E^1 \in \mathbb{R}^h$
$\mathbf{h}_E^2 \in \mathbb{R}^h$
$\mathbf{h}_F^1 \in \mathbb{R}^h$
$\mathbf{h}_F^2 \in \mathbb{R}^h$
α_{EE} α_{FE}
α_{EF} α_{FF}
β_E
β_F
$\mathbf{p}^E \in \mathbb{R}^{1,N}$
$\mathbf{p}^F \in \mathbb{R}^6$
...
$\mathcal{L}_{MT} = \gamma * \mathcal{H}(\mathbf{p}^E, \mathbf{y}^E) + \eta * \mathcal{H}(\mathbf{p}^F, \mathbf{y}^F)$
$\mathbf{y}^E \in \{0,1\}^N$
$\mathbf{y}^F \in Y_F$

Ruder, Sebastian, et al. "Latent multi-task architecture learning." In Proceedings of the AAAI 2019.
Atanasova, et al., 2020 "Generating Fact Checking Explanations." In Proceedings of ACL 2020.

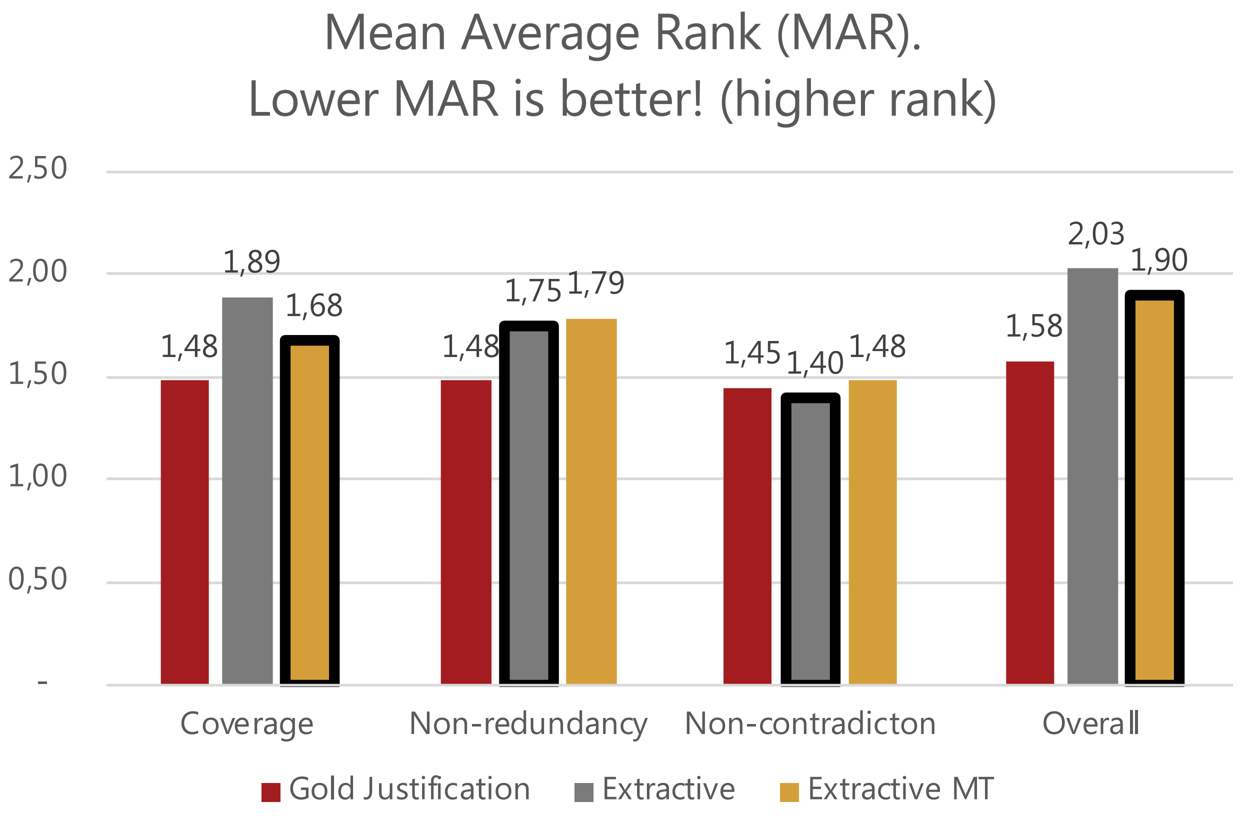
Generating Free Text Explanations: Automatic Evaluation Measures

- ***Indirectly: Fact Checking Performance***
  - ***F1.*** Predicting fact checking labels: using a joint model, from generated explanations, gold explanations, etc.
- ***Directly: Explanation Quality***
  - ***ROUGE-N.*** Overlap of n-grams between the oracle and the generated summaries.
  - ***ROUGE-L.*** Longest Common Subsequence (LCS) between the oracle and the generated summaries.

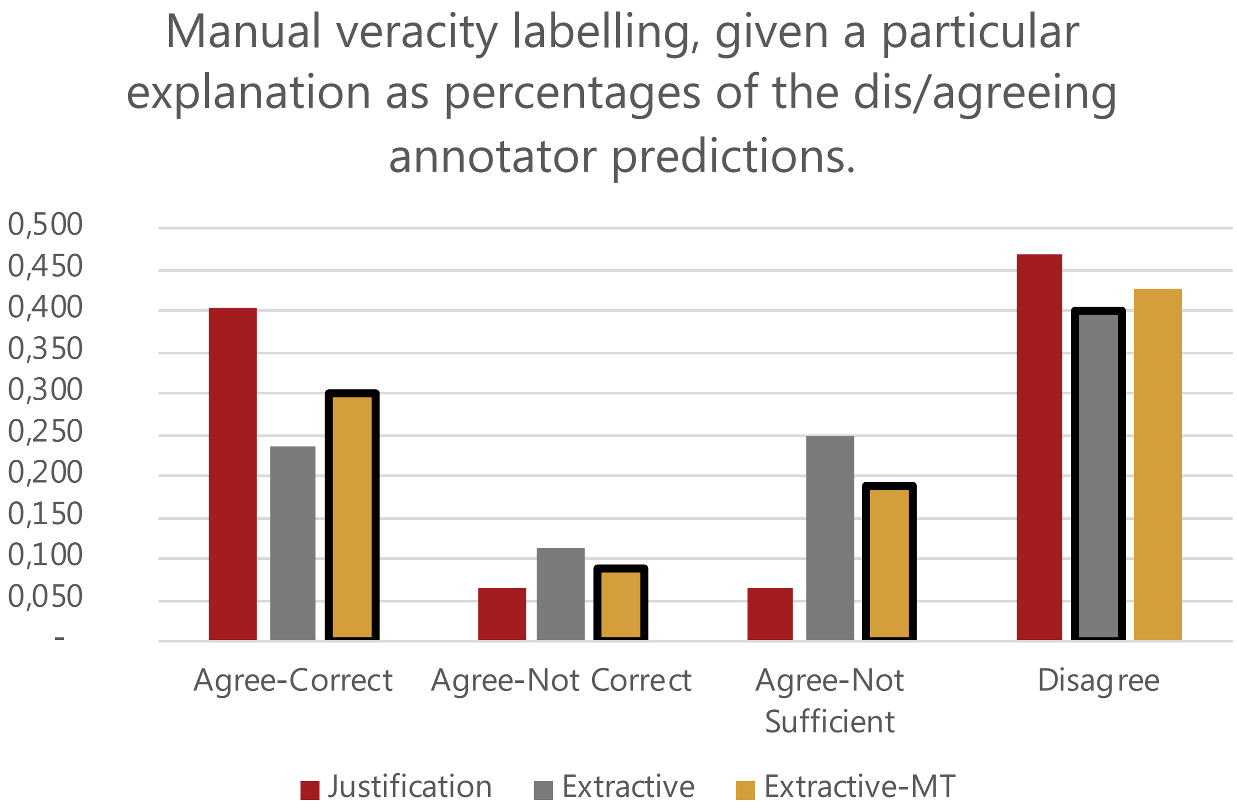
Generating Free Text Explanations: Manual Evaluation Measures

- ***Explanation Quality***
  - ***Coverage.*** The explanation contains important, salient information and does not miss any points important to the fact check.
  - ***Non-redundancy.*** The explanation does not contain any information that is redundant/repeated/not relevant to the claim or the fact check.
  - ***Non-contradiction.*** The explanation does not contain any pieces of information contradictory to the claim or the fact check.
  - ***Overall***. Overall explanation quality.
- ***Explanation Informativeness***. Veracity label for a claim provided based on the automatically generated explanations only.

Selected Results: Explanation Quality



Explanation Informativeness



Atanasova, et al., 2020 "Generating Fact Checking Explanations." In Proceedings of ACL 2020.

Summary

- First study on generating real-world veracity explanations

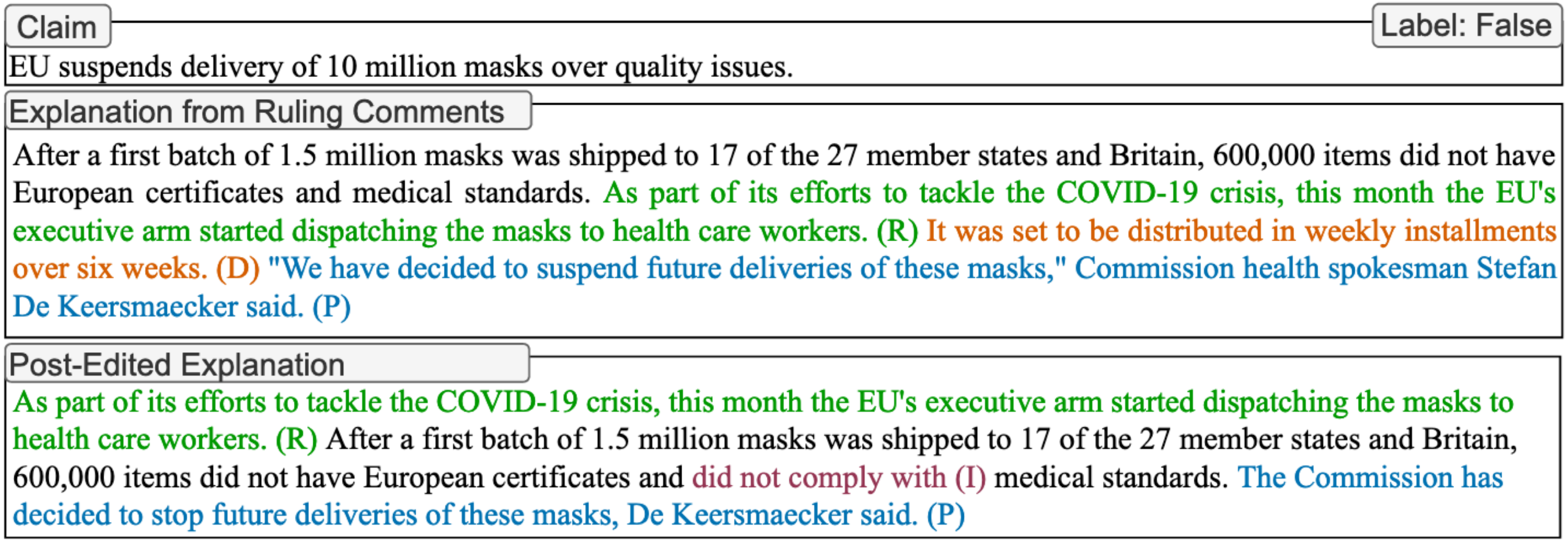
- Jointly training veracity prediction and explanation
  - improves the performance of the classification system
  - improves the coverage and overall performance of the generated explanations

Generating Fluent Fact Checking Explanations with Unsupervised Post-Editing

Shailza Jolly, Pepa Atanasova, Isabelle Augenstein

Preprint, December 2021

Ongoing Work: Unsupervised Post-Editing



Claim

Label: False

EU suspends delivery of 10 million masks over quality issues.

Explanation from Ruling Comments

After a first batch of 1.5 million masks was shipped to 17 of the 27 member states and Britain, 600,000 items did not have European certificates and medical standards. As part of its efforts to tackle the COVID-19 crisis, this month the EU's executive arm started dispatching the masks to health care workers. (R) It was set to be distributed in weekly installments over six weeks. (D) "We have decided to suspend future deliveries of these masks," Commission health spokesman Stefan De Keersmaecker said. (P)

Post-Edited Explanation

As part of its efforts to tackle the COVID-19 crisis, this month the EU's executive arm started dispatching the masks to health care workers. (R) After a first batch of 1.5 million masks was shipped to 17 of the 27 member states and Britain, 600,000 items did not have European certificates and did not comply with (I) medical standards. The Commission has decided to stop future deliveries of these masks, De Keersmaecker said. (P)

Fig. 1. Example of a post-edited explanation from PubHealth that was initially extracted from RCs. We illustrate four post-editing steps: reordering (R), insertion (I), deletion (D), and paraphrasing (P).

Shailza Jolly, Pepa Atanasova, Isabelle Augenstein (2021). Generating Fluent Fact Checking Explanations with Unsupervised Post-Editing. Preprint. https://arxiv.org/abs/2112.06924

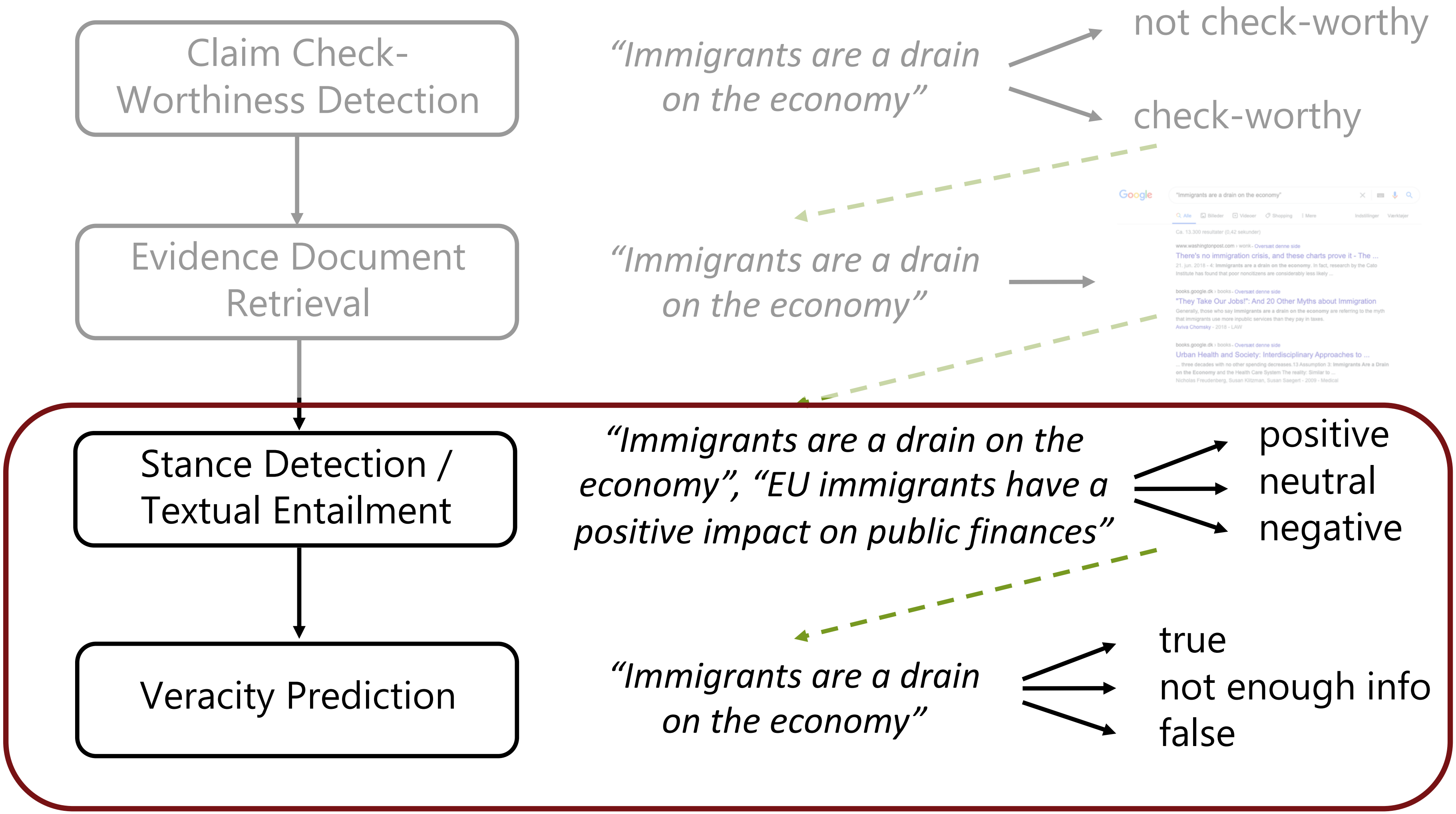
Vulnerabilities of Fact Checking Models

Fact Checking with Insufficient Evidence

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein

TACL, Vol 10 (2022), to appear

A Typical Fact Checking Pipeline



Claim Check-Worthiness Detection

"Immigrants are a drain on the economy"

not check-worthy

check-worthy

Evidence Document Retrieval

"Immigrants are a drain on the economy"

"Immigrants are a drain on the economy"

Ca. 13.300 resultater (0,42 sekunder)

www.washingtonpost.com › wonk - Oversæt denne side
There's no immigration crisis, and these charts prove it - The ...
21. jun. 2018 - 4: **Immigrants are a drain on the economy**. In fact, research by the Cato Institute has found that poor noncitizens are considerably less likely ...

books.google.dk › books - Oversæt denne side
"They Take Our Jobs!": And 20 Other Myths about Immigration
Generally, those who say **Immigrants are a drain on the economy** are referring to the myth that immigrants use more inpublic services than they pay in taxes.
Aviva Chomsky - 2018 - LAW

books.google.dk › books - Oversæt denne side
Urban Health and Society: Interdisciplinary Approaches to ...
... three decades with no other spending decreases.13 Assumption 3: **Immigrants Are a Drain on the Economy** and the Health Care System The reality: Similar to ...
Nicholas Freudenberg, Susan Klitzman, Susan Saegert - 2009 - Medical

Stance Detection / Textual Entailment

"Immigrants are a drain on the economy", "EU immigrants have a positive impact on public finances"

positive
neutral
negative

Veracity Prediction

"Immigrants are a drain on the economy"

true
not enough info
false

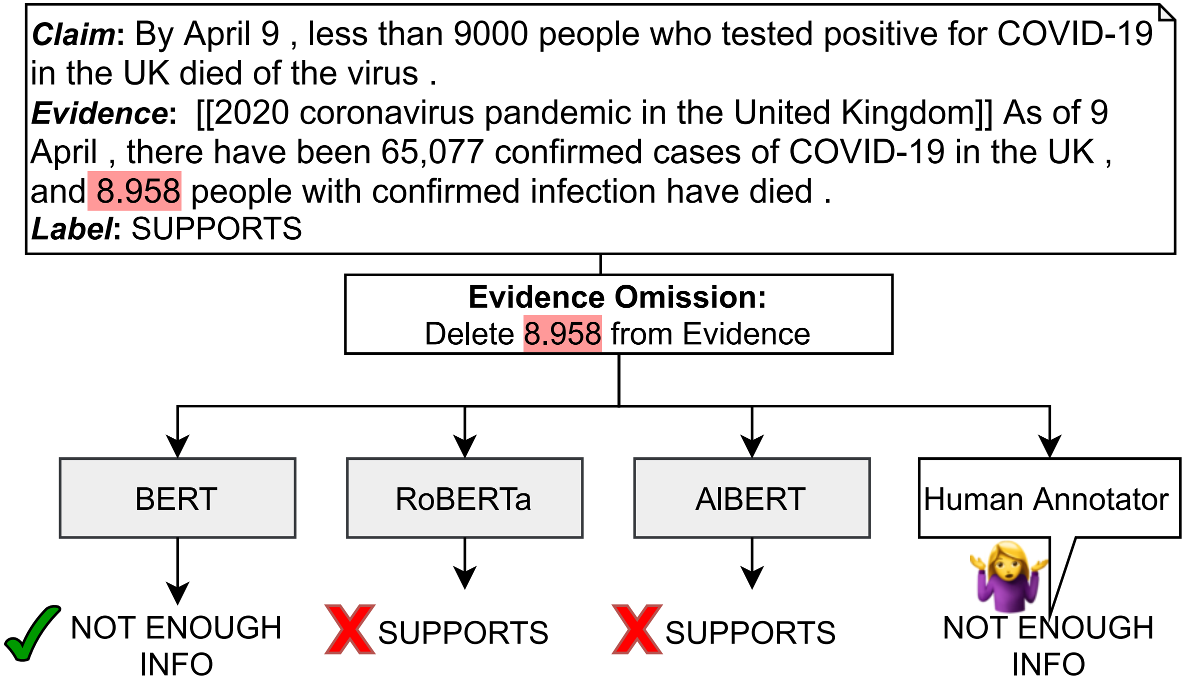
Known Vulnerabilities of Fact Checking Models

- Claim-only bias
  - 61.7 accuracy of a BERT given only the claim
- Lexical biases
  - E.g. negation and the REFUTES class
- Prior proposed solutions:
  - FEVERSymmetric dataset to measure the bias of fact checking models
  - Re-weighted training objective
    - Increase the importance of claims with different labels containing those phrases

Schuster, et al., 2019 "Towards Debiasing Fact Verification Models". In Proceedings of EMNLP 2019.

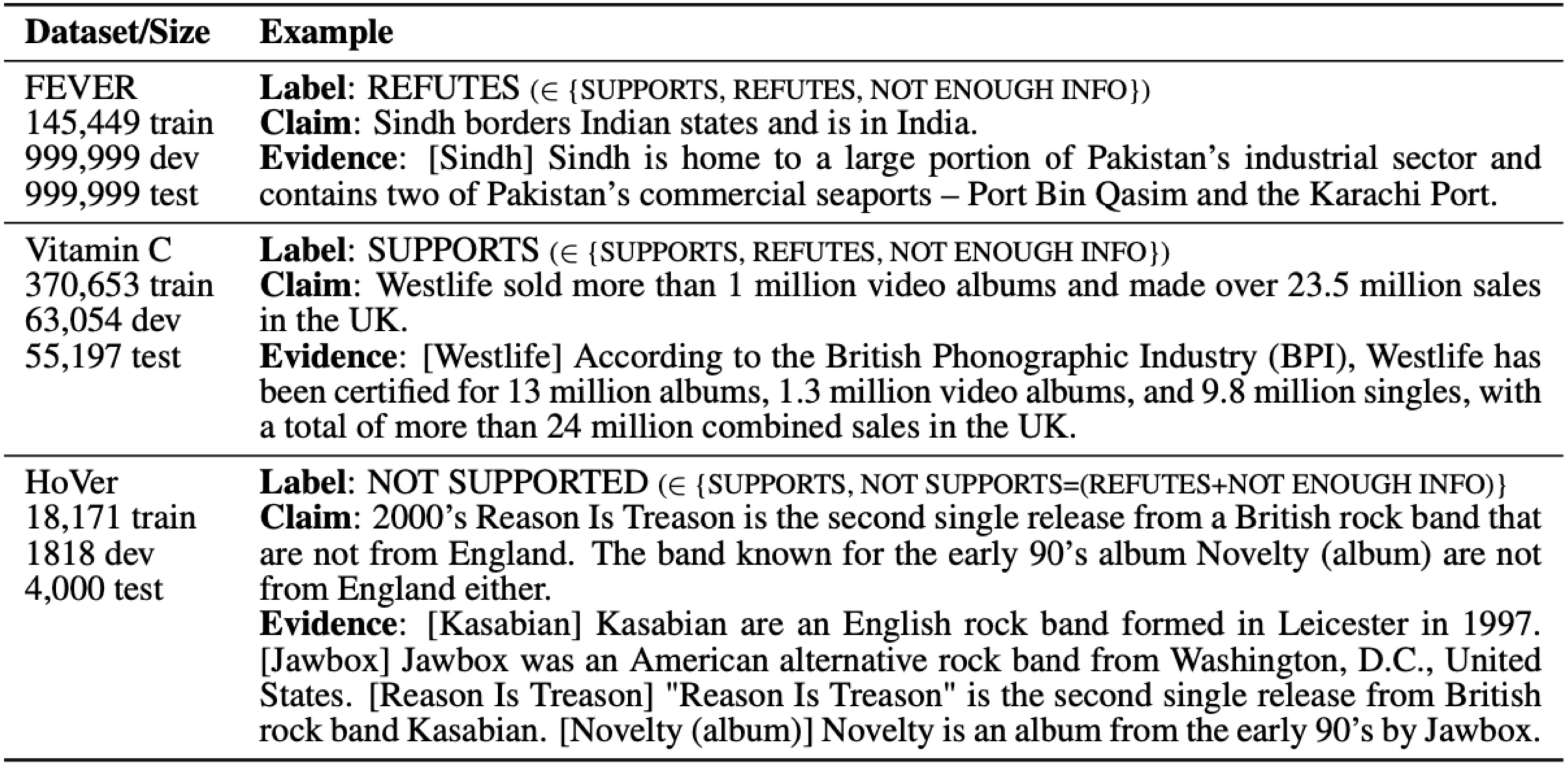
Known Vulnerabilities of Fact Checking Models

- Prediction with insufficient evidence
  - SufficientFacts dataset – fluency-preserving removal of constituents, e.g. adjective/adverb/noun/number modifiers, prepositional phrases
  - Models still predict the original label in most cases:



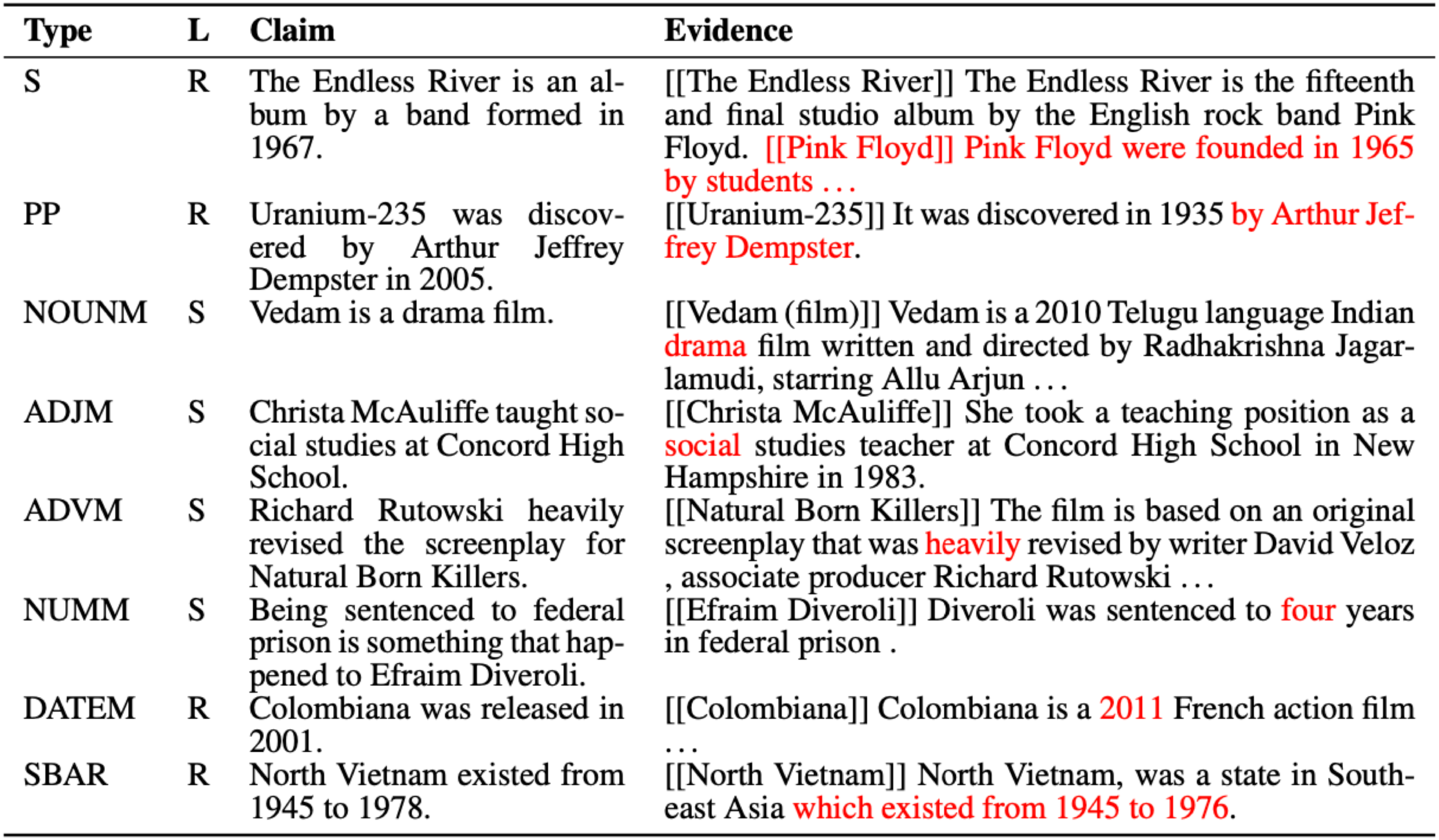
Claim: By April 9 , less than 9000 people who tested positive for COVID-19 in the UK died of the virus .
Evidence: [[2020 coronavirus pandemic in the United Kingdom]] As of 9 April , there have been 65,077 confirmed cases of COVID-19 in the UK , and 8.958 people with confirmed infection have died .
Label: SUPPORTS

Source Datasets



Dataset/Size	Example
FEVER 145,449 train 999,999 dev 999,999 test	**Label**: REFUTES (∈ {SUPPORTS, REFUTES, NOT ENOUGH INFO}) **Claim**: Sindh borders Indian states and is in India. **Evidence**: [Sindh] Sindh is home to a large portion of Pakistan's industrial sector and contains two of Pakistan's commercial seaports – Port Bin Qasim and the Karachi Port.
Vitamin C 370,653 train 63,054 dev 55,197 test	**Label**: SUPPORTS (∈ {SUPPORTS, REFUTES, NOT ENOUGH INFO}) **Claim**: Westlife sold more than 1 million video albums and made over 23.5 million sales in the UK. **Evidence**: [Westlife] According to the British Phonographic Industry (BPI), Westlife has been certified for 13 million albums, 1.3 million video albums, and 9.8 million singles, with a total of more than 24 million combined sales in the UK.
HoVer 18,171 train 1818 dev 4,000 test	**Label**: NOT SUPPORTED (∈ {SUPPORTS, NOT SUPPORTS=(REFUTES+NOT ENOUGH INFO)} **Claim**: 2000's Reason Is Treason is the second single release from a British rock band that are not from England. The band known for the early 90's album Novelty (album) are not from England either. **Evidence**: [Kasabian] Kasabian are an English rock band formed in Leicester in 1997. [Jawbox] Jawbox was an American alternative rock band from Washington, D.C., United States. [Reason Is Treason] "Reason Is Treason" is the second single release from British rock band Kasabian. [Novelty (album)] Novelty is an album from the early 90's by Jawbox.

Evidence Omission



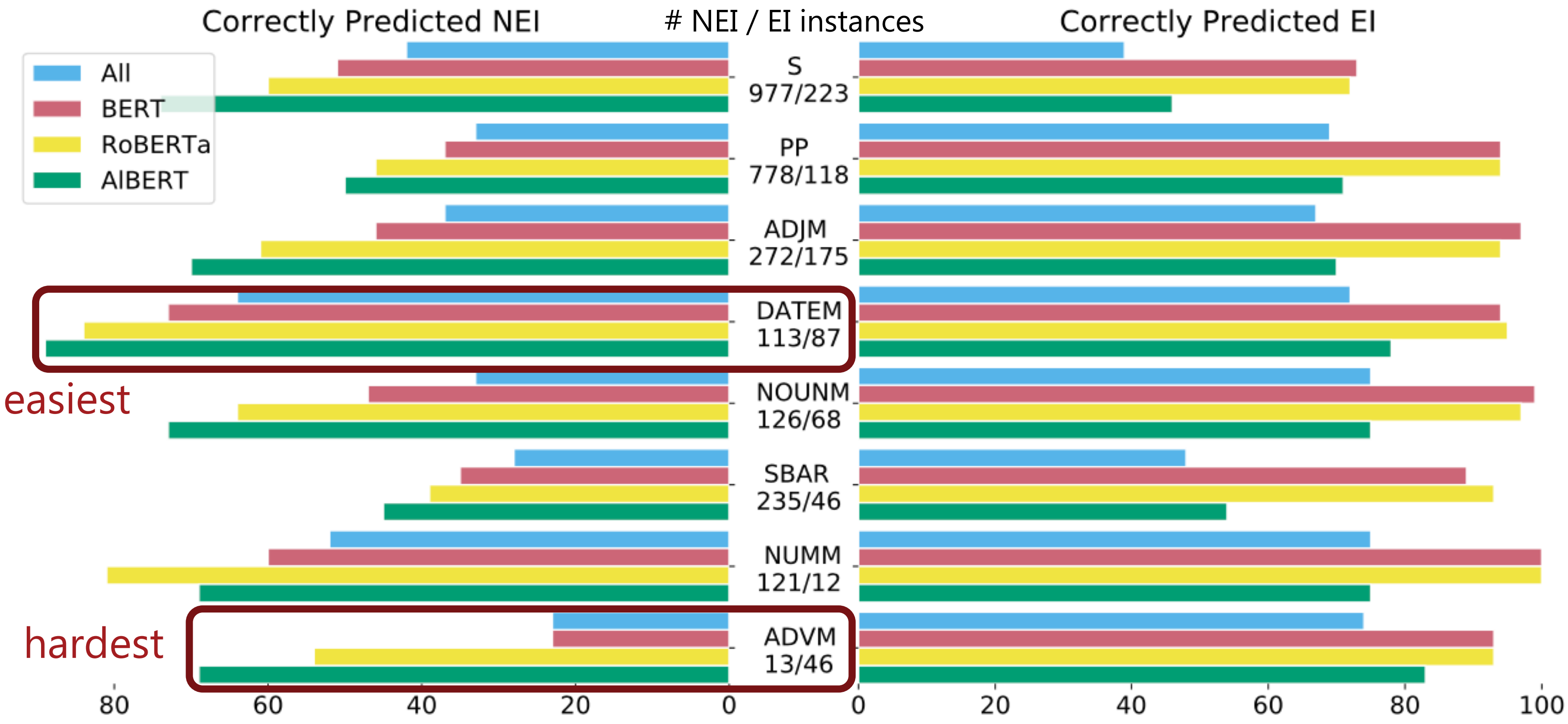
Type	L	Claim	Evidence
S	R	The Endless River is an album by a band formed in 1967.	[[The Endless River]] The Endless River is the fifteenth and final studio album by the English rock band Pink Floyd. [[Pink Floyd]] Pink Floyd were founded in 1965 by students . . .
PP	R	Uranium-235 was discovered by Arthur Jeffrey Dempster in 2005.	[[Uranium-235]] It was discovered in 1935 by Arthur Jeffrey Dempster.
NOUNM	S	Vedam is a drama film.	[[Vedam (film)]] Vedam is a 2010 Telugu language Indian drama film written and directed by Radhakrishna Jagarlamudi, starring Allu Arjun . . .
ADJM	S	Christa McAuliffe taught social studies at Concord High School.	[[Christa McAuliffe]] She took a teaching position as a social studies teacher at Concord High School in New Hampshire in 1983.
ADVM	S	Richard Rutowski heavily revised the screenplay for Natural Born Killers.	[[Natural Born Killers]] The film is based on an original screenplay that was heavily revised by writer David Veloz , associate producer Richard Rutowski . . .
NUMM	S	Being sentenced to federal prison is something that happened to Efraim Diveroli.	[[Efraim Diveroli]] Diveroli was sentenced to four years in federal prison .
DATEM	R	Colombiana was released in 2001.	[[Colombiana]] Colombiana is a 2011 French action film . . .
SBAR	R	North Vietnam existed from 1945 to 1978.	[[North Vietnam]] North Vietnam, was a state in Southeast Asia which existed from 1945 to 1976.

Sentence (S), Prepositional Phrase (PP), Noun Modifier (NOUNM), Adjective Modifier (ADJM), Adverb Modifier (ADVM), Number Modifier (NUMM), Date Modifier (DATEM), Subordinate Clause (SBAR)

SufficientFacts Dataset Construction

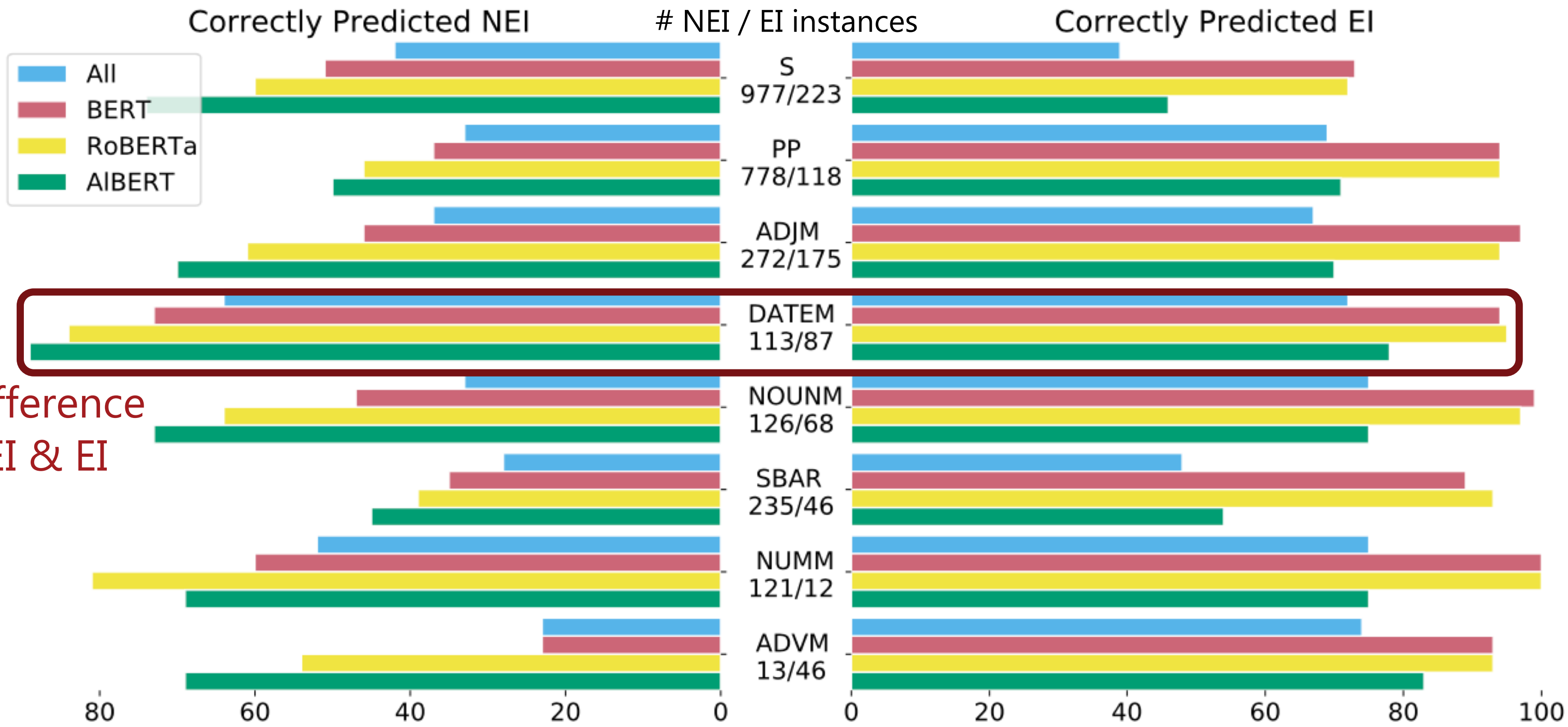
- Ensemble of 3 trained Transformer-based FC models (BERT, RoBERTa, ALBERT)
- Predict labels for instances from FC datasets (FEVER, HoVer, VitaminC) from which information has been removed
- Manually annotate instances for which:
  - Models agree that evidence is sufficient
  - Agree that evidence is insufficient
  - Disagree
- Overall findings from manual annotation
  - Model disagreements are mostly for NEI instances
  - When models agree that information is insufficient, they are correct in 97.2% of cases

Fine-Grained Analysis By Removal Type



All = all models agree on prediction; higher proportion of correct predictions -> better

Fine-Grained Analysis By Removal Type



Difference NEI & EI

All = all models agree on prediction; higher proportion of correct predictions -> better

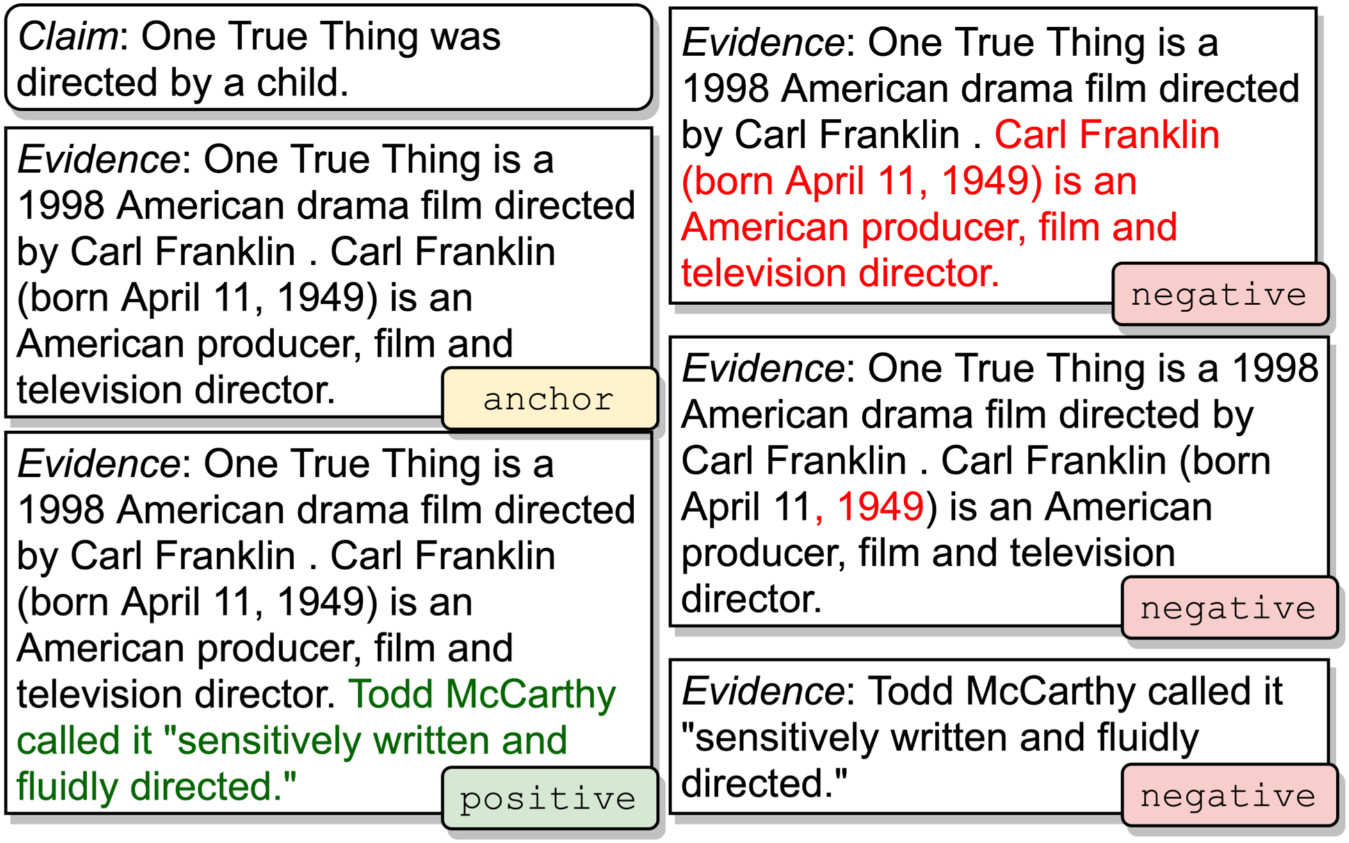
New Task: Evidence Omission Detection

- Given an instance, find a **distractor sentence** from the document of the gold evidence that is most semantically similar to claim (word overlap)
- Append distractor sentence to original evidence texts to serve as anchors
- **Omit information** from original evidence texts (using evidence omission procedure)

New Task: Evidence Omission Detection

- For each of the 3 trained Transformer-based FC models:
  - Select candidates instances for which *the other two supervised models previously predicted NEI*
  - Two types of **positive instances**:
    - Sentences from original evidence without distractor sentence
    - Sentence from original evidence + distractor sentence with highest word overlap with the claim, with one consituent omitted
  - **Negative instances**: original evidence with constituents/sentences removed that all three models predict as NEI

New Task: Evidence Omission Detection



Claim: One True Thing was directed by a child.

Evidence: One True Thing is a 1998 American drama film directed by Carl Franklin . Carl Franklin (born April 11, 1949) is an American producer, film and television director.

Evidence: One True Thing is a 1998 American drama film directed by Carl Franklin . Carl Franklin (born April 11, 1949) is an American producer, film and television director. Todd McCarthy called it "sensitively written and fluidly directed."

Evidence: One True Thing is a 1998 American drama film directed by Carl Franklin . Carl Franklin (born April 11, 1949) is an American producer, film and television director.

Evidence: One True Thing is a 1998 American drama film directed by Carl Franklin . Carl Franklin (born April 11, 1949) is an American producer, film and television director.

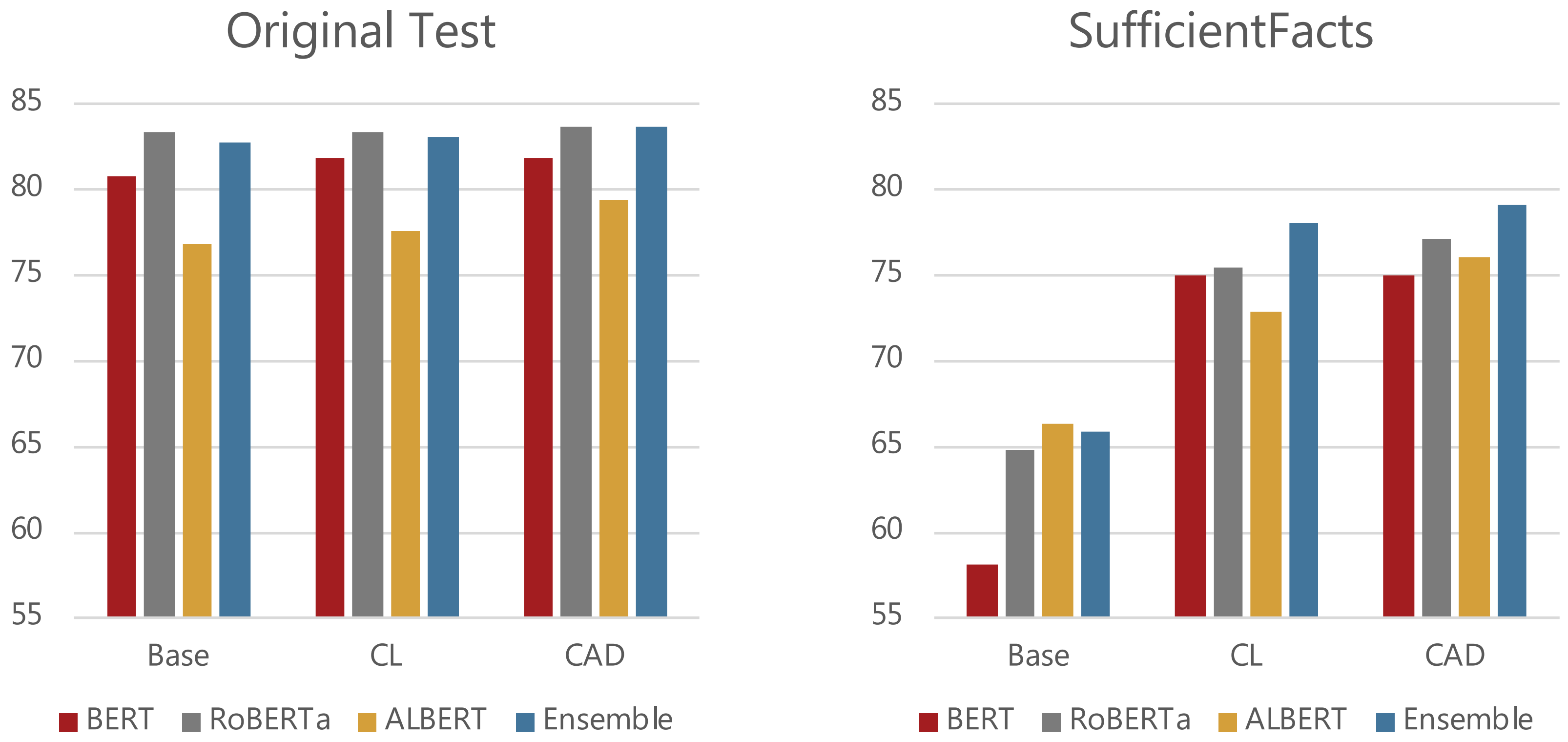
Evidence: Todd McCarthy called it "sensitively written and fluidly directed."

Contrastive instances: negative ones, where models agree that the remaining evidence is insufficient; positive ones, where a distractor sentence with high lexical overlap is added

Fact Checking with Evidence Omission Detection

- Contrastive learning (CL):
  - Auxiliary task:
    - bring the anchor and the positive instance closer in the representation space
    - drive the anchor and the negative instances further apart
- Counterfactual data augmentation (CAD):
  - Augment fact checking datasets with instances from Evidence Omission Detection dataset
  - Treat it as the same task

Selected Results for Fact Checking on HoVer



Summary of Findings

- Comprehensive study of FC with omitted information
  - Fluency-preserving omission methods
- FC models struggle most when adverbial modifiers are removed; least when date modifiers are removed
- Significant differences between Transformer architectures
  - AlBERT better at correctly predicting NEI
  - BERT and RoBERTa better at correctly predicting with EI

Summary of Findings

- SufficientFacts is challenging test bed for FC
  - Performance ~30 F1 points lower than on standard test set
- Training on additional data for evidence omission detection task improves performance on original test set and on SufficientFacts

Wrap-Up

Overall Take-Aways

- **Why** explainability?
  - understanding if a model is right for the right reasons
- Generated explanations can **help users understand**:
  - inner workings of a model (model understanding)
  - how a model arrived at a prediction (decision understanding)
- Explainability can **enable**:
  - transparency for end users
  - human-in-the-loop model development
  - human-in-the-loop data selection

Overall Take-Aways

- **Diagnostic Properties**
  - Enable automatic evaluation of explanations
  - Additional training objectives can improve the properties of the explanations and the general model performance

- **Generating Free-Text Explanations**
  - Complex task; we propose first solutions, framing the task as summarisation
  - How to generate fact checking explanations from evidence documents directly?
  - How to collect a less noisy dataset of real-world fact checks?

Overall Take-Aways

- **Detecting Vulnerabilities of Fact Checking Models**
  - Fact checking models contain many vulnerabilities
  - Easily fooled with lexical trigger words
  - Easily tricked into making predictions without sufficient evidence
  - Training on adversarially generated instances improves robustness
  - Training on instances with omitted information improves robustness
  - How to generate explanations that are useful both for debugging and for the end user?

Thank you!

isabelleaugenstein.github.io
augenstein@di.ku.dk
@IAugenstein
github.com/isabelleaugenstein

Thanks to my PhD students and colleagues!

Pepa Atanasova

Shailza Jolly

Jakob Grue Simonsen

Christina Lioma

CopeNLU

https://copenlu.github.io/

COPENLU
NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP

Presented Papers

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, **Isabelle Augenstein**. *A Diagnostic Study of Explainability Techniques for Text Classification*. In Proceedings of EMNLP 2020.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, **Isabelle Augenstein**. *Diagnostics-Guided Explanation Generation*. In Proceedings of AAAI 2022.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, **Isabelle Augenstein**. *Generating Fact Checking Explanations*. In Proceedings of ACL 2020.

Shailza Jolly, Pepa Atanasova, **Isabelle Augenstein**. *Generating Fluent Fact Checking Explanations with Unsupervised Post-Editing*. Preprint, abs/2112.06924, December 2021.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, **Isabelle Augenstein**. *Fact Checking with Insufficient Evidence*. TACL, Volume 10 (2022), to appear.

Other Recent Relevant Papers

Pepa Atanasova, Dustin Wright, **Isabelle Augenstein**. *Generating Label Cohesive and Well-Formed Adversarial Claims.* In Proceedings of EMNLP 2020.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, **Isabelle Augenstein**, Lucy Lu Wang. *Generating Scientific Claims for Automatic Scientific Fact Checking*. In Proceedings of AAAI 2022.

Dustin Wright, **Isabelle Augenstein**. *Semi-Supervised Exaggeration Detection of Health Science Press Releases*. In Proceedings of EMNLP 2021.

Shailza Jolly, Pepa Atanasova, **Isabelle Augenstein**. *Time-Aware Evidence Ranking for Fact-Checking*. Journal of Web Semantics, Special Issue on Content Credibility, Volume 71, November 2021.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, **Isabelle Augenstein**. *Multi-Hop Fact Checking of Political Claims*. In Procedings of IJCAI 2021.