Accountable and Robust Automatic Fact Checking

Isabelle Augensteinaugenstein@di.ku.dk@IAugensteinhhttp://isabelleaugenstein.github.io/

ECIR
11 April 2022
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.

Types of False Information

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"

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

❤️ 18K ◌ 5.4K ☲ Copy link to Tweet
Types of False Information

**Linkbait Title Generator**

Enter a subject and get link bait title ideas.

- 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

Types of False Information

Mars indfører indrejseforbud efter fund af britisk virusvariant

23 februar, 2021

Udland, Videnskab

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.

Types of False Information

BIRTH CONTROL MAKES WOMEN UNATTRACTIVE AND CRAZY

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?
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"

positive

neutral

negative

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

true

not enough info

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

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

<table>
<thead>
<tr>
<th>Right prediction</th>
<th>Wrong prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Claim:</strong> “In the COVID-19 crisis, ‘only 20% of African Americans had jobs where they could work from home.’”</td>
<td><strong>Claim:</strong> “Children don’t seem to be getting this virus.”</td>
</tr>
<tr>
<td><strong>Evidence:</strong> “20% of black workers said they could work from home in their primary job, compared to 30% of white workers.”</td>
<td><strong>Evidence:</strong> “There have been no reported incidents of infection in children.”</td>
</tr>
<tr>
<td><img src="true.png" alt="True" /></td>
<td><img src="mostly_false.png" alt="Mostly False" /></td>
</tr>
</tbody>
</table>

**Right reasons**

| **Claim:** “Taylor Swift had a fatal car accident.” | **Claim:** “Michael Jackson is still alive, appears in daughter’s selfie.” |
| **Reason:** overfitting to spurious patterns (celebrity death hoaxes are common) | **Reason:** overfitting to spurious patterns (celebrity death hoaxes are common) |
| ![Pants on Fire](pants_on_fire.png) | ![Pants on Fire](pants_on_fire.png) |

**Wrong reasons**
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

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"
positive
neutral
negative

"Immigrants are a drain on the economy", "EU immigrants have a positive impact on public finances"
true
not enough info
false
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

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

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

Explanation via Rationale Selection -- Words

Hypothesis: An adult dressed in black holds a stick.

Premise: An adult is walking away empty-handedly.

Label: contradiction

Post-Hoc Explainability Methods via Rationale Selection

Example from Twitter Sentiment Extraction (TSE) dataset

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?
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

Spider chart for Transformer model on e-SNLI

HA: Agreement with human rationales
F: Faithfulness
RC: Rationale Consistency
DC: Dataset Consistency
CI: Confidence indication
T: Computing Time
Post-Hoc Explainability Methods via Rationale Selection: Aggregated Results

Mean of diagnostic property measures for e-SNLI
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

1. **Claim Check-Worthiness Detection**
   - Claim: "Immigrants are a drain on the economy"
   - Check-worthiness: Not check-worthy

2. **Evidence Document Retrieval**
   - Retrieved evidence:
     - "Immigrants are a drain on the economy"

3. **Stance Detection / Textual Entailment**
   - Stance: Neutral
   - Textual Entailment:
     - "Immigrants are a drain on the economy"
     - "EU immigrants have a positive impact on public finances"

4. **Veracity Prediction**
   - Veracity:
     - True
     - False
     - Not enough info
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

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Explanation Prediction</th>
<th>Target Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td></td>
<td>False p=0.9</td>
</tr>
<tr>
<td>S15</td>
<td>False</td>
<td>0.1</td>
</tr>
<tr>
<td>S16</td>
<td>True</td>
<td>0.3</td>
</tr>
<tr>
<td>S17</td>
<td>True</td>
<td>0.2</td>
</tr>
<tr>
<td>S18</td>
<td>True</td>
<td>0.7</td>
</tr>
<tr>
<td>S19</td>
<td>True</td>
<td>0.5</td>
</tr>
<tr>
<td>S20</td>
<td>True</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Steps 2-4 illustrate the use of Faithfulness, Data Consistency diagnostic properties as well as the Confidence Indication property. Atanasova et al. (2020a) proposed guidance methods to ensure explanation generation models target a teacher model. These methods involve additional learning signals. False confidence from explanation is expected when removing non-selected explanation sentences. Indicates Faithful! Can predict confidence from explanation. Indicates Confidence! Explanation is similar for similar input. Consistent!
Selected Results (MultiRC)

- F1-Classification
- SOTA: 65.5%
- Sup+Data Consistency: 71.7%
- Sup+Faithfulness: 71.0%
- Sup+Confidence Indication: 70.6%
- Sup+All: 70.5%

- F1-Evidence
- SOTA: 51.4%
- Sup+Data Consistency: 78.3%
- Sup+Faithfulness: 79.4%
- Sup+Confidence Indication: 78.6%
- Sup+All: 78.1%

- Acc-Joint
- SOTA: 16.2%
- Sup+Data Consistency: 19.3%
- Sup+Faithfulness: 16.1%
- Sup+Confidence Indication: 16.5%
- Sup+All: 18.8%
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

---

**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.

---

**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

Remove unrelated

Supervised + DC

Supervised
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 Linklater 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

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"

positive

neutral

negative

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

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”
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

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.

**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.

---

Joint Explanation and Veracity Prediction

Which sentences should be selected for the explanation?

\[ \mathcal{L}_{MT} = \gamma \cdot \mathcal{H}(\mathbf{p}^E, \mathbf{y}^E) + \eta \cdot \mathcal{H}(\mathbf{p}^F, \mathbf{y}^F) \]

\[ \mathbf{y}^E \in \{0, 1\}^N \]

\[ \mathbf{y}^F \in Y_F \]

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

Mean Average Rank (MAR).
Lower MAR is better! (higher rank)

Coverage  Non-redundancy  Non-contradiction  Overall
Gold Justification  Extractive  Extractive MT
Explanation Informativeness

Manual veracity labelling, given a particular explanation as percentages of the dis/agreeing annotator predictions.

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
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).

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

1. **Claim Check-Worthiness Detection**
   - "Immigrants are a drain on the economy"

2. **Evidence Document Retrieval**
   - "Immigrants are a drain on the economy"

3. **Stance Detection / Textual Entailment**
   - "Immigrants are a drain on the economy", "EU immigrants have a positive impact on public finances"

4. **Veracity Prediction**
   - "Immigrants are a drain on the economy"
   - true, not check-worthy
   - not enough info, check-worthy
   - neutral, negative
   - false, positive
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

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

  **Evidence Omission:**
  Delete 8,958 from Evidence

  - BERT
    - NOT ENOUGH INFO
  - RoBERTa
    - SUPPORTS
  - AIBERT
    - SUPPORTS
  - Human Annotator
    - NOT ENOUGH INFO
# Source Datasets

<table>
<thead>
<tr>
<th>Dataset/Size</th>
<th>Example</th>
</tr>
</thead>
</table>
| **FEVER** 145,449 train<br>999,999 dev<br>999,999 test | **Label**: REFUTES ($\in\{\text{SUPPORTS, REFUTES, NOT ENOUGH INFO}\}$)  
**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<br>63,054 dev<br>55,197 test | **Label**: SUPPORTS ($\in\{\text{SUPPORTS, REFUTES, NOT ENOUGH INFO}\}$)  
**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<br>1818 dev<br>4,000 test | **Label**: NOT SUPPORTED ($\in\{\text{SUPPORTS, NOT SUPPORTS}=(\text{REFUTES+NOT ENOUGH INFO})\}$)  
**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

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

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

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 constituent omitted
  - Negative instances: original evidence with constituents/sentences removed that all three models predict as NEI
New Task: Evidence Omission Detection

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

Original Test

SufficientFacts
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!

isabelleaugenstei.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/
Presented Papers


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


