Tackling Climate Change with Machine Learning workshop
at NeurIPS 2020

Climate-FEVER: A Dataset for Verification of Real-World Climate Claims

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian,
Massimiliano Ciaramita, Markus Leippold

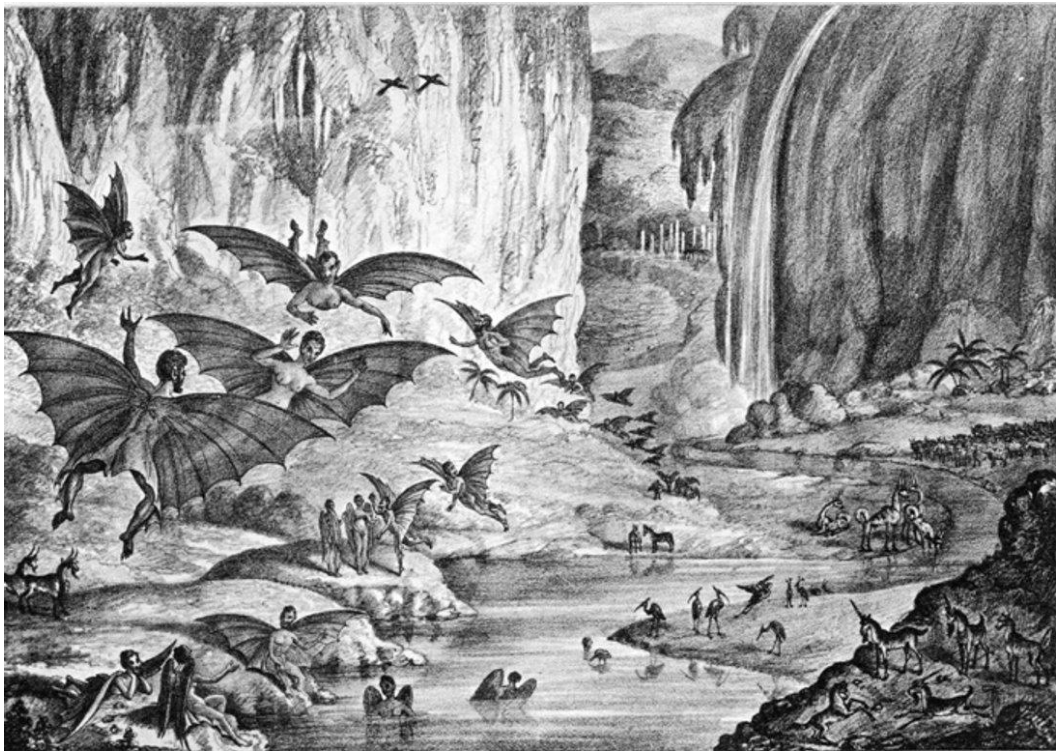


Figure 1: A lithograph of the **Great Moon Hoax**, as printed in **August 1835** (The Sun)
Source: https://en.wikipedia.org/wiki/Great_Moon_Hoax

Climate-related claims

"Doubling the concentration of atmospheric CO2 from its pre-industrial level, in the absence of other forcings and feedbacks, would likely cause a warming of ~0.3°C to 1.1°C."
Source: Craig Idso, Fred Singer, Robert Carter, Heartland Institute, 2017, published at: https://climatefeedback.org/claim-reviews/

"The rate of warming according to the data is much slower than the models used by the IPCC."
Source: Myron Ebell, BBC Newsnight, published at: https://climatefeedback.org/claim-reviews/

Record high snow cover was set in winter 2008/2009.
Source: https://skepticalscience.com/print.php

"Marine life has nothing whatsoever to fear from ocean acidification."
Source: Mike Wallace & James Delingpole, The Spectator, published at: https://climatefeedback.org/claim-reviews/

"[countering climate change:] "Why wait?" (Sadly smiling again) Do you have three more years of quiet life, you live in Switzerland?" – Reto Knutti
Source: https://earth-chronicles.com/natural-catastrophe/swiss-climatologist-scares-the-devastating-weather-anomalies-in-the-coming-years.html

"The concentration of carbon dioxide in Earth's atmosphere has climbed to a level last seen more than 3 million years ago — before humans even appeared on the rocky ball we call home."
Source: https://www.conservation.org/stories/11-climate-change-facts-you-need-to-know

Claim Validation

(Definition)

- Given a claim, provide **evidence** that either **supports** or **refutes** the claim
- Evidence candidates are retrieved from a **Knowledge Document Collection** (KDC)
- KDC is a well-vetted large corpus of **knowledge documents**

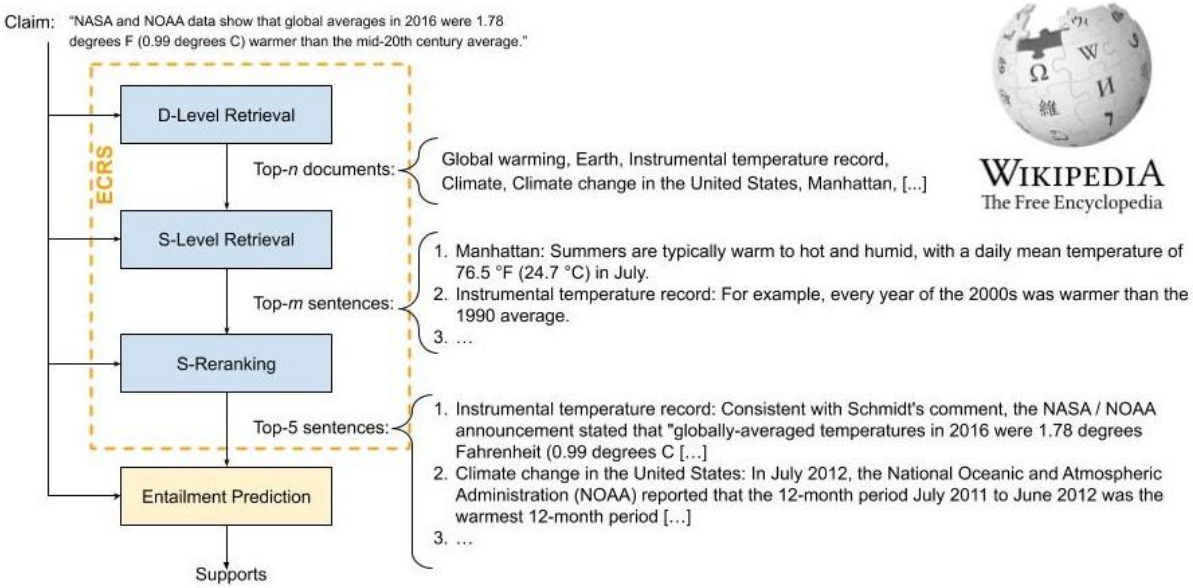
Fact Extraction and VERification (FEVER)

- **Large-scale well-vetted** dataset for claim validation (Thorne et al,. 2018)
- Diverse set of topics
- Claims are **human generated**
- Evidence sentences are from the **introductory section** of all **English Wikipedia** articles (June 2017 Wikipedia dump)

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

Figure 2: Example of a FEVER claim along with the ground-truth evidence set. (Thorne et al., 2018)

Climate-FEVER Pipeline



Building Climate-FEVER

1. Annotation Task 1: Claim Vetting
   - Reject claims containing hate-speech, private information, etc.
   - Reject **non-verifiable** claims
2. Automatic Evidence Candidate Retrieval
   - Given claim, **retrieve top-5 evidence candidates** from Wikipedia
   - Pipeline system, combining LM-based document embedding techniques and relevance prediction
3. Annotation Task 2: Evidence Labelling
   - For each claim, **label a set of five evidence candidates**
   - Each sentence of the set is labelled as supporting, refuting or not giving enough information to validate the claim
   - At least 2 voters per evidence (2.4 ± 0.7)

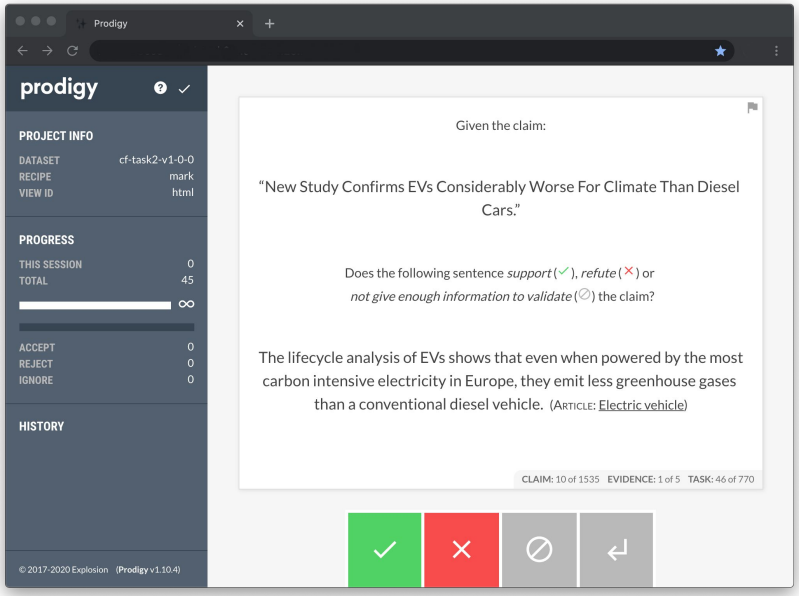


Figure 3: Screenshot of Annotation Task 2 (Evidence Labelling).

Climate-FEVER Overview

- 1,535 climate-change claims
- 7,675 annotated claim-evidence pairs
- Claim-label distribution
  - 655 (42.67%) SUPPORTED
  - 253 (16.5%) REFUTED
  - **153 (9.97%)** **DISPUTED**
  - 474 (30.88%) NOT ENOUGH INFO
- FEVER validator performance (label-acc.)
  - 77.69% on FEVER dev-set
  - 38.78% on CLIMATE-FEVER dataset

Table 3: Shows an overview about the different clusters found in the 1,535 climate claims. The cluster numbers correspond to the cluster numbers in Fig. 2. The first column denotes the cluster, second column shows how many documents belong to said cluster. The last column lists the top-5 terms in the cluster as measured with respect to term-frequency.

cluster	size	top_terms
4	550	co2, temperature, warm, year, increase
2	127	scientist, ipcc, climate, report, new
1	112	global_warming, warming, global, bad, year
3	106	ice, arctic, sea_ice, polar, sea
9	104	climate_change, change, climate, human, cause
6	78	carbon_dioxide, dioxide, carbon, atmosphere, emission
8	72	sea_level, level, sea, level_rise, rise
14	65	carbon, emission, u.s., accord, reduce
10	51	greenhouse_gas, gas, greenhouse, water, emission
16	50	global_warming, global, change, warming, increase
20	32	global_warming, cause_global, warming, global, cause
21	26	term, long_term, long, trend, cool
18	24	ice_sheet, sheet, ice, greenland, antarctica
11	21	extreme, event, weather, bad, change
13	21	review, peer_review, peer, paper, ipcc
7	19	reef, coral, great, world, year
19	19	20th, 20th_century, century, early, temperature
5	19	age, ice_age, ice, little, come
0	14	warming, time, event, trend, ice
12	9	like, temperature, degree, earth, warm
17	9	cloud, sun, predict, likely, global
15	7	large, decline, area, temperature, water

Climate-FEVER Examples

Example 1 (Supported)

“Weather and climate are different; climate predictions do not need weather detail.”

Supports `[wiki/Climate]` *“The difference between climate and weather is usefully summarized by the popular phrase "Climate is what you expect, weather is what you get.”"*

Supports `[wiki/Climate]` *“Climate is the long-term average of weather, typically averaged over a period of 30 years.”*

Supports `[wiki/Weather]` *“Weather refers to day-to-day temperature and precipitation activity, whereas climate is the term for the averaging of atmospheric conditions over longer periods of time.”*

Example 2 (Refuted)

“New Study Confirms EVs Considerably Worse For Climate Than Diesel Cars.”

Refutes `[wiki/Car]` *“Car: "EEA report confirms: electric cars are better for climate and air quality".”*

Refutes `[wiki/Electric_vehicle]` *“Electric vehicle: "UCS: Well-to-wheel, EVs cleaner than pretty much all gas cars".”*

Example 3 (Disputed)

"Positive feedback won't lead to runaway warming; diminishing returns on feedback cycles limit the amplification."

Refutes `[wiki/Greenhouse_gas]` *"Greenhouse gas: Because water vapor is a greenhouse gas, this results in further warming and so is a "positive feedback" that amplifies the original warming."*

Supports `[wiki/Greenhouse_gas]` *"Greenhouse gas: Eventually other earth processes offset these positive feedbacks, stabilizing the global temperature at a new equilibrium and preventing the loss of Earth's water through a Venus-like runaway greenhouse effect."*

Task-2 inter-annotator agreement

Table 1: For evidence candidate labelling (Task-2) the 1,535 claims were split into 10 slices of 770 claims each (except for the last slice). Each slice consists of 3,850 (770×5) claim-evidence pairs. This table lists for each slice the average number of voters, the inter-annotator agreement ($\alpha_{\text{Krippendorff}}$), the fraction of evidence sentences with total agreement and the average entropy with respect to the select class (`SUPPORTS`, `REFUTES` or `NOT_ENOUGH_INFO`).

Slice	Size	Avg. Num. Voters	$\alpha_{\text{Krippendorff}}$	Total Agreement	Avg. Entropy
0	770	2.227	0.283	**0.613**	0.266
1	770	4.019	0.399	0.423	0.380
2	770	2.000	0.522	**0.745**	0.176
3	770	3.001	0.106	0.201	0.544
4	770	2.000	0.215	0.504	0.344
5	770	2.000	0.091	0.404	0.413
6	770	2.000	0.252	0.529	0.327
7	770	2.825	0.316	0.461	0.371
8	770	2.000	0.431	**0.635**	0.253
9	745	2.000	0.229	0.545	0.315