CLIMATE DISINFORMATION TRACKER

AN OPEN-SOURCE TOOL FOR TRACING CLIMATE DENIAL NARRATIVES ON SOCIAL MEDIA

PROBLEM

- False information about climate change threatens environmental safety.
- To address this, we collaborated with the Dutch National Police, to introduce a tool that traces the earliest online origins of climate disinformation and their spread on X, helping journalists, researchers and analysists understand how these narratives begin and evolve.
- Allow Dutch National Police to have a legal pathway to step in and prosecute these organisations for crimes, consequently reducing the spread of disinformation online

RISKS AND ETHICS

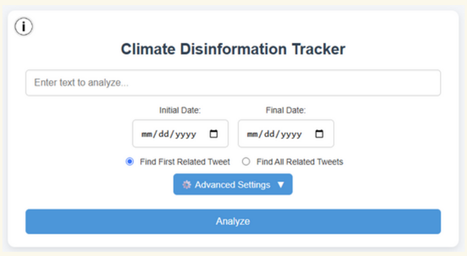
Risks

- Incorrectly identifying 'source' can lead to innocent people being blamed
- Misuse of tool

Ethical considerations

- Privacy vs. transparency
- Freedom of speech
- Usage of Nitter

SOLUTION



1 USER INPUT CLAIM

- Claim can be a climate change hoax, but there are other use cases!
- Date range
- Filters, e.g. retweets, replies

"Climate change is just caused by natural cycles of the sun"

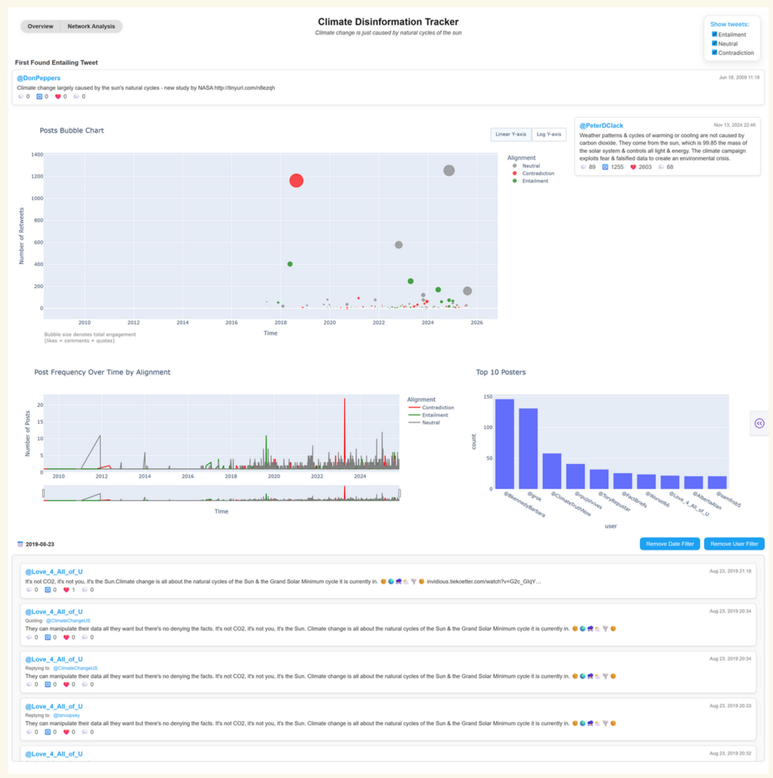
["climate", "caused", "sun", "cycles", "natural"]

2 KEYWORD EXTRACTION AND QUERY GENERATION

Used pre-trained KeyBERT model that sets a max. number of keywords extracted.

The user can select from the suggested synonyms by Wordnet and Spacy or add custom ones.

Boolean query generation with k keywords dropped.

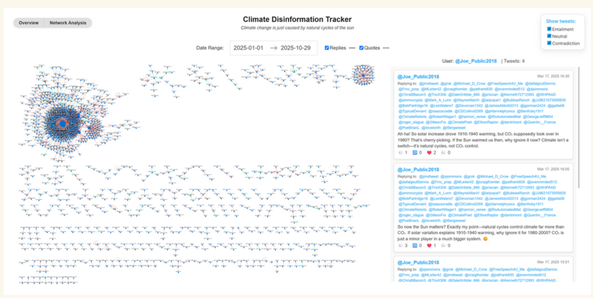


Find all 1: overview. This view shows a bubble chart of time vs engagement, the tweets frequency over time, and top 10 posters. It is possible to filter by date and/or user and by alignment.

3 TWEET RETRIEVAL

Retrieve tweets by crawling and scraping Nitter (open source alternative frontend for X) with Playwright async API and BeautifulSoup using the generated search query.

(climate AND caused AND sun AND cycles) OR (climate AND caused AND sun AND natural) OR ...



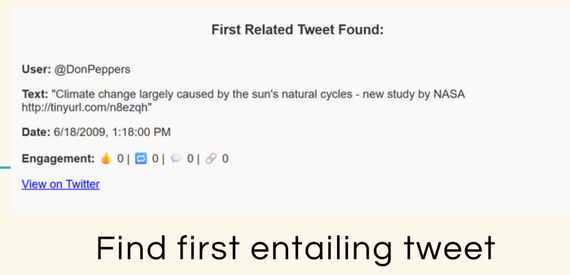
Find all 2: network analysis of the reply/quote interaction between users. It is possible to filter by alignment, date range and interaction type.

4 ALIGNMENT CLASSIFICATION

A retrieved tweet can either support the input claim (entailment), contradict it (contradiction) or is unrelated (neutral).

To determine the alignment we employ the pre-trained model mDeBERTa, which is a multilingual LLM with ~90% accuracy.

5 RESULT DISPLAY



Find first entailing tweet

TESTING AND RESULTS

- Created test cases for checking with and w/o synonyms functionality
- ~70% accuracy on avg
- Main improvement area: alignment model accuracy and keyword sensitivity.

LIMITATIONS

- Dependence on Nitter: non-deterministic and rate limit
- Accuracy bottlenecks of core models
- Query character limit

FUTURE WORK

- Broader media scope
- Deeper network analysis
- Common accounts across different topics
- Hosting tool on a server
- Pilot sessions and workshops

By JIP Group 6.1.1. - Cesar Hernando, Kasper Trouwee, Maria Paula Jimenez Moreno, Manya Atul Narkar, Shirley Li and Vincent van Vliet