FACTors: A New Dataset for Studying the Fact-checking Ecosystem

Enes Altuncu [1] Can Başkent [2] Sanjay Bhattacherjee [1] Shujun Li [1] Dwaipayan Roy [3]

1 University of Kent 2 Middlesex University London 3 IISER KOLKATA

SIGIR 2025 Padova ITALY

Quick Abstract

1. **Why?** Misinformation spreads rapidly online. WEF's 2025 *Global Risks Report* warns it may become the top threat within two years.
2. **What?** FACTors includes 118k English fact-checks (1995–2025) from 39 IFCN/EFCSN-verified organisations, covering ~3k overlapping claims.
3. **How is it different?** Addresses some key limitations of existing datasets:
   - Temporal bias
   - Source selection bias
   - Synthetic data
   - Simplified handling of overlapping claims
4. **How was it constructed?** Collected using custom scrapers, cleaned with NLP methods, and original verdicts normalised to a six-point rating scale.
5. **How can it be used?** Supports ecosystem-wide analysis with three example applications provided:
   - Statistical analysis of the fact-checking ecosystem
   - Political bias detection
   - Credibility assessment of fact-checking organisations
6. **Where can it be accessed?** Publicly available as Lucene index + CSV with author & organisation-level statistics.

Dataset Overview

INTERNATIONAL FACT-CHECKING NETWORK @ Poynter. EFCSN European Fact-Checking Standards Network

- 118,112 fact-checks from 117,993 English reports
- From 39 IFCN/EFCSN signatories, written by 1,953 authors
- Time span: 1995–2025
- 7,327 overlapping claims, referring to 2,977 unique claims
- Lucene index + CSV file
- Publicly available on GitHub

Table 1. Dataset description

Field Name	Description
Row ID	Primary Key
Report ID	ID of each unique report
Claim ID	ID of each unique claim
Claim	Textual claim fact-checked
Content	*(not published to prevent copyright infringement)*
Date published	Report publication date
Author	Report author(s)
Organisation	Publisher fact-checking org.
Original verdict	Fact-check conclusion
Title	Heading of the report
URL	Online link to the report
Normalised rating	Normalised 6-point rating from the original verdict

Raw Data Collection

- 42 IFCN/EFCSN-verified organisations attempted
- 34 IFCN-only, 1 EFCSN-only, 7 verified by both
- Reports scraped with `Scrapy` + `Playright`
- ClaimReview structure considered, when available
- ~140k reports from 39 organisations collected
- Anti-scraping measures, paywall, server connection issues for the remaining three organisations (Reuters, The Washington Post, and Deutsche Welle)

Data Processing & Normalisation

- Cleaning & Preparation
  - Non-English reports detected with `langdetect` and removed
  - Rows with missing or too short values removed
  - Duplicates found with SBERT + cosine similarity
  - 105 repetitive phrases cleaned
  - ⇒ ~22k reports removed
- Verdict Normalisation
  - Mapped 68 original verdicts to a 6-point scale (*true*, *partially true*, *false*, *misleading*, *unverifiable*, and *other*)
  - Fine-tuned a base RoBERTa on 72k samples (acc=0.849)
  - Manually revised low-confidence predictions (<0.5)
- Overlapping Claims Detection
  - SBERT + cosine similarity used
  - Similarity threshold obtained as 0.88 for 95% precision (validated on 1k pairs)

Application 1: Statistical Analysis

Author & organisation-level statistics derived from FACTors

- Fact-checking experience
- Number of fact-checks
- Percentage of unique fact-checks
- Fact-checking rate
- Number of authors per organisation
- Number of organisations per author
- Average word count per author/organisation

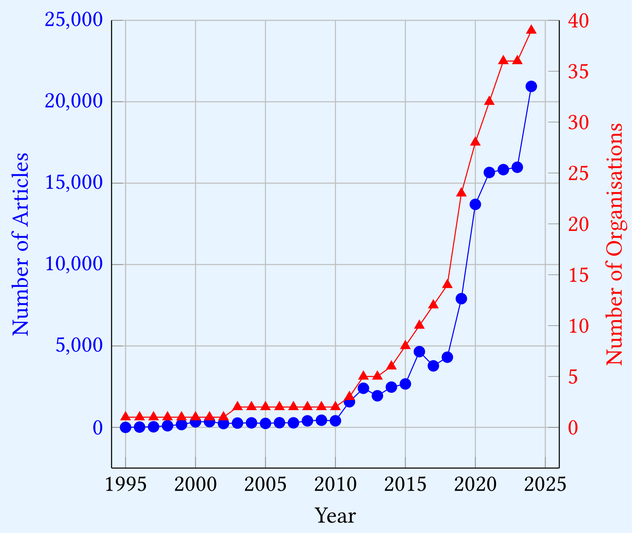

Figure 1. Number of organisations publishing fact-checks per year (red) and total number of fact-checking reports released annually (blue).

Application 2: Political Bias Detection

- Calculated mean & std of bias per organisation with *politicalBiasBERT*
- Many organisations lean left, some lean right

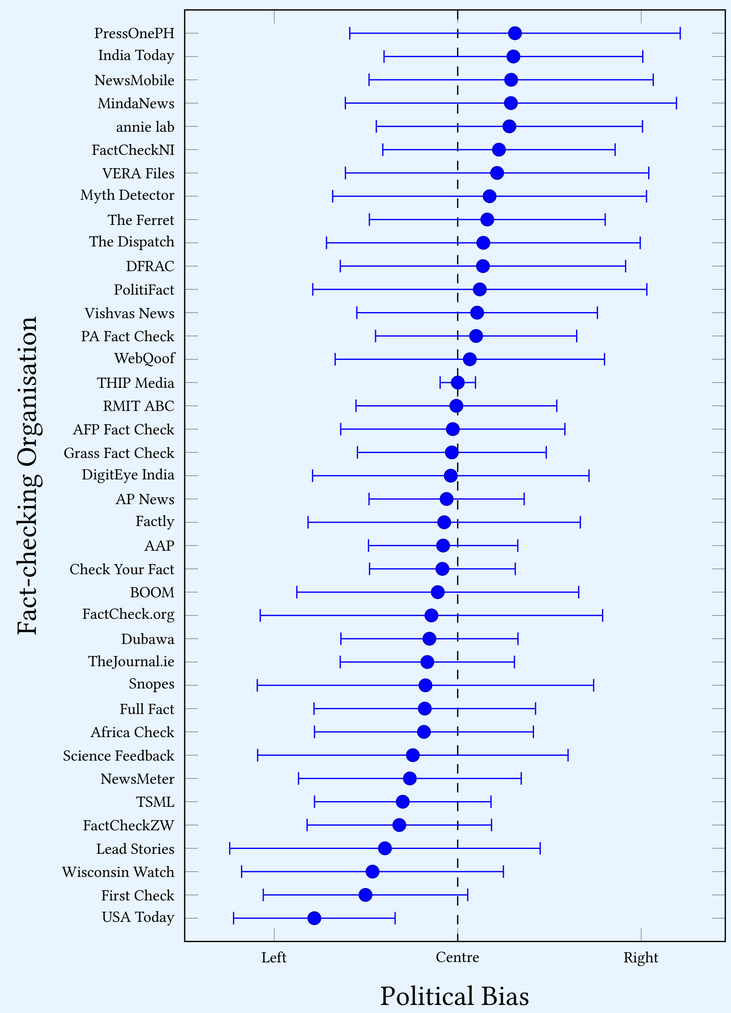

Figure 2. Predicted political bias scores of fact-checking organisations.

Application 3: Credibility Assessment

- Move beyond majority voting in conflicting verdicts
- Use historical data to compute credibility scores

$$\text{CredScore}_i = \frac{1}{m}\sum_{j=1}^{m}\frac{(-1)^{e_j}c_j}{\text{Rank}(i,j)}$$

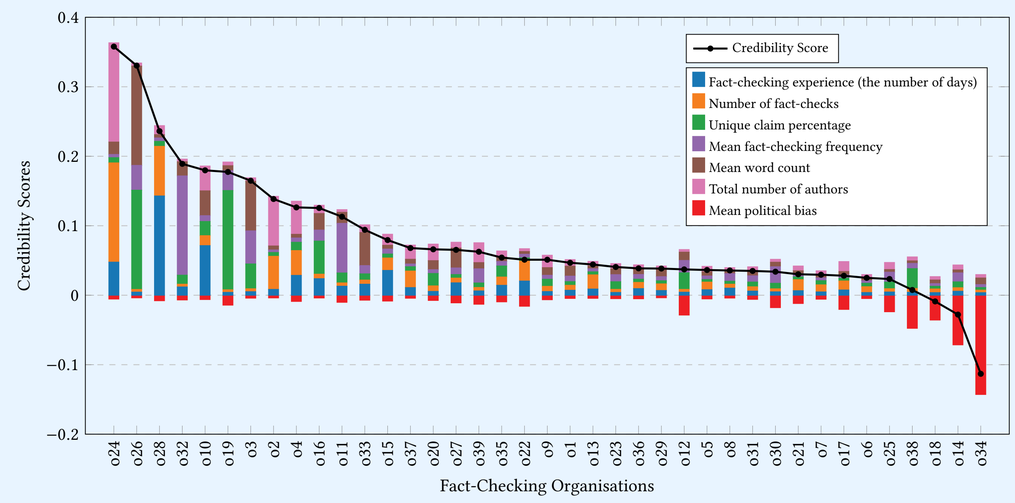

Figure 3. Credibility scores of anonymised organisations with factor contributions

Limitations & Future Work

Limitations

- Only English fact-checks
- Verdict normalisation limited by model accuracy
- Overlapping claims may include false positives/negatives
- Three organisations missing

Future Work

- Multilingual expansion
- Manual verification of normalised verdicts
- Richer metadata, e.g., original claim dates
- Credibility scoring via weighted voting games

drenesaltuncu@gmail.com

SIGIR 2025, Padua, Italy

https://github.com/altuncu/FACTors