X-FACT: A New Benchmark Dataset for Multilingual Fact-Checking

Ashim Gupta, Vivek Srikumar

ACL 2021

1. Key Points

- **Key Contribution**: A multilingual fact-checking benchmark
  - Contains 31,189 claims from 32 websites
  - Claims from 25 typologically diverse languages
  - Generalization evaluation using out-of-domain test set, and zero-shot transfer test set
- **Insights:**
  - Automated Multilingual fact-checking is hard!
  - Models exhibit poor generalization on out-of-domain examples.
  - Poor zero-shot transfer to other languages.

4. Models and Baselines

Three model types:

- **Claim-Only**
  - Rating determined by only using the claim statement.
- **Claim + Metadata**
  - Additional metadata from the fact-check used as `key:value` pairs.
  - Metadata fields: Claimant, Language, Claim Date, Claim Review Date
- **Evidence-based**
  - Uses top-5 evidences retrieved using Google Search with the claim statement.
  - Evidence aggregation using Scaled-Dot Product Attention with evidences and claim statement.
  - Aggregated evidence concatenated with BERT `[CLS]` representation of the claim text.

Majority Baseline: Always predict *False* (the majority class).

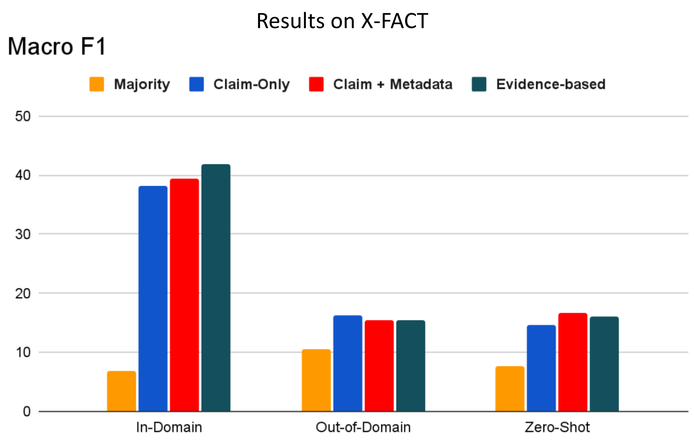
- **Other Details:**
  - Models use the BERT architecture with pretrained multilingual checkpoint (m-BERT).
  - Evaluation Metric: Macro F1 Score
  - Performance reported as average of 5 random seeds.

2. X-FACT Dataset

Claim	*Muslimische Gebete sind Pflichtprogramm an katholischer Schule.* Muslim prayers are compulsory in Catholic schools.
Label	Mostly-False (*Grösstenteils Falsch*)
Claimant	Freie Welt
Language	German
Source	de.correctiv.org
Claim Date	March 16, 2018
Review Date	March 23, 2018
Claim	*Temos, hoje, a despesa de Previdência Social representando 57% do orçamento.* Today, we have Social Security expenses representing 57% of the budget.
Label	Partly-True (*Exagerado*)
Claimant	Henrique Meirelles
Language	Portuguese (Brazilian)
Source	pt.piaui.folha.uol.com.br
Claim Date	None
Review Date	May 2, 2018

- Naturally existing real-world claims in 25 languages
- Three evaluation sets
- In-Domain Test: Language and fact-checker both in training
- Out-of-Domain Test: Language present in training but fact-checker not in training
- Zero-Shot Test: Neither language nor fact-checker in training.

5. Experimental Results



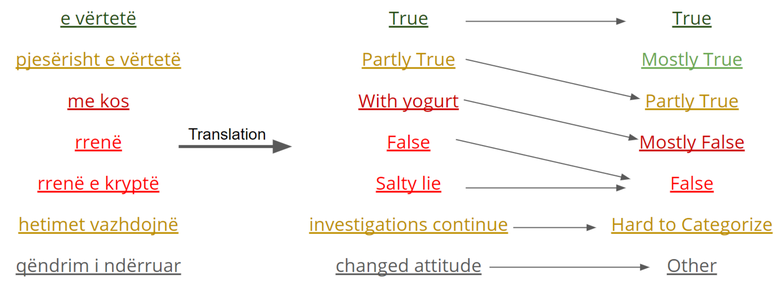
Observations:

- Claim-Only model gets more than 35%, due to existence of artifacts in the data (similar to hypothesis-only bias in Natural Language Inference).
- Augmenting claim meta-data helps the claim-only model.
- Evidence based model performs best among the three models on in-domain evaluation set.

Finding: No model performs well. We need more sophisticated fact-checking models.

3. Dataset Collection

- Claims from International Fact-Checking Network (IFCN) verified fact-checkers
- Two sources: Google Fact Explorer API, Fact-checking websites
- **Challenge**: Different rating scales used by fact-checkers
- **Solution**: Create a new rating scale encompassing all fact-checkers.



Dataset Statistics

Data Split	# Claims	# Langs
Train	19079	13
Development	2535	13
In-Domain	3826	13
Out-of-Domain	2368	4
Zero-Shot	3381	12

Languages

Portuguese	German	Polish	Marathi
Indonesian	Turkish	Bengali	Russian
Romanian	Italian	Dutch	Sinhala
Georgian	Tamil	French	Punjabi
Serbian	Hindi	Gujarati	
Norwegian	Arabic	Persian	• Train Set
Azerbaijani	Albanian	Spanish	• Zero-Shot

6. Analysis

Analysis : Can augmenting English data help? **No**

- Reason: Domain mismatch between English data and multilingual data.

Conclusion:

- Need for better models making effective use of evidence.
- Poor generalization on both out-of-domain evaluation set and zero-shot evaluation set.

Dataset and Code: https://github.com/utahnlp/x-fact
Email: ashimgupta95@gmail.com
Github: ashim95

Let's help fact-checkers with better models!