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

Anonymous ACL-IJCNLP submission

Abstract

In this work, we introduce X-FACT: the largest publicly available multilingual dataset for factual verification of naturally existing real-world claims. The dataset contains short statements in 25 languages and is labeled for veracity by expert fact-checkers. The dataset includes a multilingual evaluation benchmark that measures both out-of-domain generalization, and zero-shot capabilities of the multilingual models. Using state-of-the-art multilingual transformer-based models, we develop several automated fact-checking models that, along with textual claims, make use of additional metadata and evidence from news stories retrieved using a search engine. Empirically, our best model attains an F-score of only 40%, suggesting that our dataset is a challenging benchmark for evaluation of multilingual fact-checking models.

1 Introduction

Curbing the spread of fake news and misinformation on the web has become an important societal challenge. Several fact-checking initiatives, such as the Politifact, expend a significant amount of manual labor to investigate and determine the truthfulness of viral statements made by public figures, organizations, and social media users. Of course, this process is time-consuming and often, a large number of falsified statements go unchecked.

With the aim of assisting fact-checkers, researchers in NLP have sought to develop computational approaches to fact-checking (Vlachos and Riedel, 2014; Wang, 2017; Pérez-Rosas et al., 2018). Many such works use the FEVER dataset, which contains claims extracted from Wikipedia documents (Thorne et al., 2018). Among the works on real-world claims, Wang (2017) introduced LIAR, a dataset with 12,836 claims from politifact.com. Recently, Augenstein et al. (2019) introduced MultiFC, an even larger corpus of 34,918 claims collected from 26 fact-checking websites.

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

Table 1: Examples from X-FACT. Original labels are shown in parenthesis along with the manually mapped labels. For reference, translations are also shown.

Although misinformation transcends countries and languages (Bradshaw and Howard, 2019; Islam et al., 2020), much of the recent work focuses on claims and statements made in English. Developing Automated Fact Checking (AFC) systems in other languages is much more challenging, the primary reason being the absence of a manually annotated benchmark dataset for those languages. Also, there are fewer fact-checkers in these languages, and as a result, a monolingual dataset will be small and less effective in developing fact-checking systems. As recent research points out, a possible solution in dealing with data scarcity is to train multilingual models (Aharoni et al., 2019; Wu and Dredze, 2019; Hu et al., 2020). Indeed, this fin-

ding motivates us to construct a large multilingual resource that the research community can use to further the development of fact-checking systems in languages other than English.

Recent efforts in the construction of a multilingual dataset are limited, both in scope and in size (Shahi and Nandini, 2020; Patwa et al., 2020). For instance, FakeCovid, a dataset introduced by Shahi and Nandini (2020) contains 3066 non-English claims about COVID-19. In comparison, X-FACT contains 31,042 general domain non-English claims from 25 languages. Moreover, their data contains only two labels, namely, *False*, and *Others*. We argue that this is undesirable, as fact checking is a fine-grained classification task. Due to subtle differences in language, most claims are neither entirely true nor entirely false (Rashkin et al., 2017). In contrast, our dataset contains seven labels - we make distinctions between *true*, *mostly true*, *half-true* etc. Table 1 shows two such examples from German and Brazilian Portuguese.

In sum, our contributions are: (i) We release a multilingual fact-checking benchmark X-FACT, which includes 31,042 short statements labeled for factual correctness and covers 25 typologically diverse languages across 11 language families. X-FACT is an order of magnitude larger than any other multilingual dataset available for fact checking. (ii) Apart from the standard test set, we create two additional challenge sets to evaluate fact checking systems' generalization abilities across different domains and languages. (iii) We report results for several modeling approaches and find that these models underperform on all three test sets in our benchmark, suggesting the need for more sophisticated and robust modeling methods.

2 The X-FACT Dataset

X-FACT is constructed from several fact-checking sources. We briefly outline this process here.

Sources of Claims. We relied on a list of non-partisan fact-checkers compiled by International Fact-Checking Network (IFCN)[1], and Duke Reporter's Lab[2]. We removed all the websites that conduct fact-checks in English and are covered by previous work (Wang, 2017; Augenstein et al., 2019). As a starting point, we first queried Google's Fact Check Explorer (GFCE) for all the fact-checks done by a particular website. Then we crawled the linked article on the website and additional metadata such as claimant, URL, date of the claim. For websites not linked through GFCE, we directly crawled all the available fact-checking articles from the fact-checker's website. We left out some fact-checkers because either the claims on their websites were not well specified or the fact-checker did not use any rating scale. We performed semi-automated text processing to remove duplicate claims and examples where the label appeared in the claim itself. This resulted in data from a total of 85 fact checkers for further processing. Refer to the appendix for more details on the this process.

Filtering the Dataset. There are two major challenges in using the crawled data directly: a) the labels are in different languages, and b) each fact checker uses a different rating scale for categorization. To deal with these issues, first, we manually translated all ratings to English, followed by semi-automatic merging of labels if they were found to be synonyms. Second, in consultation with an IFCN signatory, we created a rating scale compatible with most fact-checkers. Our label set contains five labels with a decreasing level of truthfulness: true, mostly true, partly true, mostly false, false. To encompass several other cases where assigning a label is difficult due to lack of evidence or subjective interpretations, we introduced *unverifiable* as another label. A final label *Other* was used to denote cases that do not fall under the above-specified categories. Following the process described, we reviewed each fact-checker's rating system along with some examples and manually mapped these labels to our newly designed label scheme. See table 1 for examples. In our subsequent discussions, we refer to each fact-checking website as a *source*.

We found that the data from several sources was dominated by a single label ($> 80\%$). Since it is difficult to train machine learning models on highly imbalanced datasets, we removed 54 such websites. We additionally removed fact-checking websites that contained fewer than 60 examples. In total, our dataset contains 31,042 fact-checks.

A single test set is not sufficient. Recent advances in NLP have shown that multilingual models are effective for cross-lingual transfer (Kondratyuk and Straka, 2019; Wu and Dredze, 2019; Hu et al., 2020). A multilingual fact-checking system of similar transfer capabilities will certainly be an as-

[1] https://www.poynter.org/ifcn/

[2] https://reporterslab.org/fact-checking/

Data split	# claims	# languages
Train	19005	12
Dev	2506	12
α_1	3781	12
α_2	2369	12
α_3	3381	13

Table 2: Dataset details. Three challenge sets are denoted by α_1, α_2, and α_3

set, especially in languages with no or few fact-checkers. From this perspective, we seek to provide a robust evaluation benchmark that can help researchers understand generalization abilities of their fact-checking systems.

With this objective, we construct three test sets, namely α_1, α_2, and α_3. The first test set (α_1) is distributionally similar to the training set. The α_1 set contains fact-checks from the same languages and sources as the training set.

Second, the *out-of-domain* test set (α_2), contains claims from the same languages as the training set but are from a different source. A model that performs well on both α_1 and α_2 can be presumed to generalize across different source distributions.

Third test set is the *zero-shot* (α_3) set, which seeks to measure the cross-lingual transfer abilities of the fact-checking systems. The α_3 set contains claims from languages not contained in the training set. Models that overfit language-specific artifacts will underperform on α_3.

Languages For training and development, we choose top twelve languages based on the number of labeled examples. Average number of examples per language is 1784, with Serbian being the smallest (835). We split the data into training (75%), development (10%), and α_1 test set (15%). This leaves us with 13 languages for our zero-shot test set (α_3). The remaining set of sources form our out-of-domain test set (α_2). See table 2 for details.

In total, X-FACT covers the following 25 languages (shown with their ISO 639-1 code for brevity): ar, az, bn, de, es, fa, fr, gu, hi, id, it, ka, mr, no, nl, pa, pl, pt, ro, ru, si, sr, sq, ta, tr. Please refer to the appendix for more details.

3 Experiments and Results

3.1 Experimental Setting

The goal of our experiments is to study how different modeling choices address the task of multilingual fact-checking. All our experiments use mBERT, the multilingual variant of BERT (Devlin et al., 2019) and use macro F1 score as the evaluation metric. We implement the following multilingual models as baselines for future work:

1. **Claim Only Model (Claim-Only):** We provide textual claim as the only input to the model, in effect treating the problem as a simple sentence classification problem.

2. **Attention-based Evidence Aggregator (Attn-EA):** Typically, to determine the veracity of a claim, fact-checkers first gather relevant evidence by performing a web search and then aggregate this evidence to reach their final decision. We emulate this procedure by developing an attention-based evidence aggregation model that operates on evidence documents retrieved after performing web search with the claim using Google. For each claim, we obtain the top five results and use them as evidence. Using full text from web pages is not feasible, as the mBERT model has a restricted input sequence length of 512. Following previous work (Augenstein et al., 2019), we use snippets from search results as our evidence.

   For a given claim and a collection of n evidence documents, we first encode the claim and evidences separately using mBERT by extracting the output of the CLS token, denoted as: $\mathbf{c}$, $[\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}]$. We first apply dot-product attention (Luong et al., 2015) to obtain the attention weights $[\alpha_1, \alpha_2, ..., \alpha_n]$, and then compute a linear combination using these attention coefficients: $\mathbf{e} = \sum_i \alpha_i \mathbf{e_i}$. This representation is then concatenated with $\mathbf{c}$ and fed to the classification layer.

3. **Augmenting metadata (+Meta):** We concatenate additional metadata with the claim text by representing it with a string of type: `Key : Value` (Chen et al., 2019).

All the models are trained in a multilingual setting, i.e., a single model is trained for all languages. The trained monolingual models were unstable due to the small size of data for each language.

3.2 Results

The results are shown in table 3. We will discuss results by answering a series of research questions.

Model	α_1	α_2	α_3
Majority	8.1	12.7	8.7
Claim-Only	35.7	**15.3**	**15.9**
Claim + Meta	38.3	14.1	14.9
Attn-EA (Random)	34.3	14.4	**15.8**
Attn-EA	38.7	14.5	14.6
Attn-EA + Meta	**39.9**	14.3	14.8

Table 3: F1 score of the models studied in this work. Attn-EA (Random) denotes the evidence-based model when it is trained with random search snippets.

Model	α_1	α_2	α_3
X-FACT			
Claim + Meta	38.3	14.1	**14.9**
Attn-EA + Meta	**39.9**	14.3	14.8
X-FACT + English			
Claim + Meta	35.1	**15.2**	14.0
Attn-EA + Meta	36.3	14.2	13.1

Table 4: Performance comparison when augmenting the dataset with 12,311 English claims.

Does the dataset exhibit claim-only bias? Before moving to more sophisticated systems, let us first examine if the model can predict a statement's veracity by only using the textual claim. Note that this setting is similar to that of hypothesis only models for the task of NLI (Poliak et al., 2018). From table 3, we see that a claim-only model outperforms a majority baseline by a large margin. We can draw two inferences: a) A significant number of examples in α_1 can be labeled by just relying on the textual claim, and b) the claim-only model has learned spurious correlations from the dataset.

Do search snippets improve fact-checking? First, results from table 3 show that augmenting models with metadata is helpful. Second, using search snippets as evidence with an attention-based model improves performance by 3 points. To further validate that snippets indeed help the evidence-based model, we perform another experiment in which we pair each claim with random search snippets of the same language. Since there are no relevant evidences, the performance is indeed similar to the claim-only model. This again confirms our finding that the dataset exhibits some claim-only bias.

While the Attn-EA model provides some performance improvement on in-domain test set, surprisingly, the claim-only model outperforms the evidence-based model by more than one point on α_3. This might be due to the evidence-based overfitting the in-domain data.

Note that we used snippets to summarize the retrieved search results. To gauge the relevance of these snippets, we manually examine 100 examples from α_1 test set for Hindi. Our preliminary analysis reveals that only 45% of snippets provide sufficient information to classify the claim, indicating why the performance increase with the evidence-based model is small. Our same analysis suggests that for 83% of the examples, using full text of the web pages provides sufficient evidence to determine veracity of the claim. Hypothetically, this means, were the models able to ingest large documents (web pages), their performance increase could have been much more significant.

Do the models generalize across sources and languages? We observe that performance on α_2 and α_3 is worse than on α_1, not only highlighting the difficulty of these challenge sets, but also showing that models overfit both source-specific patterns (α_2) and language-specific patterns(α_3).

Importantly, these results underscore the utility of our challenge sets in assessing model generalizability as well as diagnosing overfitting.

Can we improve performance by augmenting training data with English claims? Since X-FACT does not contain any examples from English, we answer this question by augmenting the training set with 12,311 claims from the Politifact subset of the MultiFC (Augenstein et al., 2019). Results are shown in table 4. Interestingly, we see that augmenting the models with English data significantly hurts model performance. A possible cause is that the augmented data mostly contains political claims, while our dataset contains general claims.

4 Conclusion

We presented X-FACT, the currently largest multilingual dataset for fact-checking. Compared to the prior work, X-FACT is an order of magnitude larger, enabling the exploration of large transformer-based multilingual approaches to fact-checking. We presented results for several multilingual modelling methods and showed that the models find this new dataset challenging. We envision our dataset as an important benchmark in development and evaluation of multilingual approaches to fact-checking.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *EMNLP-IJCNLP*.

Samantha Bradshaw and Philip N Howard. 2019. *The global disinformation order: 2019 global inventory of organised social media manipulation*. Project on Computational Propaganda.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an infodemic: Covid-19 fake news dataset. *arXiv preprint arXiv:2011.03327*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid-a multilingual cross-domain fact check news dataset for covid-19.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

A Details on Dataset Construction

1. As mentioned in the paper, we omit several fact-checking websites from our data. A large number of these websites are not amenable to crawling and scraping the data. For instance, AFP[3] is a prominent fact-checker for many Indo-European Romance languages, but the template on its website does not lend itself to automatic data extraction tools. We can try to access this websites using GFCE, but case many times, the ratings assigned are sentences instead of a single label.

2. Another common reason is that on a number of these websites, the claim statements are not well-specified. Take for example Faktograf[4], a website performing fact-checking in Croatian. On this website , we can neither properly extract the claim statements nor do they clearly mention the rating assigned to the articles.

3. For a small percentage of the claim statements, Google search did not yield any results. We omitted all of these claims from our training, development, and test sets. These are only a very small percentage of claims, so we remove them from all models.

Because of these reasons, a large number of websites in a number of languages could not be crawled.

There are two ways we obtain our claims, labels, and other metadata. One is the Google's Fact Check Explorer (GFCE)[5], and the other is by crawling from the respective fact-checking website. In case, the links are available on GFCE, we download other metadata by visiting the website. Also, we will release the label mapping we created along with the dataset. Appendix A provides more details on the dataset we collected.

B Reproducibility

In this section, we provide details on our hyperparameter settings along with some comments on reproducibility.

[3] https://factuel.afp.com/

[4] https://faktograf.hr/ocjena-tocnosti/

[5] https://toolbox.google.com/factcheck/explorer

Dataset	Model	RunTime
X-FACT	Claim	1.5 hr
X-FACT	Claim+Meta	1.5 hr
X-FACT	Attn-Evd	2.3 hr
X-FACT	Attn-Evd + Meta	2.3 hr
X-FACT + Eng	Claim+Meta	2.5 Hr
X-FACT + Eng	Attn-Evd + Meta	4.1 Hr

Table 5: Average Training time of the models trained

B.1 Models and Code

As described in the main paper, we used multilingual BERT for performing our experiments. We implemented all our models in PyTorch using the transformers library (Wolf et al., 2019).

B.2 Computing Infrastructure Used

All of our experiments required access to GPU accelerators. We ran our experiments on three machines: Nvidia Tesla V100 (16 GB VRAM), Nvidia Tesla P100 (16 GB VRAM), Tesla A100 (40 GB VRAM). Our experiments for the claim-only model were run on V100, and P100 GPUs and evidence-based models required larger VRAM, so they were run on A100 GPUs.

B.3 Hyperparameters and Fine-tuning Details

1. We used the mBERT-*base* model for all of our experiments. This model has 12 layers each with hiddem size of 768 and number of attention heads equal to 12. Total number of parameters in this model is 125 million. We set all the hyper-parameters as suggested by Devlin et al. (2019), except the batch size which is fixed to 8.

2. All our models were run with four random seeds (seed $= [1, 2, 3, 4]$) and the numbers reported in paper are the means of these four runs. We fine-tuned all models for ten epochs and the model performing the best on development set across all epochs was chosen as the final model.

3. Due to constraints on the VRAM of the GPUs, we restricted the number of evidence documents to five.

Average Run times Average training times are presented in table 5.

Language	ISO 639-1 code	FactChecker	Language Family	Train	Dev	α_1	α_2	α_3
Arabic	ar	misbar.com	Afro-Asiatic	✓	✓	✓		
Bengali	bn	dailyo.in	IE: Indo-Aryan					✓
Spanish	es	chequeado.com	IE: Romance	✓	✓	✓		
Persian	fa	factnameh.com	IE: Iranian					✓
Indonesian	id	cekfakta.com	Austronesian	✓	✓	✓		
Indonesian	id	cekfakta.tempo.co	Austronesian				✓	
Italian	it	pagellapolitica.it	IE: Romance	✓	✓	✓		
Italian	it	agi.it	IE: Romance				✓	
Hindi	hi	aajtak.in	IE: Indo-Aryan	✓	✓	✓		
Hindi	hi	hindi.newschecker.in	IE: Indo-Aryan				✓	
Gujarati	gu	gujarati.newschecker.in	IE: Indo-Aryan					✓
Marathi	mr	marathi.newschecker.in	IE: Indo-Aryan					✓
Punjabi	pa	punjabi.newschecker.in.txt	IE: Indo-Aryan					✓
Polish	pl	demagog.org.pl	IE: Slavic	✓	✓	✓		
Portuguese	pt	piaui.folha.uol.com.br	IE: Romance	✓	✓	✓		
Portuguese	pt	poligrafo.sapo.pt	IE: Romance	✓	✓	✓		
Romanian	ro	factual.ro	IE: Romance	✓	✓	✓		
Norwegian	no	faktisk.no	IE: Germanic					✓
Sinhala	si	srilanka.factcrescendo.com	IE					✓
Serbian	sr	istinomer.rs	IE: Slavic	✓	✓	✓		
Tamil	ta	youturn.in	Dravidian	✓	✓	✓		
Albanian	sq	kallxo.com	IE: Albanian					✓
Albanian	sq	faktoje.al	IE: Albanian					✓
Russian	ru	factcheck.kz	IE: Slavic					✓
Turkish	tr	dogrulukpayi.com	Turkic	✓	✓	✓		
Turkish	tr	teyit.org	Turkic				✓	
Azerbaijani	az	faktyoxla.info	Turkic					✓
Portuguese	pt	aosfatos.org	IE: Romance					✓
German	de	correctiv.org	IE: Germanic	✓	✓	✓		
Dutch	nl	nieuwscheckers.nl	IE: Germanic					✓
French	fr	fr.africacheck.org	IE: Romance					✓

Table 6: Details of the X-FACT dataset. Our dataset belongs to 25 typologically diverse languages across 11 language families. The table shows the composition of training, development, and three challenge sets. IE: denotes Indo-Aryan