Multilingual Previously Fact-Checked Claim Retrieval

Anonymous ACL submission

Abstract

Fact-checkers are often hampered by the sheer amount of online content that needs to be fact-checked. NLP can help them by retrieving already existing fact-checks relevant to the content being investigated. This paper introduces a new multilingual dataset for previously fact-checked claim retrieval. We collected 28k posts in 27 languages from social media, 206k fact-checks in 39 languages written by professional fact-checkers, as well as 31k connections between these two groups. This is the most extensive and the most linguistically diverse dataset of this kind to date. We evaluated how different unsupervised methods fare on this dataset and its various dimensions. We show that evaluating such a diverse dataset has its complexities and proper care needs to be taken before interpreting the results. We also evaluated a supervised fine-tuning approach, improving upon the unsupervised method significantly.

1 Introduction

Fact-checking organizations have made progress in recent years in manually and professionally fact-checking viral content (Micallef et al., 2022; Full Fact, 2020). To reduce some of the fact-checkers' manual efforts and make their work more effective, several studies have recently examined their needs and pain points and identified tasks that could be automated (Nakov et al., 2021; Full Fact, 2020; Micallef et al., 2022; Dierickx et al., 2022; Hrckova et al., 2022). These include searching for the source of evidence for verification, searching for other versions of misinformation, and searching within existing fact-checks. These tasks were identified as particularly painful for fact-checkers working in low-resource languages (Hrckova et al., 2022).

We focus on one of these needs – *previously fact-checked claim retrieval* (PFCR) (Shaar et al., 2020). Given a text making an *input claim* (e.g., a social media post) and a set of *fact-checked claims*, our task is to rank the *fact-checked claims* so that those that are the most relevant w.r.t. the *input claim* (and thus the most useful from the fact-checker's perspective) are ranked as high as possible.

Previously, this task was mostly done in English. Other languages that have been considered include Arabic (Nakov et al., 2022), Bengali, Hindi, Malayalam, and Tamil (Kazemi et al., 2021). However, many other languages or even entire major language families have not been considered at all. Additionally, so far only *monolingual PFCR* has been tackled, when the input claim and the fact-checked claims are in the same language. To address these shortcomings, we introduce in this paper a new extensive multilingual dataset. Our two main contributions are:

1. Multilingual dataset for PFCR. We collected and made available[1] a novel multilingual dataset for PFCR. The dataset consists of 205,751 fact-checks in 39 languages and 28,092 social media posts (from now on just *posts*) in 27 languages. For most of these languages, this is the first time this task has been considered at all. This is also the biggest dataset of fact-checks released to date.

All the posts were previously reviewed by professional fact-checkers who also assigned appropriate fact-checks to them. We collected these assignments and gathered 31,305 pairs consisting of a post and a fact-check reviewing the claim made in the post. 4,212 of these pairs are crosslingual (i.e., the language of the fact-check and the language of the post are different). This dataset introduces *crosslingual PFCR* as a new task that has not been tackled before. This is the biggest collection of such pairs that were confirmed by professional fact-checkers.

The dataset also includes OCR transcripts of the images attached to the posts and machine translation of all the data into English.

[1]The dataset will be published at Zenodo after acceptance. Access will be granted upon request *for research purposes only*. We will also release code and detailed results.

2. In-depth multilingual evaluation. We evaluated the performance of various text embedding models and BM25 for both the original multilingual data and their English translations. We describe several pitfalls related to the complexity of evaluating such a linguistically diverse dataset. We also explore the performance across several other data dimensions, such as post length or publication date. Finally, we show that we can improve text embedding methods further by using supervised training with our data.

2 Related Work

Other names are used for PFCR or similar tasks for various reasons, e.g., fact-checking URL recommendation (Vo and Lee, 2018), fact-checked claims detection (Shaar et al., 2020), verified claim retrieval (Barrón-Cedeño et al., 2020), searching for fact-checked information (Vo and Lee, 2020), or claim matching (Kazemi et al., 2021).

Datasets. *CheckThat!* datasets (Barrón-Cedeño et al., 2020; Shaar et al., 2021b) have the most similar collection approach to ours. They collect English and Arabic tweets mentioned in fact-checks to create preliminary pairs and then manually filter them. Compared to this work, we broaden the scope of data collection and omit the manual cleaning in favor of using fact-checkers' reports. Shaar et al. (2020) collected data from fact-checking of English political debates done by fact-checkers. The *CrowdChecked* dataset (Hardalov et al., 2022) was created by searching for fact-check URLs on Twitter and collecting English tweets from retrieved threads. The process is inherently noisy and, the authors propose different noise filtering techniques.

Kazemi et al. (2021) collected several million chat messages from public chat groups and tiplines in English, Bengali, Hindi, Malayalam, and Tamil and 150k fact-checks. Then they sampled roughly 2,300 pairs based on their embedding similarity and manually annotated them. In the end, they obtained only roughly 250 positive pairs. Jiang et al. (2021) matched COVID-19 tweets and 90 COVID-19 claims in a similar manner. Their data could be used for PFCR, but the authors worked on classification instead.

PFCR datasets are summarized in Table 1. Our dataset has the highest number of fact-checked claims. It also has the second-highest number of input claims and pairs after *CrowdChecked*, but that dataset is significantly noisier.

	Input claims	FC claims	Pairs	Languages
Kazemi et al., 2021	*NA*	150,000	258	5
Jiang et al., 2021	*NA*	90	1,573	1
Shaar et al., 2020	*NA*	27,032	1,768	1
Shaar et al., 2021b	2,259	44,164	2,440	2
Hardalov et al., 2022	316,564	10,340	332,660	1
Our Dataset	28,092	205,751	31,305	27/39

Table 1: PFCR datasets. FC claims are *fact-checked*. *NA* means that we were not able to identify the correct number of input claims. The number should be similar to the number of pairs in most cases.

Methods. Methods used for PFCR are usually either BM25 (and other similar information retrieval algorithms) or various text embedding-based approaches (Vo and Lee, 2018; Shaar et al., 2022, 2021a, i.a.). Reranking is often used to combine several methods to side-step compute requirements or as a sort of ensembling (Shaar et al., 2020, i.a.). PFCR task is also a target of the *CLEF's CheckThat!* challenge, with many teams contributing with their solutions (Nakov et al., 2022). Other methods use visual information from images (Mansour et al., 2022; Vo and Lee, 2020), abstractive summarization (Bhatnagar et al., 2022), or key sentence identification (Sheng et al., 2021) to improve the results.

3 Our Dataset

Our dataset consists of fact-checks, social media posts and pairings between them.

Fact-checks. We have collected the majority of fact-checks listed in the Google Fact Check Explorer, as well as fact-checks from additional manually identified major sources (e.g., Snopes) that were missing. Overall, we have collected 205,751 fact-checks from 142 fact-checking organizations covering 39 languages. We publish the *claim, title, publication date, and URL* of each fact-check. We do not publish the full body of the articles. The claim is usually a one sentence long summarization of the information being fact-checked.

Social media posts. We used two ways to find relevant social media posts from Facebook, Instagram and Twitter for the fact-checks. (1) Some fact-checks use the *ClaimReview* schema[2], which has a field for reviewed items. We selected all the links to the social media platforms from this field and

[2] https://schema.org/ClaimReview

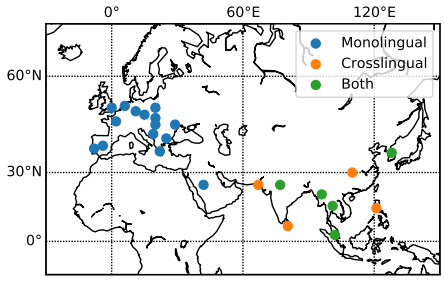

Figure 1: Major languages from our dataset. Crosslingual languages all have English fact-checks.

used them to form the pairs. (2) We searched for appropriate URLs in the main body of fact-check texts and visited the links to Facebook and Instagram. Then, we looked for fact-checking warnings that these two platforms show. These warnings contain links to fact-checking articles, which we used to establish the *pairs*. In both cases, we can be assured that it was professional fact-checkers that assigned the fact-checks to the posts, one way or another. We only processed fact-checks written by AFP news agency[3], though pairs with other fact-checks might have been established from the warnings.

In total, we collected 28,092 posts from 27 languages, as well as 31,305 fact-check-to-post pairs. 26,774 of these pairs are monolingual and 4,212 are crosslingual. Each post in our dataset has at least one fact-check assigned. Figure 1 shows the major (more than 100 samples) languages. All the crosslingual cases have the visualized language for posts and English for fact-checks. We can see that there is a clear distinction between these two groups, probably caused by different fact-checking cultures in different regions.

We publish the *text, OCR of the attached images (if any), publication date, social media platform, and fact-checker's rating* of each post. The *rating* is the reason why the post was flagged (see Section 4.2 for more details). We do not publish URLs in an effort to protect the users and their privacy as much as possible. For detailed information about the implementation of this dataset collection pipeline, see Appendix B. For a more detailed breakdown of dataset statistics (by languages and sources), see Appendix C.

[3]We chose them because they are an established fact-checking organization with high editing standards and are also a part of Meta's *Third-Party Fact-Checking Program*

Dataset versions. We machine-translated all the published texts into English, resulting in two parallel versions of our dataset: the *original version* and the *English version*. We also identified the languages of all the texts. Both translations and language identifications are published as well.

Noise ratio. We manually checked 100 randomly selected pairs from our dataset and evaluated their validity. Three authors rated these pairs and assessed whether the claim from the fact-check was made in the post. In case of disagreement, they discussed the annotation until an agreement was reached. Based on our assessment, 87 out of 100 pairs were correct. The remaining 13 pairs were not errors made by social media platforms or fact-checkers, but rather posts that required visual information (either from video or image) to fully match the assigned fact-check. The Agresti-Coull 95% confidence interval for correct samples in our dataset is 79-92%.

4 Unsupervised Evaluation

We formulate the task we are solving with our dataset as a ranking task, i.e., for each post, the methods rank all the fact-checks. Then, we evaluate the performance based on the rank of the desired fact-checks by using success-at-K (S@K) as the main evaluation metric. We define it in this case as the percentage of pairs when the fact-check ends up in the top K. Throughout the paper, we report this metric with a 95% confidence interval.

For unsupervised evaluation, we evaluated text embedding models and the BM25 algorithm to understand how they are able to handle pairs in different languages or even crosslingual pairs. We were able to gain additional insights into our dataset based on the results as well. Fact-checks are represented with their claims only. Posts are represented with their main texts concatenated with the OCR transcripts.

Text embedding models (TEMs). We use various neural TEMs (Reimers and Gurevych, 2019) that encode texts into a vector space. These are usually based on pre-trained transformer language models fine-tuned as Siamese networks to generate well-performing text embeddings. We use these models to embed both social media posts and fact-checked claims into a common vector space. The retrieval is then reduced to calculating and sorting distances between vectors.

BM25. With BM25 (Robertson and Zaragoza, 2009), we use the posts as queries and fact-checked claims as documents. The score is then calculated based on the lexical overlap between the query and all the documents.

4.1 Main Results

We compare the performance of 15 English TEMs, 5 multilingual TEMs, and BM25. The English TEMs were only evaluated with the English version. The multilingual TEMs and BM25 were evaluated with both the original and the English versions. BM25 with different versions will be denoted as BM25-Original and BM25-English, respectively.

In this section, we use different strategies to evaluate monolingual and crosslingual pairs. For monolingual pairs, we only search within the pool of fact-checks written in the same language as the post (e.g., for a French post we only rank the French fact-checks). For crosslingual pairs, we search in all the fact-checks. In both cases, we report the average performance for individual languages. We only report for languages with more than 100 pairs. For crosslingual pairs, we also consider a separate *Other* category for all the leftover pairs.

We present the main results in Table 2 and we visualize them in Figure 2. We conclude that: **(1)** English TEMs are the best performing option for both monolingual and crosslingual claim retrieval. **(2)** Machine translation significantly improved the performance of both BM25 and TEMs. The difference between the best performing English version method and the best performing original version method is 35% for crosslingual and 14% for monolingual S@10. Currently, machine translation systems also have better language coverage than multilingual TEMs. **(3)** TEMs have a strong correlation between monolingual and crosslingual performance (Pearson's $\rho = 0.98$, $P = 4e{-}10$ for English TEMs). These two capabilities do not conflict. **(4)** There is almost no correlation (Pearson's $\rho = 0.03$, $P = 0.89$ for English TEMs) between model size and performance. The training procedure is much more important. GTR is an exceptionally well-performing family, with all three models being Pareto optimal w.r.t. model size and performance. Another notable model is MiniLM – a surprisingly powerful model for its size (33M).

Even though multilingual TEMs also perform better with the English version, we will report for them the results of the original version from now on to show how the models would perform without using machine translation.

Method	Size [M]	Ver.	Mono	Cross	SLB
BM25					
BM25		En	0.78	0.39	0.18
BM25		Og	0.62	0.06	0.68
English TEMs					
DistilRoBERTa	82	En	0.76	0.43	0.18
GTR-T5-Base	110	En	0.81	0.51	0.19
GTR-T5-Large	336	En	0.83	**0.56**	0.20
GTR-T5-XL	1242	En	**0.83**	**0.56**	0.20
MPNet-Base	109	En	0.78	0.47	0.18
MSMARCO-BERT-Base	109	En	0.78	0.46	0.18
MiniLM-L12	33	En	0.80	0.48	0.18
MultiQA-MPNet-Base	109	En	0.80	0.50	0.18
SGPT-125M	125	En	0.63	0.25	0.13
SGPT-2.7B	2700	En	0.77	0.50	0.19
Sentence-T5-Base	110	En	0.73	0.37	0.14
Sentence-T5-Large	336	En	0.75	0.41	0.15
Sentence-T5-XL	1242	En	0.78	0.46	0.16
Multilingual TEMs					
DistilUSE-Base-Multilingual	135	En	0.74	0.40	0.15
		Og	0.66	0.20	0.16
LaBSE	472	En	0.63	0.22	0.13
		Og	0.69	**0.22**	0.17
MPNet-Base-Multilingual	278	En	0.75	0.40	0.16
		Og	**0.70**	0.21	0.17
MiniLM-L2-Multilingual	118	En	0.74	0.38	0.15
		Og	0.63	0.15	0.17
XLM-R	278	En	0.72	0.33	0.15
		Og	0.66	0.15	0.16

Table 2: Results for methods showing both *mono*lingual and *cross*lingual S@10. Ver. denotes either the original (Og) or the English (En) version of our dataset. The best results for these two versions are bolded. SLB denotes *same language bias*.

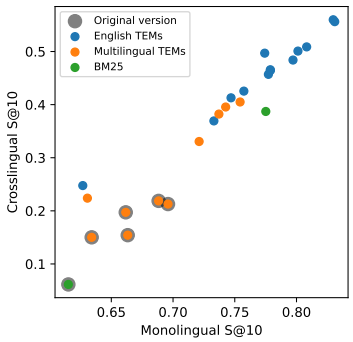

Figure 2: Comparison of different method families.

Languages. Performance for individual languages is shown in Figure 3. We show the results for the best performing TEMs for both versions (**GTR-T5-Large** for the English and **MPNet-Base-Multilingual** for the original, which are denoted as GTR-T5 and MPNet from now on) and both BM25s. We cannot compare the performance across different monolingual languages, since they all use different pools of fact-checks. This is also why smaller languages seem to have better scores.

BM25-Original, despite its seemingly weak overall performance, is actually competitive in some

languages, e.g., Spanish, Portuguese, or Malay. It is better than multilingual TEMs for 7 out of 20 monolingual cases. Its overall monolingual performance is significantly decreased by Thai and Myanmar, due to their use of *scriptio continua*. On the other hand, unlike multilingual TEMs, BM25-Original is not capable of any crosslingual retrieval by design.

False positive rate. We noticed that BM25-Original seems to perform better for languages with larger fact-check pools. We conducted an experiment to measure how pool size affects the results. We randomly selected 100 pairs for 7 of our languages with the largest fact-check pools. We then measured the performance for these 100 pairs while gradually increasing the pool size from 100 to 2,100 by gradually adding random fact-checks.

We found that our initial observation was correct and that BM25-Original performs better than the MPNet model as the pool size increases (especially for Spanish, Portuguese, and French). The relative comparison between BM25 methods and TEMs is shown in Figure 4. This suggests, that MPNet has a higher *false positive rate*, i.e., it is more likely to assign high scores to irrelevant fact-checks. As the number of fact-checks grows, the risk of selecting irrelevant fact-checks also grows. **Different methods may be appropriate for different languages based on the number of fact-checks available.** We did not find the same pattern when comparing the methods using the English version.

Same language bias. The fact that we reduce the fact-checks pool to one language in monolingual evaluation is motivated by what we call *same language bias* (SLB) – a tendency of methods to retrieve fact-checks that have the same language as the post. We approximate SLB by calculating the percentage of top 10 fact-checks that have the same language as the input post. This number is reported in Table 2.

BM25-Original has the highest SLB score of all the methods, and it was the reason we analyzed this score in the first place. BM25-Original has an implicit language filter that effectively reduces the number of fact-checks taken into consideration. This reduction makes the task easier, but it violates our requirement that the method should take all the fact-checks into consideration. We used language-filtered fact-checks in monolingual evaluation to reduce the effect the SLB has on the results. Without this filtering, BM25-Original would clearly outperform MPNet (S@10 51.9 vs 38.5), even though our results in Figure 3 show that for many languages, its language understanding capabilities are actually worse.

However, it is not necessarily true that a higher SLB leads to worse crosslingual performance. As shown in Figure 5, TEMs with the highest SLB actually have the best performance for crosslingual evaluation. Even more strikingly, the relative crosslingual performance compared to monolingual performance increases with SLB as well. We theorize that a certain amount of SLB is healthy, as long as the methods focus on meaningful similarities in texts written in the same language, such as local topics, named entities, and events, rather than on superfluous lexical overlaps. SLB can also be useful to localize claims that are not specific enough. For example, it is impossible to identify the country of origin for the following claim translated to English: *Educational institutions are reopening from January 18*. However, as soon as we use the fact that the original language was Bengali, we can guess that it is about Bangladeshi institutions.

4.2 Other Dimensions

In this section, we report results for various data splits. Since we often work with small splits, we are not able to report the results as an average per language as in the previous section. Instead, we report the average score across the samples. This will give more weight to the more common languages, penalizing the methods with high *false positive rate* (e.g., multilingual TEMs).

Time. We grouped the posts for which we were able to obtain the publication date (N = 26,337) into 20-quantiles and measured the performance of individual methods. The results are shown in Figure 6. There is a visible drop-off for all the methods at the start of 2020, largely caused by the COVID-19 pandemic. We confirmed this by measuring how well the methods worked on posts with the substrings *corona*, *covid* or *korona*.[4] The results are shown in Table 3 (top panel).

The relative differences between individual methods seem stable. We hypothesized that TEMs might have problems with aging, since many of

[4]We chose this as a very simple, high-precision filtering technique. Many other COVID-19-related posts were not retrieved.

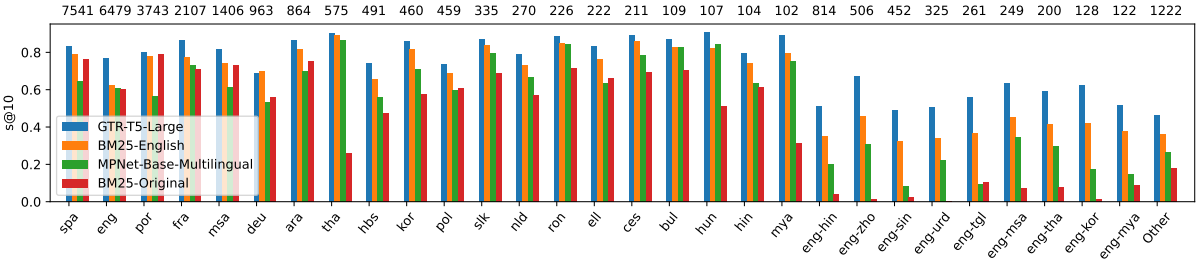

Figure 3: Performance of selected methods for individual languages. For crosslingual pairs (e.g., *eng-hin*), the first language is for the fact-checks and the second is for the posts. The number of pairs is shown at the top.

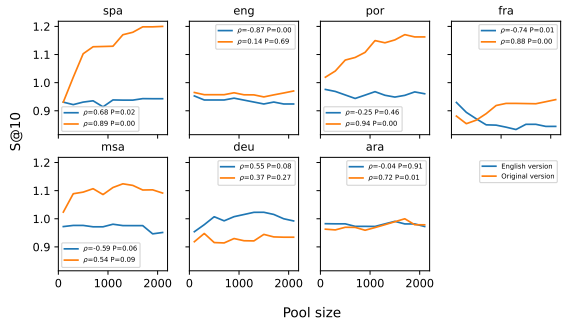

Figure 4: Relative performance (S@10) between BM25 methods and TEMs for different fact-check pool sizes. For both versions we compare the best performing TEMs (GTR-T5 and MPNet) with BM25. Positive ρ means that BM25 gets better with the growing pool size.

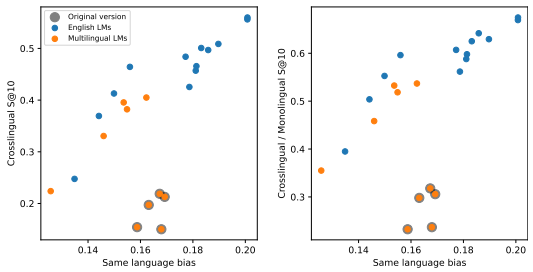

Figure 5: Relation between *same language bias* and performance for TEMs.

the foundation language models were originally trained before 2020. We correlated the average post time for each quantile with the difference between GTR-T5 and BM25-English performance and found a negative, but statistically insignificant correlation (Pearson $\rho = -0.33$, $P = 0.17$ for monolingual S@10). Similar results were measured for crosslingual performance. In both cases, the direction signals that the GTR model is indeed getting worse over time. We found no such signal comparing methods using the original version.

There is a risk that the fact-check was written based on the very post we are using, and an information leak might have happened (e.g., the fact-checker might have used parts of the post verbatim). To test this, we compared pairs where the post is newer with the pairs where the post is older. We found that the two groups have virtually the same performance for all the methods (e.g., 80.02 vs 80.04 monolingual S@10 for GTR-T5). If there is an information leak happening, we were not able to measure it.

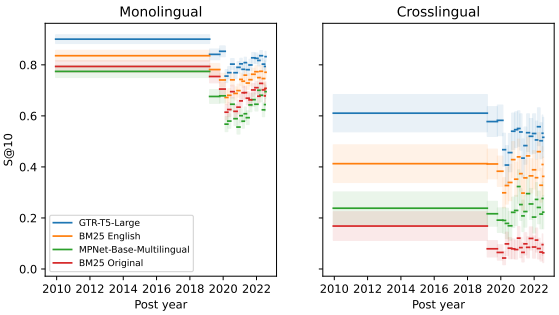

Figure 6: Performance of selected methods for posts from different time intervals. Shaded areas are confidence intervals.

Post rating. In the case of Facebook and Instagram posts, fact-checkers use the so-called *ratings* to describe the type of fallacy present. We show the results for the most common ratings in Table 3 (middle panel). *Missing context* has a slightly lower score than *(Partially) False information*. This might be caused by the fact that the rating is defined by what is *not* written in the post, making it harder to match with an appropriate fact-check. *Altered photo / video* rating has an even lower score. This is an expected behavior, since our purely text-based models cannot handle cases when the crux of the post is in its visual aspect.

Post length. We show how the length of the posts influence the results in Figure 7. In general, the performance peaks at around 500 characters. Posts

	Monolingual					Crosslingual				
	N	GTR-T5	BM25-En	MPNet	BM25-Og	*N*	GTR-T5	BM25-En	MPNet	BM25-Og
COVID-related	4159	0.72 ± 0.01	0.68 ± 0.01	0.50 ± 0.02	0.60 ± 0.01	514	0.40 ± 0.04	0.29 ± 0.04	0.17 ± 0.03	0.06 ± 0.02
Otherwise	22615	0.83 ± 0.00	0.75 ± 0.01	0.66 ± 0.01	0.70 ± 0.01	3698	0.55 ± 0.02	0.39 ± 0.02	0.23 ± 0.01	0.08 ± 0.01
False information	14812	0.82 ± 0.01	0.75 ± 0.01	0.65 ± 0.01	0.69 ± 0.01	2155	0.52 ± 0.02	0.37 ± 0.02	0.22 ± 0.02	0.09 ± 0.01
Partly false information	4498	0.82 ± 0.01	0.75 ± 0.01	0.63 ± 0.01	0.70 ± 0.01	669	0.53 ± 0.04	0.39 ± 0.04	0.21 ± 0.03	0.08 ± 0.02
Missing context	1993	0.77 ± 0.02	0.70 ± 0.02	0.61 ± 0.02	0.63 ± 0.02	268	0.53 ± 0.06	0.35 ± 0.06	0.19 ± 0.05	0.05 ± 0.03
Altered photo/video	753	0.73 ± 0.03	0.66 ± 0.03	0.52 ± 0.04	0.64 ± 0.03	142	0.47 ± 0.08	0.34 ± 0.08	0.17 ± 0.06	0.12 ± 0.05
Facebook	24668	0.81 ± 0.00	0.74 ± 0.01	0.64 ± 0.01	0.68 ± 0.01	3927	0.52 ± 0.02	0.37 ± 0.02	0.22 ± 0.01	0.08 ± 0.01
Instagram	1473	0.78 ± 0.02	0.74 ± 0.02	0.56 ± 0.03	0.75 ± 0.02	44	0.56 ± 0.14	0.37 ± 0.14	0.19 ± 0.11	0.19 ± 0.11
Twitter	682	0.84 ± 0.03	0.74 ± 0.03	0.69 ± 0.03	0.70 ± 0.03	244	0.64 ± 0.06	0.49 ± 0.06	0.38 ± 0.06	0.06 ± 0.03
Total	26774	0.81 ± 0.00	0.74 ± 0.01	0.64 ± 0.01	0.68 ± 0.01	4212	0.53 ± 0.02	0.38 ± 0.01	0.23 ± 0.01	0.08 ± 0.01

Table 3: Performance (S@10) for various splits and methods.

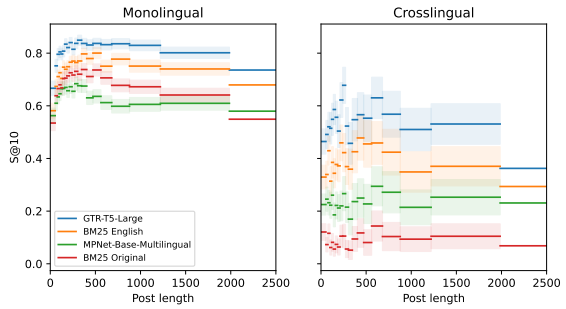

Figure 7: Performance of selected methods for posts with different lengths. Shaded areas are confidence intervals.

that are too short are too difficult to match (and extremely short posts may even indicate noise in the data). On the other hand, for posts longer than 500 characters, the methods gradually lose their effectiveness. The relative performance of methods seems to be relatively stable.

Social media platforms. The results for social media platforms are in Table 3 (bottom pannel). We can see that Twitter has the best performance overall. We believe that this is, to a large extent, caused by the limited length of the posts on the platform.

5 Supervised Training

To validate that our dataset can be used as a training set, we fine-tuned TEMs and evaluated their performance. We split the posts randomly into 80:10:10% train, development, and test sets. We used *cosine* or *contrastive* training losses to fine-tune the models. In both cases, both positive and negative pairs are required for training. We used our data as positive samples and random pairs as negative samples. We performed a hyperparameter search with GTR-T5 and MPNet TEMs (see §D). Here, we report the best performing fine-tuned model we were able to achieve for both TEMs.

The overall results for the test set are reported in Table 4. We can see that GTR-T5 achieved only modest improvements. On the other hand, MPNet improved significantly in both monolingual and crosslingual performance, even surpassing the performance of BM25-English. We observed that the improvements were global across all languages.

We also observed that the TEMs were able to saturate the training set quite quickly, achieving 99.5%+ average precision after only a few epochs. This shows that our naive random selection of negative samples was too easy. The model can learn only a limited amount of information from such samples, and we would need a more elaborate scheme for generating more challenging negative samples. This could lead to further performance improvements.

6 Post-Hoc Results Analysis

The pairs, we obtained from the fact-checks, are only a subset of all the potentially valid pairs. This incompleteness limits our understanding of the dataset and also our evaluation. We decided to manually annotate a subset of the results generated by the methods to better understand what is missing from our data. We generated the top 10 fact-checks for the 87 test set posts that we knew had valid fact-checks (see §3). We used the 4 unsupervised and 2 supervised methods from Section 5.

These methods generated 3,390 unique pair predictions for these 87 posts. Three authors went through each prediction and marked, whether they agreed with it, i.e., whether they found the fact-check to be valid and useful for the post. We consider pairs where at least two annotators agreed to be *correct*. In total, the methods were able to find 719 correct pairs. 96 of these were present in our original dataset. This suggests that there is roughly 7× more pairs in our dataset than we had previously identified. The methods were not

Model	Section 5 (S@10) Monolingual	Section 5 (S@10) Crosslingual	Section 6 (S@10) Our dataset	Section 6 (S@10) Annotated	Section 6 (R@10) Our dataset	Section 6 (R@10) Annotated
Unsupervised						
GTR-T5-Large	0.82 ± 0.01	0.55 ± 0.05	0.7 ± 0.09	0.93 ± 0.05	0.69 ± 0.09	0.59 ± 0.04
BM25-English	0.74 ± 0.02	0.40 ± 0.05	0.67 ± 0.10	0.85 ± 0.07	0.67 ± 0.09	0.48 ± 0.04
MPNet-Base-Multilingual	0.63 ± 0.02	0.23 ± 0.04	0.51 ± 0.10	0.7 ± 0.09	0.47 ± 0.09	0.32 ± 0.03
BM25 Original	0.68 ± 0.02	0.09 ± 0.03	0.6 ± 0.10	0.71 ± 0.09	0.58 ± 0.09	0.26 ± 0.03
Supervised						
GTR-T5-Large	0.84 ± 0.01	0.59 ± 0.05	0.71 ± 0.09	0.92 ± 0.05	0.7 ± 0.09	0.65 ± 0.03
MPNet-Base-Multilingual	0.76 ± 0.02	0.42 ± 0.05	0.62 ± 0.10	0.85 ± 0.07	0.6 ± 0.09	0.45 ± 0.04

Table 4: Test set performance (§5) and annotated results performance (§6) of unsupervised and supervised methods.

able to find 9 fact-checks out of 105 that were already in our dataset. Of the 719 correct pairs, only 247 were monolingual, 136 were crosslingual with an English fact-check, and 336 were crosslingual with a non-English fact-check. The last category in particular is almost completely missing from our dataset.

In Table 4, we show the results for individual methods. We compare S@10 (now defined as how many posts have at least one correct fact-check produced) as approximated with our dataset and the true S@10 obtained by the annotation. We can see that the score for our dataset is significantly lower and true performance of our methods is better then what was measured previously. We also compare recall-at-10 (R@10), defined as the percentage of expected pairs a method was able to produce in the top 10. In this case, both our dataset and manual annotation are only estimates, since they do not contain *all* the valid pairs, they both contain only a subset obtained by different methods. Here we can see that our dataset actually provides higher estimates. We assume that our annotation is more precise, so we conclude that the recall calculated from our dataset is overinflated (possibly due to selection bias). **It also seems that our dataset has a bias in favor of BM25**, compared to the results obtained from annotated data.

7 Discussion

Complexity of crosslingual evaluation. Phenomena such as *same language bias* or *false positive rate* make the evaluation of multilingual and crosslingual datasets inherently complex. If we were to abstract the whole evaluation into a single number, as is often done in practice, we would have completely missed these pitfalls. Without an in-depth evaluation, we might have been misled while applying our methods in practice, e.g., while developing helpful tools for fact-checkers. Our evaluation procedures were previously impossible to develop in the absence of linguistically diverse PFCR datasets.

Machine translation beats multilingual TEMs. These two technologies represent the two main multilingual and crosslingual learning paradigms – label transfer and parameter transfer (Pikuliak et al., 2021). Machine translation is a clear winner in our case. English TEMs significantly outperform multilingual approaches for both monolingual and crosslingual retrieval.

COVID-19. As shown in Table 3, it seems that the performance for COVID-19 is significantly worse than for the rest of the dataset. However, this might not necessarily mean that the methods are having issues with the domain shift. The sheer amount of fact-checks written about COVID-19 makes it hard for the methods to pick the desired fact-check in the presence of thousands of other very similar ones. This is evident considering that BM25 also has worse results, even though it should be less prone to domain shift based on its design.

8 Conclusions

In this paper, we introduced a new multilingual *previously fact-checked claim retrieval* dataset. Our collection process yielded a unique and diverse dataset with a relatively small amount of noise in it. We believe that the evaluation of various methods is also insightful and can lead to the development of better fact-checking tools in the future. We summarize the limitations of our work discussed throughout the paper in Appendix E.

We believe that our dataset opens up many interesting research directions. We have barely scraped the surface of crosslingual learning in this work. Applying various transfer learning methods (especially for low-resource languages) is an important future direction.

9 Ethical Considerations

We analyzed the likelihood and impact of ethical and societal risks for the most affected stakeholders, such as social media users and profile owners, fact-checkers, researchers, or social media platforms. For the most severe risks, we proposed respective countermeasures, following the guidelines and arguments in (Franzke et al., 2020; Townsend and Wallace, 2016; Mancosu and Vegetti, 2020).

Data collection process. While Twitter posts were collected using a publicly available API, the Terms of Service (ToS) of Facebook and Instagram do not currently allow for the accessing or collecting of data using automated means. To minimize the harm to these social media platforms and their users, we made sure to only collect publicly available posts that are accessible even without logging in. This complies with the ToS.

Even if we admit the risk that such research activities could potentially violate the ToS, we argue that ignoring posts from Facebook and Instagram would prohibit research that seeks to address key current issues such as disinformation on these platforms (Bruns, 2019). These are some of the main platforms for disinformation dissemination in many countries. We consider the collection of such public data and its usage for research purposes to be in the public interest, especially considering the status of disinformation as a hybrid security threat (ENISA, 2022), which could justify minor harms to social media platforms.

Other considerable risks include the risk of accessibility privacy intrusion (Tavani, 2016) of social media users by observing them in an environment where they do not want to be observed. We did not obtain explicit consent from social media users to collect their posts. However, the criteria for considering social media data private or public depend on the assumption of whether social media users can reasonably expect to be observed by strangers (Townsend and Wallace, 2016). Twitter is considered an open platform. The collected posts on Facebook or Instagram are not only public, but the users can also expect that their posts will be widely shared, commented or reacted to and they can end up being fact-checked if it is the case.

Data publication. To minimize the risk of third-party misuse, the dataset is available only to researchers for research purposes. The full texts of the fact-checks are not published to avoid possible copyright violations.

Automatic translation has the risk of unintentional harm from misinterpretation of the original claims. To counter this risk, we always provide the original text as well.

We assessed the risk of re-identification, as well as the risk of revealing incorrect, highly sensitive or offensive content regarding social media users. At the same time, we had to take into account the fact that social media platforms remove some posts after they have been flagged as disinformation. Therefore, we decided to include the original texts of the posts in the dataset to prevent it from decaying. Otherwise, it would become progressively less usable and research based on it less reproducible. This also allows us avoid publishing the URLs of posts, which would directly reveal the identities of the users. It is not possible to guarantee complete anonymity, since the posts are still linked in the fact-checks. The posts could also theoretically be found by full-text search.

On the other hand, all the posts released in our dataset are already mentioned in a publicly available space in the context of fact-checking efforts. Our publication of these posts does not significantly increase their already existing public exposure, especially considering the limited access options of our dataset.

To support users' rights to rectification and erasure in case of the publication of incorrect or sensitive information, we provide a procedure for them to request the removal of their posts from the dataset. However, we assess that the risk of wrongfully assigned fact-checks has a low probability (see §3).

As the dataset can also be used for supervised training (see §5), there is a risk of propagating biases present in the data (see §E). We recommend performing a proper linguistic analysis of any supervised model w.r.t. all the languages for which the model is intended. The results shown in this paper may not reflect the performance of the methods on other languages. We are also aware of the risk of propagating the biaseas of the fact-checkers, as it is they who decide what to fact-check. Although they should generally follow principles of fact-checking ethics (see, e.g., the IFCN's Code of Principles), there may still be present some human or systemic biases (Schwartz et al., 2022) that could affect the results when using the dataset for other purposes.

References

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 12260 of *Lecture Notes in Computer Science*, pages 215–236, Cham. Springer International Publishing.

Varad Bhatnagar, Diptesh Kanojia, and Kameswari Chebrolu. 2022. Harnessing abstractive summarization for fact-checked claim detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2934–2945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Axel Bruns. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11):1544–1566.

Laurence Dierickx, Ghazaal Sheikhi, Duc Tien Dang Nguyen, and Carl-Gustav Lindén. 2022. Report on the user needs of fact-checkers. Technical report, NORDIS – NORdic observatory for digital media and information DISorders.

ENISA. 2022. ENISA Threat Landscape 2022.

Aline Shakti Franzke, Anja Bechmann, Michael Zimmer, Charles Ess, and the Association of Internet Researchers. 2020. Internet research: Ethical guidelines 3.0.

Full Fact. 2020. The challenges of online fact checking.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs.CL].

Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. Crowd-Checked: Detecting previously fact-checked claims in social media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 266–285, Online only. Association for Computational Linguistics.

Andrea Hrckova, Robert Moro, Ivan Srba, Jakub Simko, and Maria Bielikova. 2022. Automated, not automatic: Needs and practices in European fact-checking organizations as a basis for designing human-centered AI systems. arXiv:2211.12143 [cs.CY].

Ye Jiang, Xingyi Song, Carolina Scarton, Ahmet Aker, and Kalina Bontcheva. 2021. Categorising Fine-to-Coarse Grained Misinformation: An Empirical Study of COVID-19 Infodemic. arXiv:2106.11702 [cs].

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517, Online. Association for Computational Linguistics.

Moreno Mancosu and Federico Vegetti. 2020. What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media + Society*, 6(3). SAGE Publications Ltd.

Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2022. Did I See It Before? Detecting Previously-Checked Claims over Twitter. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 367–381, Cham. Springer International Publishing.

Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or false: Studying the work practices of professional fact-checkers. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, Michael Wiegand, Melanie Siegel, and Juliane Köhler. 2022. Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 infodemic and fake news detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 495–520, Cham. Springer International Publishing.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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

3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. Towards a standard for identifying and managing bias in artificial intelligence. NIST Special Publication 1270, National Institute of Standards and Technology.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The role of context in detecting previously fact-checked claims. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1619–1631, Seattle, United States. Association for Computational Linguistics.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2021a. Assisting the Human Fact-Checkers: Detecting All Previously Fact-Checked Claims in a Document. arXiv:2109.07410 [cs.CL].

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Shaden Shaar, Fatima Haouari, Watheq Mansour, Maram Hasanain, Nikolay Babulkov, Firoj Alam, Giovanni Da San Martino, Tamer Elsayed, and Preslav Nakov. 2021b. Overview of the CLEF-2021 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims in Tweets and Political Debates. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, volume 2936. CEUR-WS.

Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

H.T. Tavani. 2016. *Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing*, 5th edition. Wiley.

Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics.

Nguyen Vo and Kyumin Lee. 2018. The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 275–284, New York, NY, USA. Association for Computing Machinery.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? Searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

A Computational Resources

We calculated all the results on an AWS-based virtual machine located in the Ohio AWS data center. The machine has one NVIDIA Tesla T4 GPU installed. The unsupervised experiments would take approximately 2 GPU days to replicate. The supervised experiments would take approximately 3 GPU days to replicate. Additional roughly 4 GPU days were spent on other experiments that were discarded or are not reported in this paper.

B Dataset Pipeline Details

B.1 Dataset Collection

Archiving services. Since the content from social media networks may disappear in time, fact-checkers tend to use various content archiving services (e.g., `perma.cc`). We extract the content from these services as well.

AI APIs. We use following services to process our samples:

- *Google Vision API.* We use Google Vision API to extract text from images attached to the post. The API also returns a list of languages found in each image with their percentage.

- *Google Translate API.* We use Google Translate API to translate all the texts into English. The API also returns a most probable language.

B.2 Dataset Pre-Processing

We performed several cleaning and pre-processing steps with our dataset. All the pre-processing is available in the released code repository.

Removing noisy claims. We removed fact-checks that had no claim or where the claim was shorter than 10 characters.

Fact-check deduplication. We unified fact-checks with identical claims.

Noise in social media posts. We removed texts or OCR transcripts that we deemed noisy (shorter than 25 characters or more than 50% non-alphabetical characters). We then only kept posts where at least one text was considered not noisy. We also removed noisy lines from OCR transcripts (Lines shorter than 5 characters or with more than 50% non-alphabetical characters). We also removed URLs.

Post deduplication. We unified posts that ended up with identical text contents after the cleaning process.

Machine translation. We translated all the texts into English. The only exceptions were fact-check claims coming from English-language providers (e.g., Snopes) that we considered English by default, and fact-check claims where CLD3[5] identified English language. We confirmed experimentally that CLD3 has a high precision on English texts.

Language identification normalization. We observed that there are some systematic errors in the language identification models we used. We found out that the model often selected less common languages based on spurious patterns, e.g., mentions of Filipino politicians sometimes led to Ilocano language prediction. Based on data analysis, we changed some predictions automatically, e.g., all Ilocano predictions were changed into English. Sometimes we only did it when the script did not match the language, e.g., for posts with Latin script identified as Oromo. We do not recommend using this process automatically on any data. In other contexts, the generated predictions might be less noisy. Even in our case, we have different rules for posts and fact-checks based on the characteristics of these two domains. If the predictions proved to be too noisy, we unified several languages or language varieties into one. This is the case of Croatian, Bosnian and Serbian, as well as Indonesian and Malay.

C Dataset Statistics

We show the number of fact-checks and posts per language in Table 5. For fact-checks, we only take into consideration the language of claim, since we mostly only work with claims in this work.

[5] https://github.com/google/cld3

Posts can have more than one language detected based on its overall compositions. We calculated percentage for each language based on the language prediction methods. We consider all languages with at least 20% to be relevant. 25,482 posts have only one language detected, while 2,549 has two, 59 has three, 1 has four and 1 has zero.

Table 6 shows the sources of our fact-checks. Here we only show the statistics for the fact-checks we actually used in our experiments. There are additional 6k fact-checks that we have not used because they we were not able to fill their *claim* field.

D Hyperparameter Search

Table 7 show the range of hyperparmeters in our hyperparameter search, as well as the best performing hyperparameters.

E Limitations

E.1 Dataset

Noise. Based on our annotation (see §3), we expect that around 13% of our dataset is not correct. These are the cases when the claim being made on the social media is based on visual information. Note, that the methods might still be able to retrieve correct fact-checks for some of these posts, based on spurious correlations, e.g., overlaps in named entities.

AI APIs. We use out-of-the-box AI services to perform optical character recognition, machine translation to English and language detection. All of these have limited precision and might inject noise into our data.

- *OCR* was too sensitive and was often reading imaginary character, watermarks, etc. We had to address this by a more aggressive text cleaning.
- *Machine translation to English* is not perfect and the quality of translations depends on source language, particular topics or even the writing style.
- *Language detection* is an important component in our pipeline as we use it to group samples by language and then reason about these languages. Noise in language detection might have influenced our results and insights.

Code	Language	# fact-checks	# posts
ara	Arabic	14201	931
asm	Assamese	60	5
aze	Azerbaijani	178	2
bul	Bulgarian	162	114
ben	Bengali	4143	113
cat	Catalan	574	100
ces	Czech	254	265
dan	Danish	648	6
deu	German	4996	932
ell	Greek	1821	175
eng	English	85814	7307
spa	Spanish	14082	7319
fas	Farsi	418	17
fin	Finnish	109	103
tgl	Tagalog	462	439
fra	French	4355	2146
hbs	Serbo-Croatian	2451	481
hin	Hindi	7149	833
hun	Hungarian	139	113
ita	Italian	3047	65
heb	Hebrew	202	2
jpn	Japanese	62	7
khm	Khmer	144	6
kor	Korean	510	474
mkd	Macedonian	1125	1
mal	Malayalam	1206	4
msa	Malay	8424	1389
mya	Myanmar	92	172
nld	Dutch	1232	257
nor	Norwegian	440	5
pol	Polish	6912	453
por	Portuguese	21569	3366
ron	Romanian	204	238
rus	Russian	2715	28
sin	Sinhala	825	534
slk	Slovak	260	363
sqi	Albanian	726	1
tam	Tamil	1612	29
tel	Telugu	2450	11
tha	Thai	382	626
tur	Turkish	6676	7
ukr	Ukrainian	68	6
urd	Urdu	0	378
zho	Chinese	2586	595
	Others	266	343

Table 5: List of languages with at least 50 fact-checks or 50 posts.

Selection bias. There is a possibility that selection bias influences our results. First, sometimes fact-checkers writing the fact-checks base their writing on a particular post and the fact-check might contain parts of it verbatim. We tried to measure the size of this effect by comparing cases when the fact-checks are newer and older then posts (see §4.2), but we did not find a signal that this is the case. However, we know that there are at least few samples with this problem.

Second, there might be a bias towards social media posts that the social media platform or fact-checkers are already able to detect. Other, more difficult cases might still elude us.

Linguistic bias. Although our dataset is quite diverse, compared to most published datasets, there is still a bias towards major languages and Indo-European language family in particular. Crosslingual pairs consist mostly of East or South Asian posts with non-Latin script mapped to English fact-checks. It is hard to estimate how our results would generalize to other language pairs. We visualize the languages in Figure 1. The annotation efforts in Section 6 shows that there are many crosslingual pairs that our methodology was not able to detect.

E.2 Methods

Language support. The methods we use have different degrees of support for different languages. BM25 requires a proper tokenization to work. We have languages that use *scriptio continua* – Thai and Myanmar – where this is a problem. BM25-Original performance for these two is subpar, but could be improved by implementing custom tokenization models.

Multilingual TEMs we use do not support Sinhala and Tagalog languages, i.e., they were not trained with their data. The performance for these two languages is again subpar. Additionally, all methods depending on machine translation are naturally only able to handle languages that have a machine translation system available, although we believe that this was not a significant problem in our dataset.

Hidden positive pairs. The results we report might be deflated from the practical point of view because of unmarked correct pairs that are in the dataset. We have information only about a small subset of all the pairs. Our attempt to approximate true performance in in Section 6.

Name	Lang.	N	Name	Lang.	N	Name	Lang.	N
snopes.com	eng	18376	washingtonpost.com	eng	1413	agi.it	ita	246
politifact.com	eng	9029	dogrulukpayi.com	tur	1360	verify-sy.com	ara	242
misbar.com	ara	9027	stopfake.org	rus	1307	cbsnews.com	eng	242
boomlive.in	eng	7949	colombiacheck.com	spa	1271	factchecknederland.afp.com	nld	234
factcheck.afp.com	eng	6853	tempo.co	id	1143	butac.it	ita	220
cekfakta.com	id	6523	vistinomer.mk	mkd	1141	efe.com	spa	219
altnews.in	eng	6199	faktograf.hr	hr	1094	br.de	deu	214
factly.in	eng	5818	dubawa.org	eng	1066	annielab.org	eng	204
leadstories.com	eng	5319	factcheck.kz	rus	1044	globes.co.il	heb	202
sapo.pt	por	5200	istinomer.rs	sr	958	factcheckhub.com	eng	200
demagog.org.pl	rus	4292	boombd.com	ben	937	ghanafact.com	eng	199
fullfact.org	eng	4260	bufale.net	ita	928	telemundo.com	spa	195
factual.afp.com	spa	4051	apublica.org	pt-pt	915	apa.at	deu	185
uol.com.br	por	3908	rappler.com	tgl	874	verificat.afp.com	ron	177
checkyourfact.com	eng	3620	verificat.cat	spa	821	efectococuyo.com	spa	170
teyit.org	tur	3289	kallxo.com	sqi	728	factcheckni.org	eng	157
newsmobile.in	eng	3265	aap.com.au	eng	687	proveri.afp.com	bul	152
newtral.es	spa	3256	projetocomprova.com.br	por	686	icirnigeria.org	eng	142
dpa-factchecking.com	nld	2839	tjekdet.dk	dan	648	tenykerdes.afp.com	hun	138
indiatoday.in	eng	2799	dogrula.org	tur	634	liberation.fr	fra	134
factcheck.org	eng	2716	faktencheck.afp.com	deu	629	factcheckgreek.afp.com	ell	129
aosfatos.org	por	2596	thip.media	ben	598	radio-canada.ca	fra	123
boatos.org	por	2553	dailyo.in	ben	591	maharat-news.com	ara	121
aajtak.in	hin	2493	univision.com	spa	582	factcheckmyanmar.afp.com	mya	119
dabegad.com	ara	2342	periksafakta.afp.com	id	563	jachai.org	ben	113
factcheck.afp.com/ar	ara	2292	lemonde.fr	fra	558	nieuwscheckers.nl	nld	111
estadao.com.br	por	2197	check4spam.com	eng	524	europapress.es	spa	108
factuel.afp.com	fra	2178	healthfeedback.org	eng	499	faktantarkistus.afp.com	fin	107
thequint.com	eng	2058	mygopen.com	zho	494	tagesschau.de	deu	103
tfc-taiwan.org.tw	zho	1960	sprawdzam.afp.com	pol	458	scroll.in	eng	100
observador.pt	por	1930	faktisk.no	nor	444	thelallantop.com	hin	99
usatoday.com	eng	1901	presseportal.de	deu	439	theferret.scot	eng	96
oko.press	pol	1872	20minutes.fr	fra	419	france24.com	fra	92
fatabyyano.net	ara	1844	cinjenice.afp.com	sr	387	voachinese.com	zho	92
factcrescendo.com	ben	1808	factcheckthailand.afp.com	tha	382	comprovem.afp.com	cat	90
correctiv.org	deu	1783	factcheckkorea.afp.com	kor	382	factandfurious.com	fra	82
maldita.es	spa	1748	asianetnews.com	mal	365	factchecker.in	eng	74
ellinikahoaxes.gr	ell	1688	newsweek.com	eng	364	telugupost.com	tel	73
checamos.afp.com	por	1672	factnameh.com	fas	356	zimfact.org	eng	72
facta.news	ita	1652	fakenews.pl	pol	320	factcheckbangla.afp.com	ben	62
youturn.in	tam	1609	fastcheck.cl	spa	313	buzzfeed.com	jpn	56
malumatfurus.org	tur	1572	newsmeter.in	eng	290	verificado.com.mx	spa	55
polygraph.info	eng	1527	factrakers.org	eng	276	ripplesnigeria.com	eng	52
metafact.io	eng	1526	semakanfakta.afp.com	ms	267	poynter.org	eng	52
africacheck.org	eng	1468	fakty.afp.com	slk	260	globo.com	por	52
animalpolitico.com	spa	1468	napravoumiru.afp.com	ces	255	radiofarda.com	fas	51
verafiles.org	tgl	1414	factograph.info	rus	253	stern.de	deu	50

Table 6: Fact-checking sources with at least 50 fact-checks in our dataset.

Hyperparameter	**Range**	**GTR-T5-Large**	**MPNet-Base-Multilingual**
Loss	contrastive cosine online-contrastive	online-contrastive	online-contrastive
Learning rate	$[1e{-}3, 1e{-}7]$	$1e{-}5$	$5e{-}6$
Learning rate schedule	cosine linear	cosine	cosine
Warmup steps	$[100, 3200]$	800	1600
Weight decay rate	$[1e{-}7, 1e{-}4]$	$1e{-}5$	$8e{-}5$
Ratio of positive to negative samples	$[10, 50\%]$	10%	30%
Margin	$[0.1, 0.5]$	0.5	0.4
Batch size	Maximum possible	2	8

Table 7: Range of hyperparameters used in our supervised hyperparameter search and the hyperparameters of our most successful models. The ranges adjusted during the experimentation according to the preliminary results.

Supervised learning overfitting. It is possible that our supervised training yielded model that is overfitted on the particular languages and time frame that are represented in our dataset. The increase in performance might not transfer to out-of-domain pairs.

F Other Ideas

Here we discuss some additional ideas that were tried and that we decided not to include in the main text for various reasons.

Sliding window embedding. Figure 7 shows that the performance for methods decreases for posts with certain length. The decrease is generally starting at around 500 characters. We experimented with using sliding windows with various sizes (both based on the number of characters and the number of sentences) and strides. TEMs then encode only this sliding window and the final vector similarity is calculated as the maximum similarity of any of the windows. We found out that this technique can slightly ($+0.01 - 0.02$ S@10) improve the results for TEMs.

Using fact-check titles alongside claims. We represent fact-checks with the *claim* field obtained from the data in our main text experiments. We also experimented with the *title* field that we were able to obtain for the majority of the fact-checks. We found out that representing the fact-check as a concatenation of a claim and a title improves the results slightly ($+0.00 - 0.01$ S@10) for BM25 methods.

Topic detection. We attempted to run a topic detection over our posts to better understand how different methods handle different topics and themes in our data. We experimented with both original and English versions, with both multilingual and monolingual topic detection models, such as LDA (Blei et al., 2003) or BERTopic (Grootendorst, 2022). Ultimately we were not content with the quality of topic detection, as the models failed to reliably identify even the most frequent topics in our data, such as the COVID-19 pandemic or Russo-Ukrainian war. We believe that this is caused by the short length of the majority of the posts, as well as their relatively noisy nature.

Mixing original and English versions. We experimented with representing both fact-checks and posts as a concatenation of both the original language texts and the English translations, so that the multilingual methods can use both sources of information. However, this increased the *same language bias* significantly while the performance decreased significantly across the board.