Multilingual Previously Fact-Checked Claim Retrieval

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smolen
Martin Melisek, Ivan Vykopal, Jakub Simko, Juraj Podrouzek, Maria Bielikova
Kempelen Institute of Intelligent Technologies
{name.surname}@kinit.sk

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 – MultiClaim – 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 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 tasks include searching for the source of evidence for verification, searching for other versions of misinformation, and searching within existing fact-checks. They were identified as particularly challenging for fact-checkers working in low-resource languages (Hrckova et al., 2022).

In this work, we focus on 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. MultiClaim – Multilingual dataset for PFCR. We collected and made available 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

The dataset is available at Zenodo upon request for research purposes only: https://zenodo.org/record/7737983. The source code is available at: https://github.com/kinit-sk/multiclaim.
OCR transcripts of the images attached to the posts and machine translation of all the data into English.

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., 2021) 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.

<table>
<thead>
<tr>
<th>Input claims</th>
<th>FC claims</th>
<th>Pairs</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kazemi et al., 2021</td>
<td>NA</td>
<td>150,000</td>
<td>258</td>
</tr>
<tr>
<td>Jiang et al., 2021</td>
<td>NA</td>
<td>90</td>
<td>1,573</td>
</tr>
<tr>
<td>Shaar et al., 2020</td>
<td>NA</td>
<td>27,032</td>
<td>1,768</td>
</tr>
<tr>
<td>Shaar et al., 2021</td>
<td>2,259</td>
<td>44,164</td>
<td>2,440</td>
</tr>
<tr>
<td>Hardalov et al., 2022</td>
<td>316,564</td>
<td>10,340</td>
<td>332,660</td>
</tr>
<tr>
<td>MultiClaim (our)</td>
<td>28,092</td>
<td>205,751</td>
<td>31,305</td>
</tr>
</tbody>
</table>

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.

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. Finally, our dataset has by far the most languages, while the second biggest dataset in this regard has 5 languages with only 50 samples per language.

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., 2022a,b, 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 MultiClaim 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 (in 88.2% of the cases) 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. In both cases, it was professional fact-checkers that assigned the fact-checks to the posts. (1) Some fact-checks use the ClaimReview schema\(^2\), which has a field for reviewed items. All the links to the three social media platforms from this field are used to collect the posts and form the pairs. (2) We searched for URLs to Facebook and Instagram in the main body of the fact-checks. This is our pool of potentially relevant posts. Then we use the fact-checking warnings these two platforms provide. These warnings contain links to relevant fact-checking articles. We use these links to establish additional pairs. We used these links to establish additional pairs.

In total, we collected 28,092 posts from 27 languages. There are 31,305 fact-check-to-post pairs, each post in our dataset is paired with at least one fact-check. 26,774 of these pairs are monolingual and 4,212 are crosslingual (as predicted by the language identification, see below). 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. Examples from our dataset can be seen in Appendix G.

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 used automatic language identification on 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 95% Agresti-Coull confidence interval (Agresti and Coull, 1998) 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 as the percentage of pairs when the desired fact-check ends up in the top K. Throughout the paper, we report this metric with the 95% Agresti-Coull 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. Fact-checks are represented with their claims only. Posts are represented with their main texts concatenated with the OCR transcripts. We use either the original texts or their English translations, depending on the version of the dataset that is reported.

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

---

\(^2\)https://schema.org/ClaimReview
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 is 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\(^3\). 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).

### 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 directly compare the performance numbers across different monolingual languages, since they use different pools of fact-checks with different sizes. 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 by design not capable of any crosslingual retrieval and the results are shown only for completeness.

### False positive rate.

We noticed that BM25-Original seems to perform better for languages with larger fact-check pools. We conducted an

---

\(^3\)The index created for BM25 is multilingual as well.

<table>
<thead>
<tr>
<th>Method</th>
<th>Size [M]</th>
<th>Ver.</th>
<th>Mono</th>
<th>Cross</th>
<th>SLB</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English TEMs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilRoBERTa</td>
<td>En 82</td>
<td>0.76</td>
<td>0.43</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>GTR-T5-Base</td>
<td>En 110</td>
<td>0.81</td>
<td>0.51</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>GTR-T5-Large</td>
<td>En 336</td>
<td>0.83</td>
<td>0.56</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>GTR-T5-XL</td>
<td>En 1242</td>
<td>0.83</td>
<td>0.56</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>MPNet-Base</td>
<td>En 109</td>
<td>0.78</td>
<td>0.47</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>MMSAMCO-BERT-Base</td>
<td>En 109</td>
<td>0.78</td>
<td>0.46</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>Multilingual TEMs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilUSE-Base-Multilingual</td>
<td>En 135</td>
<td>0.74</td>
<td>0.40</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>LaBSE</td>
<td>En 472</td>
<td>0.63</td>
<td>0.22</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>MPNet-Base-Multilingual</td>
<td>En 278</td>
<td>0.75</td>
<td>0.41</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>MiniLM-L12-Multilingual</td>
<td>En 118</td>
<td>0.74</td>
<td>0.38</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>XLM-R</td>
<td>En 278</td>
<td>0.72</td>
<td>0.33</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

| MultiQA-MPNet-Base | En 109 | 0.80 | 0.50 | 0.18  |     |
| GPT-125M     | En 125 | 0.63 | 0.25 | 0.14  |     |
| GPT-2-7B     | En 2700| 0.77 | 0.50 | 0.19  |     |
| Sentence-T5-Base | En 110 | 0.73 | 0.37 | 0.14  |     |
| Sentence-T5-Large | En 336 | 0.75 | 0.41 | 0.15  |     |
| Sentence-T5-XL    | En 1242| 0.78 | 0.47 | 0.16  |     |

| MultiQA-MPNet-Base | En 109 | 0.80 | 0.50 | 0.18  |     |
| GPT-125M     | En 125 | 0.63 | 0.25 | 0.14  |     |
| GPT-2-7B     | En 2700| 0.77 | 0.50 | 0.19  |     |
| Sentence-T5-Base | En 110 | 0.73 | 0.37 | 0.14  |     |
| Sentence-T5-Large | En 336 | 0.75 | 0.41 | 0.15  |     |
| Sentence-T5-XL    | En 1242| 0.78 | 0.47 | 0.16  |     |

Table 2: Results for methods showing both monolingual and crosslingual 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.
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 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 when we use the full pool. This number is reported in Table 2.

BM25-Original has the highest SLB score of all the methods, as it has an implicit language filter that effectively removes fact-checks from other languages from the pool. This reduction makes the task easier, but it violates our requirement that the method should take fact-checks in all the languages 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 know 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.* 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 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.

**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 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 panel). 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 Twitter posts.

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 Section 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\(^5\). 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 Section 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. The agreement rates between the annotators were sufficiently high: 82.2%, 85.5% and 92.9%. 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. No method was 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.

\(^5\)This is largely caused by the fact that we had to use smaller batch size due to (1) this model being larger, and (2) this model not supporting average mixed precision. We believe that with larger batch size the performance could be even better.
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 in the top 10) 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 the true performance of our methods is better than 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 believe that our dataset opens up many interesting research directions. We have, for example, 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 Limitations

9.1 Dataset

Noise. Based on our annotation (see Section 3), we expect that \( \sim 13\% \) of posts in our dataset do not contain the claim in the textual form. 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.

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 Section 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 show that there are many crosslingual pairs that our data collection methodology was not able to collect.

9.2 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 is provided in Section 6.

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.

10 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 (at the time of collection) 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. Following the discussion and arguments presented in (Mancosu and Vegetti, 2020) and 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. 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.

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 to 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 Section 3).

As the dataset can also be used for supervised training (see Section 5), there is a risk of propagating biases present in the data (see Section 9). 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 biases 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.
Acknowledgements

This work was partially supported by the **Central European Digital Media Observatory (CEDMO)**, a project funded by the European Union under the Contract No. 2020-EU-1A-0267, and by **DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies**, a project funded by European Union under the Horizon Europe, GA No. 101079164.

We thank **Weights & Biases** for providing us a license for their experiment tracking platform.

References


ENISA. 2022. **ENISA Threat Landscape 2022**.


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


Preslav Nakov, Alberto Barrón-Cedeño, Giovanni da San Martin, Firoj Alam, Julia Maria Strüß, 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...**


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

The dataset was collected via our research platform Monant (Srba et al., 2019).

Crawling. We use a Selenium-based web crawler that visits the links, extracts the HTML content and parses it with the Beautiful Soup\(^6\) library.

Source of fact-checks. We only processed fact-checks written by the AFP news agency. 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. Pairs with fact-checks from other organizations might have been established from the warnings.

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\(^7\) 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.

\(^6\)https://www.crummy.com/software/BeautifulSoup/

\(^7\)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.

Table 7 show the number of fact-check-to-post pairs for different language combinations.

Figure 8 show the density of lengths for both the fact-checked claims and the posts. Both have long tail distributions, but the claims are in general much shorter. 99% of claims are shorter than 379 characters. For social media posts, it is 4129 characters.

D Hyperparameters

D.1 BM25

We use the default PyTerrier values for BM25 algorithm: $k_1 = 1.2, b = 0.75$. Our preliminary results show that the results are not very sensitive towards these two hyperparameters, probably because of the relatively short length of the documents that we retrieve. Most claims in our dataset have only one sentence.

D.2 Supervised Training

Table 8 show the range of hyperparameters in our hyperparameter search done for supervised training in Section 5, as well as the best performing hyperparameters.
### Table 6: Fact-checking sources with at least 50 fact-checks in our dataset.

<table>
<thead>
<tr>
<th>Name</th>
<th>Lang</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>snopes.com</td>
<td>eng</td>
<td>18376</td>
</tr>
<tr>
<td>politifact.com</td>
<td>eng</td>
<td>9029</td>
</tr>
<tr>
<td>minbar.com</td>
<td>ara</td>
<td>9027</td>
</tr>
<tr>
<td>boomblive.in</td>
<td>eng</td>
<td>7949</td>
</tr>
<tr>
<td>factcheck.afp.com</td>
<td>eng</td>
<td>6853</td>
</tr>
<tr>
<td>cerkafakta.com</td>
<td>ind</td>
<td>6523</td>
</tr>
<tr>
<td>alnews.in</td>
<td>eng</td>
<td>6199</td>
</tr>
<tr>
<td>factly.in</td>
<td>ara</td>
<td>5818</td>
</tr>
<tr>
<td>leadstories.com</td>
<td>eng</td>
<td>5319</td>
</tr>
<tr>
<td>sapo.pt</td>
<td>por</td>
<td>5200</td>
</tr>
<tr>
<td>demagog.org_pl</td>
<td>pol</td>
<td>4292</td>
</tr>
<tr>
<td>fullfact.org</td>
<td>eng</td>
<td>4260</td>
</tr>
<tr>
<td>factnulialp.com</td>
<td>spa</td>
<td>4051</td>
</tr>
<tr>
<td>uol.com.br</td>
<td>por</td>
<td>3908</td>
</tr>
<tr>
<td>checkcrapfact.org</td>
<td>eng</td>
<td>3620</td>
</tr>
<tr>
<td>teyoi.org</td>
<td>tur</td>
<td>3289</td>
</tr>
<tr>
<td>newsmobile.in</td>
<td>eng</td>
<td>3265</td>
</tr>
<tr>
<td>bonatos.org</td>
<td>por</td>
<td>2596</td>
</tr>
<tr>
<td>aajtak.in</td>
<td>hin</td>
<td>2493</td>
</tr>
<tr>
<td>dabregal.com</td>
<td>ana</td>
<td>2342</td>
</tr>
<tr>
<td>factcheck.com/alp/ar</td>
<td>ara</td>
<td>2292</td>
</tr>
<tr>
<td>estudios.com.br</td>
<td>eng</td>
<td>2197</td>
</tr>
<tr>
<td>factuel.afp.fr</td>
<td>fra</td>
<td>2178</td>
</tr>
<tr>
<td>thequint.com</td>
<td>eng</td>
<td>2058</td>
</tr>
<tr>
<td>flicsawarn.org.tw</td>
<td>zho</td>
<td>1900</td>
</tr>
<tr>
<td>usaratv.org</td>
<td>por</td>
<td>1930</td>
</tr>
<tr>
<td>usatoday.com</td>
<td>eng</td>
<td>1901</td>
</tr>
<tr>
<td>olko.press</td>
<td>pol</td>
<td>1872</td>
</tr>
<tr>
<td>factbyjanyo.net</td>
<td>ara</td>
<td>1844</td>
</tr>
<tr>
<td>factcheckersend/com</td>
<td>ben</td>
<td>1808</td>
</tr>
<tr>
<td>connectus.org</td>
<td>den</td>
<td>1783</td>
</tr>
<tr>
<td>maltida.es</td>
<td>spa</td>
<td>1748</td>
</tr>
<tr>
<td>ellilinkautors.gr</td>
<td>ell</td>
<td>1688</td>
</tr>
<tr>
<td>checamos.afp.com</td>
<td>por</td>
<td>1672</td>
</tr>
<tr>
<td>facta.news</td>
<td>ita</td>
<td>1652</td>
</tr>
<tr>
<td>youattain.com</td>
<td>tam</td>
<td>1609</td>
</tr>
<tr>
<td>mahamuffars.org</td>
<td>tur</td>
<td>1572</td>
</tr>
<tr>
<td>polygraph.info</td>
<td>eng</td>
<td>1527</td>
</tr>
<tr>
<td>metafact.io</td>
<td>por</td>
<td>1517</td>
</tr>
<tr>
<td>africacheck.org</td>
<td>eng</td>
<td>1468</td>
</tr>
<tr>
<td>awajulpolitico.org</td>
<td>spa</td>
<td>1460</td>
</tr>
<tr>
<td>verafacts.info</td>
<td>tgl</td>
<td>1414</td>
</tr>
</tbody>
</table>

### Table 7: Number of fact-check-to-post pairs for different language combinations. Note that one post can have more than one language assigned.

<table>
<thead>
<tr>
<th>Social media post language</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>zho</td>
<td>urd</td>
</tr>
<tr>
<td>urd</td>
<td>tha</td>
</tr>
<tr>
<td>tha</td>
<td>spa</td>
</tr>
<tr>
<td>spa</td>
<td>slk</td>
</tr>
<tr>
<td>slk</td>
<td>ron</td>
</tr>
<tr>
<td>ron</td>
<td>por</td>
</tr>
<tr>
<td>por</td>
<td>pol</td>
</tr>
<tr>
<td>pol</td>
<td>nld</td>
</tr>
<tr>
<td>nld</td>
<td>deu</td>
</tr>
<tr>
<td>deu</td>
<td>ell</td>
</tr>
<tr>
<td>ell</td>
<td>eng</td>
</tr>
<tr>
<td>eng</td>
<td>fin</td>
</tr>
<tr>
<td>fin</td>
<td>fra</td>
</tr>
<tr>
<td>fra</td>
<td>hbs</td>
</tr>
<tr>
<td>hbs</td>
<td>hin</td>
</tr>
<tr>
<td>hin</td>
<td>hun</td>
</tr>
<tr>
<td>hun</td>
<td>kor</td>
</tr>
<tr>
<td>kor</td>
<td>msa</td>
</tr>
<tr>
<td>msa</td>
<td>mya</td>
</tr>
<tr>
<td>mya</td>
<td>nld</td>
</tr>
<tr>
<td>nld</td>
<td>pol</td>
</tr>
<tr>
<td>pol</td>
<td>por</td>
</tr>
<tr>
<td>por</td>
<td>sin</td>
</tr>
<tr>
<td>sin</td>
<td>slk</td>
</tr>
<tr>
<td>slk</td>
<td>spa</td>
</tr>
<tr>
<td>spa</td>
<td>tha</td>
</tr>
<tr>
<td>tha</td>
<td>Other</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ara</th>
<th>ben</th>
</tr>
</thead>
<tbody>
<tr>
<td>ben</td>
<td>cat</td>
</tr>
<tr>
<td>cat</td>
<td>ces</td>
</tr>
<tr>
<td>ces</td>
<td>deu</td>
</tr>
<tr>
<td>deu</td>
<td>ell</td>
</tr>
<tr>
<td>ell</td>
<td>eng</td>
</tr>
<tr>
<td>eng</td>
<td>fin</td>
</tr>
<tr>
<td>fin</td>
<td>fra</td>
</tr>
<tr>
<td>fra</td>
<td>hbs</td>
</tr>
<tr>
<td>hbs</td>
<td>hin</td>
</tr>
<tr>
<td>hin</td>
<td>kor</td>
</tr>
<tr>
<td>kor</td>
<td>msa</td>
</tr>
<tr>
<td>msa</td>
<td>mya</td>
</tr>
<tr>
<td>mya</td>
<td>nld</td>
</tr>
<tr>
<td>nld</td>
<td>pol</td>
</tr>
<tr>
<td>pol</td>
<td>por</td>
</tr>
<tr>
<td>por</td>
<td>sin</td>
</tr>
<tr>
<td>sin</td>
<td>slk</td>
</tr>
<tr>
<td>slk</td>
<td>spa</td>
</tr>
<tr>
<td>spa</td>
<td>tha</td>
</tr>
<tr>
<td>tha</td>
<td>Other</td>
</tr>
</tbody>
</table>

---

16491
Table 8: 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.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Range</th>
<th>GTR-T5-Large</th>
<th>MPNet-Base-Multilingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss</td>
<td>contrastive cosine</td>
<td>online-contrastive</td>
<td>online-contrastive</td>
</tr>
<tr>
<td>Learning rate</td>
<td>[1e−3, 1e−7]</td>
<td>1e−5</td>
<td>5e−6</td>
</tr>
<tr>
<td>Learning rate schedule</td>
<td>cosine linear</td>
<td>cosine</td>
<td>cosine</td>
</tr>
<tr>
<td>Warmup steps</td>
<td>[100, 3200]</td>
<td>800</td>
<td>1600</td>
</tr>
<tr>
<td>Weight decay rate</td>
<td>[1e−7, 1e−4]</td>
<td>1e−5</td>
<td>8e−5</td>
</tr>
<tr>
<td>Ratio of positive to negative samples</td>
<td>[10, 50%]</td>
<td>10%</td>
<td>30%</td>
</tr>
<tr>
<td>Margin</td>
<td>[0.1, 0.5]</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>Batch size</td>
<td>Maximum possible</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

E Additional Results

E.1 Additional Metrics

Table 9 show additional IR metrics calculated for the experiments done in Section 4.1. There is a strong correlation between all these metrics, as shown in Table 10. This is caused by the fact that most posts have only one fact-check assigned and the calculations for such cases are very similar for different metrics. We ultimately decided to use S@10 as our main evaluation metric in this work as we find it to be the most interpretable measure (for how many pairs the expected fact-checked claim ended up in the top 10).

E.2 Detailed Per-language Results

Table 11 shows additional language-specific results for all the methods from Section 4 including confidence intervals.

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.

G Examples

This Appendix contains 5 randomly selected fact-check-post pairs from our dataset. We show here all the information present in our dataset for these samples.
## Table 9: The results for different ranking methods. This table shows the same experiment as Table 2, but also calculates additional information retrieval metrics: MRR, MAP, NDCG, MAP@10.

<table>
<thead>
<tr>
<th>Method</th>
<th>S@10 Ver.</th>
<th>MRR Mono</th>
<th>MAP Mono</th>
<th>NDCG Mono</th>
<th>MAP@10 Mono</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BM25</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM25</td>
<td>0.61</td>
<td>0.22</td>
<td>0.39</td>
<td>0.61</td>
<td>0.22</td>
</tr>
<tr>
<td>BM25 Og</td>
<td>0.48</td>
<td>0.04</td>
<td>0.06</td>
<td>0.47</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>English TEMs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilRoBERTa En</td>
<td>0.59</td>
<td>0.24</td>
<td>0.43</td>
<td>0.59</td>
<td>0.24</td>
</tr>
<tr>
<td>DistilRoBERTa Og</td>
<td>0.50</td>
<td>0.10</td>
<td>0.20</td>
<td>0.49</td>
<td>0.10</td>
</tr>
<tr>
<td>GTR-T5-Base En</td>
<td>0.65</td>
<td>0.30</td>
<td>0.51</td>
<td>0.64</td>
<td>0.30</td>
</tr>
<tr>
<td>GTR-T5-Base Og</td>
<td>0.53</td>
<td>0.12</td>
<td>0.22</td>
<td>0.53</td>
<td>0.12</td>
</tr>
<tr>
<td>GTR-T5-Large En</td>
<td><strong>0.67</strong></td>
<td>0.33</td>
<td><strong>0.83</strong></td>
<td><strong>0.66</strong></td>
<td><strong>0.33</strong></td>
</tr>
<tr>
<td>GTR-T5-Large Og</td>
<td>0.58</td>
<td>0.19</td>
<td>0.33</td>
<td>0.55</td>
<td>0.19</td>
</tr>
<tr>
<td>MPNet-Base En</td>
<td>0.61</td>
<td>0.27</td>
<td>0.47</td>
<td>0.60</td>
<td>0.27</td>
</tr>
<tr>
<td>MPNet-Base Og</td>
<td>0.53</td>
<td>0.12</td>
<td>0.22</td>
<td>0.53</td>
<td>0.12</td>
</tr>
<tr>
<td>MSMARCO-BERT-Base En</td>
<td>0.62</td>
<td>0.26</td>
<td>0.46</td>
<td>0.61</td>
<td>0.26</td>
</tr>
<tr>
<td>MSMARCO-BERT-Base Og</td>
<td>0.53</td>
<td>0.12</td>
<td>0.22</td>
<td>0.53</td>
<td>0.12</td>
</tr>
<tr>
<td>MiniLM-L12 En</td>
<td>0.64</td>
<td>0.29</td>
<td>0.48</td>
<td>0.63</td>
<td>0.29</td>
</tr>
<tr>
<td>MultiQA-MPNet-Base En</td>
<td>0.64</td>
<td>0.29</td>
<td>0.48</td>
<td>0.63</td>
<td>0.29</td>
</tr>
<tr>
<td>MultiQA-MPNet-Base Og</td>
<td>0.50</td>
<td>0.10</td>
<td>0.20</td>
<td>0.49</td>
<td>0.10</td>
</tr>
<tr>
<td>SGPT-125M En</td>
<td>0.47</td>
<td>0.14</td>
<td>0.25</td>
<td>0.47</td>
<td>0.14</td>
</tr>
<tr>
<td>SGPT-125M Og</td>
<td>0.50</td>
<td>0.10</td>
<td>0.20</td>
<td>0.49</td>
<td>0.10</td>
</tr>
<tr>
<td>Sentence-T5-Base En</td>
<td>0.57</td>
<td>0.21</td>
<td>0.37</td>
<td>0.56</td>
<td>0.21</td>
</tr>
<tr>
<td>Sentence-T5-Base Og</td>
<td>0.54</td>
<td>0.19</td>
<td>0.33</td>
<td>0.55</td>
<td>0.19</td>
</tr>
<tr>
<td>Sentence-T5-Large En</td>
<td>0.58</td>
<td>0.24</td>
<td>0.41</td>
<td>0.58</td>
<td>0.23</td>
</tr>
<tr>
<td>Sentence-T5-Large Og</td>
<td>0.53</td>
<td>0.12</td>
<td>0.22</td>
<td>0.53</td>
<td>0.12</td>
</tr>
<tr>
<td>Sentence-T5-XL En</td>
<td>0.61</td>
<td>0.27</td>
<td>0.47</td>
<td>0.60</td>
<td>0.26</td>
</tr>
<tr>
<td>Sentence-T5-XL Og</td>
<td>0.53</td>
<td>0.12</td>
<td>0.22</td>
<td>0.53</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>Multilingual TEMs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilUSE-Base-Multilingual En</td>
<td>0.58</td>
<td>0.22</td>
<td>0.40</td>
<td>0.58</td>
<td>0.22</td>
</tr>
<tr>
<td>DistilUSE-Base-Multilingual Og</td>
<td>0.50</td>
<td>0.10</td>
<td>0.20</td>
<td>0.49</td>
<td>0.10</td>
</tr>
<tr>
<td>LaBSE En</td>
<td>0.47</td>
<td>0.13</td>
<td>0.22</td>
<td>0.46</td>
<td>0.13</td>
</tr>
<tr>
<td>LaBSE Og</td>
<td>0.53</td>
<td>0.12</td>
<td>0.22</td>
<td>0.53</td>
<td>0.12</td>
</tr>
<tr>
<td>MPNet-Base-Multilingual En</td>
<td><strong>0.60</strong></td>
<td>0.24</td>
<td><strong>0.75</strong></td>
<td><strong>0.59</strong></td>
<td><strong>0.23</strong></td>
</tr>
<tr>
<td>MPNet-Base-Multilingual Og</td>
<td>0.53</td>
<td>0.11</td>
<td>0.21</td>
<td>0.53</td>
<td>0.11</td>
</tr>
<tr>
<td>MiniLM-L2-Multilingual En</td>
<td>0.58</td>
<td>0.22</td>
<td>0.38</td>
<td>0.57</td>
<td>0.22</td>
</tr>
<tr>
<td>MiniLM-L2-Multilingual Og</td>
<td>0.48</td>
<td>0.08</td>
<td>0.15</td>
<td>0.47</td>
<td>0.08</td>
</tr>
<tr>
<td>XLM-R En</td>
<td>0.55</td>
<td>0.19</td>
<td>0.33</td>
<td>0.55</td>
<td>0.19</td>
</tr>
<tr>
<td>XLM-R Og</td>
<td>0.50</td>
<td>0.08</td>
<td>0.15</td>
<td>0.50</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Table 10: Pearson correlation coefficient between different metrics as calculated in Table 9. Both monolingual and crosslingual scores are taken into consideration.

<table>
<thead>
<tr>
<th>S@10</th>
<th>MRR</th>
<th>MAP</th>
<th>NDCG</th>
<th>MAP@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>S@10</td>
<td>1.000</td>
<td>0.986</td>
<td>0.986</td>
<td>0.993</td>
</tr>
<tr>
<td>MRR</td>
<td>0.986</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td>MAP</td>
<td>0.986</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td>NDCG</td>
<td>0.993</td>
<td>0.998</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>MAP@10</td>
<td>1.000</td>
<td>0.988</td>
<td>0.988</td>
<td>0.995</td>
</tr>
</tbody>
</table>

G.1 Example #1

G.1.1 Fact-check

ID: 104315
Published at: 2021-09-27 factcheck.afp.com

Claim

Original text: Photo shows meeting of five international intelligence agencies in Delhi
Translated text: Photo shows meeting of five international intelligence agencies in Delhi
Detected languages: eng: 100.0%

Title

Original text: This photo shows a delegation-level meeting between Indian and Russian national security advisors
Translated text: This photo shows a delegation-level meeting between Indian and Russian national security advisors
Detected languages: eng: 100.0%

G.1.2 Social Media Post

ID: 16806
Published at: Facebook 2021-09-16
Verdicts: Partly false information

Main text

Original text: Post Giri IyerNow in Delhi !!India RAWIsrael MOSSADAmercia CIARussia KGBEngland Mi6First time ever that the top five intelligence agency of the world are sitting together for a high level meeting in Delhi. This is the power of new India
Translated text: Post Giri IyerNow in Delhi !!India RAWIsrael MOSSADAmercia CIARussia KGBEngland Mi6First time ever that the top five intelligence agency of the world are sitting together for a high level meeting in Delhi. This is the power of new India
Detected languages: eng: 100.0%

G.2 Example #2

G.2.1 Fact-check

ID: 34296
Published at: 2019-10-29 factuel.afp.com

Claim

Original text: COMMENT TRAITER L’HÉPATITE B PAR LES PLANTES
Translated text: HOW TO TREAT HEPATITIS B WITH HERBS
Detected languages: fra: 100.0%

Title

Original text: Non, cette boisson à base de papaye, de citron, de racines de cocotier et de moringa bouillis, ne guérit pas l’hépatite B
Translated text: No, this drink made from boiled papaya, lemon, coconut palm roots and moringa does not cure hepatitis B
Detected languages: fra: 100.0%

G.2.2 Social Media Post

ID: 11569
Published at: Facebook 2019-07-06
Verdicts: False information

Main text

Original text: HÉPATITE B ET Cet très rassurant faites cette expérience et révenez témoigné !!! L’hépatite n’ex qu’un vieux souvenir après !!REMÈDES POUR TRAITER ET ÉRADIQUER L’HÉPATITE B DU CORPSL’hépatite B est une infection virale qui s’attaque au foie.Le virus se transmet par le sang ou lors des rapports sexuels. En effet, les seules sécrétions ou liquides corporels qui permettent de transmettre le virus sont le sang, le sperme, les sécrétions vaginales, la salive et les liquides issus d’une plaiIngrédients Une Papaye non mur Les RACINES de Papayer Femelle Les feuilles fraîches de Papayer femelle racines de Moringa feuilles fraiche Moringa 4 citrons à couper en deux. racines de cocotierPréparation-Mettez les tous dans la marmite, les feuilles en dernier position. Ajoutez de l’eau et faites bouillir le mélange.Mode d’emploiBoire 2 à 3 verres par jour.Ajoutez de l’eau à chaque fois et faites bouillir une fois par jour . Suivez le traitement pendant un mois.Faites vous examiné par un médecin et revenez témoigner .Bonne guérison...Aimes ton prochain par le partage de ce messageLa Boutique
Table 11: S@10 performance with confidence intervals for individual languages. This table has the same results as Figure 3, but also calculated for additional methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>En</th>
<th>spa</th>
<th>eng</th>
<th>por</th>
<th>pol</th>
<th>slk</th>
<th>ndl</th>
<th>nld</th>
<th>ell</th>
<th>ces</th>
<th>bul</th>
<th>hun</th>
<th>hin</th>
<th>mya</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.79±0.01</td>
<td>0.62±0.01</td>
<td>0.78±0.01</td>
<td>0.77±0.02</td>
<td>0.74±0.02</td>
<td>0.70±0.03</td>
<td>0.82±0.03</td>
<td>0.89±0.03</td>
<td>0.66±0.04</td>
<td>0.81±0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilRoBERTa</td>
<td>0.72±0.01</td>
<td>0.72±0.01</td>
<td>0.74±0.01</td>
<td>0.76±0.02</td>
<td>0.74±0.02</td>
<td>0.72±0.03</td>
<td>0.77±0.03</td>
<td>0.80±0.03</td>
<td>0.65±0.04</td>
<td>0.82±0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTR-T5-Base</td>
<td>0.80±0.01</td>
<td>0.74±0.01</td>
<td>0.75±0.01</td>
<td>0.84±0.02</td>
<td>0.80±0.02</td>
<td>0.66±0.03</td>
<td>0.85±0.03</td>
<td>0.90±0.02</td>
<td>0.71±0.04</td>
<td>0.85±0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTR-T5-Large</td>
<td>0.84±0.01</td>
<td>0.77±0.01</td>
<td>0.80±0.01</td>
<td>0.86±0.02</td>
<td>0.82±0.02</td>
<td>0.69±0.03</td>
<td>0.86±0.02</td>
<td>0.90±0.02</td>
<td>0.74±0.04</td>
<td>0.86±0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTR-T5-XL</td>
<td>0.83±0.01</td>
<td>0.77±0.01</td>
<td>0.80±0.01</td>
<td>0.86±0.01</td>
<td>0.83±0.02</td>
<td>0.70±0.03</td>
<td>0.87±0.02</td>
<td>0.91±0.02</td>
<td>0.73±0.04</td>
<td>0.86±0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPNet-Base</td>
<td>0.75±0.01</td>
<td>0.68±0.01</td>
<td>0.67±0.02</td>
<td>0.81±0.02</td>
<td>0.79±0.02</td>
<td>0.57±0.03</td>
<td>0.80±0.03</td>
<td>0.83±0.03</td>
<td>0.70±0.04</td>
<td>0.80±0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSMARCO-BERT-Base</td>
<td>0.78±0.01</td>
<td>0.63±0.01</td>
<td>0.75±0.01</td>
<td>0.80±0.02</td>
<td>0.79±0.02</td>
<td>0.59±0.03</td>
<td>0.83±0.02</td>
<td>0.87±0.03</td>
<td>0.66±0.04</td>
<td>0.85±0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPNet-L12</td>
<td>0.78±0.01</td>
<td>0.70±0.01</td>
<td>0.72±0.01</td>
<td>0.82±0.02</td>
<td>0.78±0.02</td>
<td>0.64±0.03</td>
<td>0.84±0.02</td>
<td>0.89±0.03</td>
<td>0.72±0.04</td>
<td>0.83±0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MultiQA-MPNet-Base</td>
<td>0.79±0.01</td>
<td>0.70±0.01</td>
<td>0.71±0.01</td>
<td>0.84±0.02</td>
<td>0.81±0.02</td>
<td>0.64±0.03</td>
<td>0.82±0.03</td>
<td>0.90±0.02</td>
<td>0.71±0.04</td>
<td>0.84±0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGPT-2TB</td>
<td>0.77±0.01</td>
<td>0.65±0.01</td>
<td>0.72±0.01</td>
<td>0.79±0.02</td>
<td>0.79±0.02</td>
<td>0.60±0.03</td>
<td>0.82±0.03</td>
<td>0.91±0.02</td>
<td>0.70±0.04</td>
<td>0.83±0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence-T5-Base</td>
<td>0.71±0.01</td>
<td>0.60±0.01</td>
<td>0.63±0.02</td>
<td>0.75±0.02</td>
<td>0.77±0.02</td>
<td>0.43±0.03</td>
<td>0.76±0.03</td>
<td>0.89±0.03</td>
<td>0.59±0.04</td>
<td>0.83±0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence-T5-Large</td>
<td>0.74±0.01</td>
<td>0.64±0.01</td>
<td>0.65±0.02</td>
<td>0.76±0.02</td>
<td>0.76±0.02</td>
<td>0.47±0.03</td>
<td>0.80±0.03</td>
<td>0.89±0.02</td>
<td>0.59±0.04</td>
<td>0.81±0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence-T5-XL</td>
<td>0.77±0.01</td>
<td>0.68±0.01</td>
<td>0.70±0.01</td>
<td>0.80±0.02</td>
<td>0.79±0.02</td>
<td>0.53±0.03</td>
<td>0.82±0.03</td>
<td>0.91±0.02</td>
<td>0.68±0.04</td>
<td>0.81±0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilUSE-Base-Multilingual</td>
<td>Oq</td>
<td>0.64±0.01</td>
<td>0.56±0.01</td>
<td>0.58±0.02</td>
<td>0.69±0.02</td>
<td>0.60±0.03</td>
<td>0.50±0.03</td>
<td>0.74±0.03</td>
<td>0.72±0.04</td>
<td>0.57±0.04</td>
<td>0.74±0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LabSE</td>
<td>En</td>
<td>0.56±0.01</td>
<td>0.44±0.01</td>
<td>0.53±0.02</td>
<td>0.60±0.02</td>
<td>0.59±0.03</td>
<td>0.34±0.03</td>
<td>0.68±0.03</td>
<td>0.82±0.04</td>
<td>0.44±0.04</td>
<td>0.72±0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LabSE</td>
<td>Oq</td>
<td>0.64±0.01</td>
<td>0.44±0.01</td>
<td>0.66±0.02</td>
<td>0.72±0.02</td>
<td>0.67±0.02</td>
<td>0.48±0.03</td>
<td>0.77±0.03</td>
<td>0.79±0.03</td>
<td>0.57±0.04</td>
<td>0.77±0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPNet-Base-Multilingual</td>
<td>Oq</td>
<td>0.72±0.01</td>
<td>0.62±0.01</td>
<td>0.64±0.02</td>
<td>0.79±0.02</td>
<td>0.73±0.02</td>
<td>0.58±0.03</td>
<td>0.83±0.03</td>
<td>0.89±0.03</td>
<td>0.64±0.04</td>
<td>0.80±0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPNet-Base-Multilingual</td>
<td>Oq</td>
<td>0.64±0.01</td>
<td>0.61±0.01</td>
<td>0.57±0.02</td>
<td>0.72±0.02</td>
<td>0.57±0.02</td>
<td>0.53±0.03</td>
<td>0.54±0.03</td>
<td>0.48±0.03</td>
<td>0.53±0.03</td>
<td>0.51±0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mani-L12-Multilingual</td>
<td>En</td>
<td>0.69±0.01</td>
<td>0.59±0.01</td>
<td>0.60±0.02</td>
<td>0.75±0.02</td>
<td>0.71±0.02</td>
<td>0.54±0.03</td>
<td>0.82±0.03</td>
<td>0.86±0.03</td>
<td>0.65±0.04</td>
<td>0.79±0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mani-L12-Multilingual</td>
<td>Oq</td>
<td>0.57±0.01</td>
<td>0.57±0.01</td>
<td>0.51±0.02</td>
<td>0.66±0.02</td>
<td>0.54±0.03</td>
<td>0.49±0.03</td>
<td>0.49±0.03</td>
<td>0.82±0.03</td>
<td>0.55±0.04</td>
<td>0.61±0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XLM-R</td>
<td>En</td>
<td>0.64±0.01</td>
<td>0.53±0.01</td>
<td>0.60±0.02</td>
<td>0.72±0.02</td>
<td>0.69±0.02</td>
<td>0.53±0.03</td>
<td>0.86±0.03</td>
<td>0.58±0.04</td>
<td>0.79±0.04</td>
<td>0.74±0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XLM-R</td>
<td>Oq</td>
<td>0.58±0.01</td>
<td>0.52±0.01</td>
<td>0.55±0.02</td>
<td>0.70±0.02</td>
<td>0.57±0.03</td>
<td>0.54±0.03</td>
<td>0.51±0.03</td>
<td>0.82±0.03</td>
<td>0.54±0.04</td>
<td>0.63±0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11: S@10 performance with confidence intervals for individual languages. This table has the same results as Figure 3, but also calculated for additional methods.
HEPATITIS B AND Cand very reassuring made this experience and dreamed witnessed !!!Hepatitis is just an old memory afterwards!!REMEDIES TO TREAT AND ERADICATE HEPATITIS B FROM THE BODYHepatitis B is a viral infection that attacks the liver. The virus is transmitted through blood or during sexual intercourse. Indeed, the only secretions or bodily fluids that can transmit the virus are blood, semen, vaginal secretions, saliva and fluids from a wound.Ingredients An unripe Papaya The ROOTS of Female Papaya Fresh female papaya leaves Moringa roots fresh Moringa leaves 4 lemons to be cut in half. coconut rootsPreparationPut them all in the pot, the leaves last. Add water and boil the mixture.ManualDrink 2-3 glasses a day. Add water each time and boil once a day. Follow the treatment for a month. Get examined by a doctor and come back to testify.Good recovery...Love your neighbor by sharing this message.
translated text: Ukraine’s borders are recognized and do not need to be registered with the UN.
area is an internal Russian affair. Any interference will be seen as a measure against Russia. Based on that, they can nullify the May 25, 2014 elections and do whatever the people want! According to the Budapest Memorandum and other agreements, Ukraine has no borders. The state of Ukraine does not exist (and never did!).” Alexandre Panin

G.5 Example #5

G.5.1 Fact-check

ID: 61827
Published at: 2019-11-26 periksaafakta.afp.com

Claim
Original text: Gadis gembala di Maroko menjadi menteri pendidikan Prancis setelah dewasa
Translated text: Shepherd girl in Morocco becomes French education minister as an adult
Detected languages: msa: 100.0%

Title
Original text: Ini adalah foto seorang anak Maroko, bukan mantan menteri pendidikan Prancis
Translated text: This is a photo of a Moroccan child, not a former French education minister
Detected languages: msa: 100.0%

G.5.2 Social Media Post

ID: 10815
Published at: Facebook
Verdicts: None

Main text


Translated text: The girl on the left is herding sheep in Morocco, the woman on the right is the same girl 20 years later as the French Minister of Education. Never stop dreaming and never stop working hard for your dreams.... *** Energetic. That is the picture of Najat Vallaud-Belkacem. In the past, he wore modest clothes with his hair in a ponytail, carried a stick, and herded sheep. Everyday she is a shepherd girl in a small village near Nador, Morocco. At that time no one expected that his life as an adult would change much for the better. Became the French minister of education and research. Of course that position didn’t just come from the sky. Belkacem tried extra hard to reach it. In his dictionary, there is nothing that cannot be realized. In the past, when he wanted to study at the Paris Institute of Political Studies, his school teacher forbade him to enroll. The reason, the school is expensive and difficult to enter. However, the step of the second child among the seven siblings did not subside. Belkacem continued to apply, studied hard, and was finally accepted. He also had to work part-time at two places to pay for his tuition. On the same campus, he met Boris Vallaud, who is now an adviser to French President Francois Hollande. They are both active in the Socialist Party. The two were married on August 27, 2005. Long before that, Belkacem was also used to living hard. When he was four years old, his father took him, his mother and eldest sister, Fatiha, to Amiens, a suburb of France. “My father had no problems. But, we, me, mother and brother, are desperately adapting to a new life,” he was quoted as saying by Vogue. He even had time to be surprised when he saw the car. A rare thing in their home country. Not to mention the discrimination that comes from the surrounding environment. Even when he was already a member of parliament in the Rhone-Alpes. In an article, Belkacem recounted that at that time he held a dinner banquet and invited guests who did not know him well. When the guest arrived, Belkacem greeted him and helped him take off his coat. The guest then asked where the owner of the house was. "Until now in France, if a woman of color opened the door to a house in a luxury area, it was always considered a maid," wrote the mother of twins Louis-Adel Vallaud and Nour-Chloe Vallaud. Since then, he has been steadily devoting his life to eliminating discrimination. The spotlight on Belkacem’s illustrious career began when President Francois Hollande appointed him as government spokesman and minister for women’s rights on 16 May 2012. Months after that, Hollande gave him the responsibility to fight homophobia. Belkacem took office as minister of education and research on August 25, 2014, two days before her ninth wedding anniversary. The
appointment makes him the youngest education minister France has ever had. The election of Belkacem seems to be proof that an immigrant can also be a valuable asset for the country. Moreover, he is a Muslim. About Belkacem When he was a child, the best moment in his life was when a bibliobus (mobile library car) visited the area where he lived. Because, he can read a variety of books. Have dual citizenship. One of them is Morocco because he is from there. In addition, France gave him the status of a citizen while still in college. The second of seven children, Najat Belkacem was born in Morocco in 1977 in Bni Chiker, a village near Nador in the Rif region. In 1982 he rejoined his father, a construction worker, with his mother and sister Fatiha, and grew up in the suburb of Amiens. He graduated from the Institut d’études politiques de Paris (Paris Institute of Political Studies) in 2002. At the Institute he met Boris Vallaud, whom he married on 27 August 2005. He joined the Socialist Party in 2002 and joined the team of Gérard Collomb, Mayor of Lyon, in 2003 to promote strong local democracy, fight against discrimination, promote civil rights, and access to jobs and housing. Elected to the Rhone-Alpes County Council in 2004, he chaired the Culture Commission, resigning in 2008. In 2005, he became an adviser to the Socialist Party. In 2005 and 2006 he was columnist for the cultural program C’est tout vu at Télé Lyon Municipale together with Stéphane Cayrol. In February 2007 he joined the Ségolène Royal campaign team as a spokesperson, along with Vincent Peillon and Arnaud Montebourg. In March 2008 he was elected conseillère générale of the Rhone department in cantonal elections with 58.52% of the vote in the second round, under the banner of the Socialist Party in the canton of Lyon-XIII. On 16 May 2012, she was appointed to the cabinet of French President François Hollande as Minister of Women’s Rights and spokesperson for the government. https://m.facebook.com/story.php...