Breaking Language Barriers with MMTweets: Advancing Cross-Lingual Debunked Narrative Retrieval for Fact-Checking

Iknoor Singh[1], Carolina Scarton[1], Xingyi Song[1] and Kalina Bontcheva[1]

[1]University of Sheffield, Sheffield, United Kingdom

Abstract

Finding previously debunked narratives involves identifying claims that have already undergone fact-checking. The issue intensifies when similar false claims persist in multiple languages, despite the availability of debunks for several months in another language. Hence, automatically finding debunks (or fact-checks) in multiple languages is crucial to make the best use of scarce fact-checkers' resources. Mainly due to the lack of readily available data, this is an understudied problem, particularly when considering the cross-lingual scenario, i.e. the retrieval of debunks in a language different from the language of the online post being checked. This study introduces cross-lingual debunked narrative retrieval and addresses this research gap by: (i) creating Multilingual Misinformation Tweets (MMTweets): a dataset that stands out, featuring cross-lingual pairs, images, human annotations, and fine-grained labels, making it a comprehensive resource compared to its counterparts; (ii) conducting an extensive experiment to benchmark state-of-the-art cross-lingual retrieval models and introducing multistage retrieval methods tailored for the task; and (iii) comprehensively evaluating retrieval models for their cross-lingual and cross-dataset transfer capabilities within MMTweets, and conducting a retrieval latency analysis. We find that MMTweets presents challenges for cross-lingual debunked narrative retrieval, highlighting areas for improvement in retrieval models. Nonetheless, the study provides valuable insights for creating MMTweets datasets and optimising debunked narrative retrieval models to empower fact-checking endeavours. The dataset and codebook are publicly available at https://doi.org/10.5281/zenodo.10637161.

Keywords

Misinformation Detection, Disinformation Detection, Cross-lingual Information Retrieval,

1. Introduction

Automated fact-checking systems play a vital role in both countering false information on digital media and alleviating the burden on fact-checkers [1, 1, 2, 3, 4, 5, 6]. A key task of these systems is the detection of previously fact-checked similar claims – an information retrieval problem where claims serve as queries to retrieve from a corpus of debunks [7, 8, 9]. This task aims to detect claims that spread even after they have already been debunked by at least one professional fact-checker. Previous work has focused on training retrieval models, primarily focusing on monolingual retrieval, where the language of the query claim matches the language of the debunk [7, 8, 9, 10]. Moreover, these monolingual retrieval models assume that the debunks exist exclusively in one language. However, previous studies [11, 12, 13] demonstrate that similar false claims continue to spread in multiple languages, despite the availability of debunks for several months in another language. Hence, automatically finding debunks in multiple languages is crucial to make the best use of scarce fact-checkers' resources.

In this study, we define the task of **cross-lingual Debunked Narrative Retrieval (X-DNR)** as a cross-lingual information retrieval problem where a claim is used as a query to retrieve from a corpus of debunks in multiple languages (see Figure 1). In this paper, we use the term "debunked narrative retrieval" over the previously used term "fact-checked claim retrieval" because the term "debunked

Disinformation, Misinformation and Learning in the Age of Generative AI: Joint Proceedings of the 1st International Workshop on Disinformation and Misinformation in the Age of Generative AI (DISMISS-FAKE'25) and the 4th International Workshop on Investigating Learning during Web Search (IWILDS'25) co-located with 18th International ACM WSDM Conference on Web Search and Data Mining (WSDM 2025)

i.singh@sheffield.ac.uk (I. Singh); c.scarton@sheffield.ac.uk (C. Scarton); x.song@sheffield.ac.uk (X. Song); k.bontcheva@sheffield.ac.uk (K. Bontcheva)

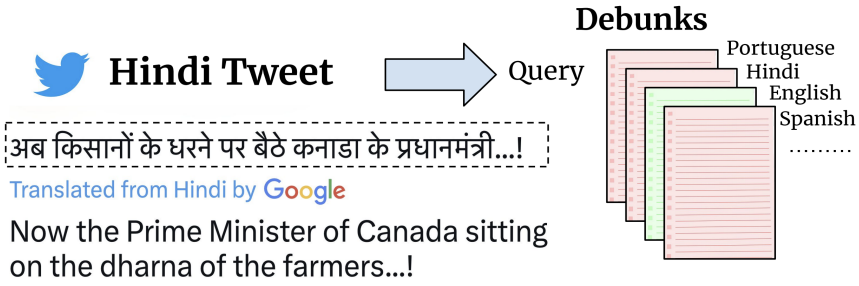


Figure 1: Cross-lingual debunked narrative retrieval: Query tweet is in Hindi and the relevant debunk is in English.

Table 1
Sample query tweets and their corresponding debunks from the **MMTweets** dataset.

Fields	Hindi Query Tweet - English Debunk	English Query Tweet - Spanish Debunk
Tweet	अब किसानों के धरने पर बैठे कनाडा के प्रधानमंत्री...! (English translation: Now the Prime Minister of Canada sitting on the farmers' dharna..!)	I Sultue you Sir. You are So intelligent. RUSSIA: Vladimir Putin has Dropped 800 tigers and Lions all over the Country to push people to stay Home...Stay Safe Everyone!!
Debunk title	Old Photo Passed Off As Justin Trudeau Sitting In An Anti-Farm Laws Protest	La foto del león en la calle fue tomada en Sudáfrica en 2016 y no tiene relación con la pandemia del COVID-19
Debunk claim	Justin Trudeau sits in protest in support of the protesting farmers.	Publicaciones compartidas más de 35.000 veces en redes sociales desde el 22 de marzo último aseguran que Rusia liberó..

narrative" better captures the range of false narratives or stories related to a claim that has already been debunked. This term acknowledges that a single claim can have multiple narratives, all needing debunking, unlike fact-checked claim retrieval, which focuses narrowly on verified claims without addressing their diverse associated narratives. Therefore, the term "debunked narrative retrieval" is more fitting for this task, as the primary objective of X-DNR is to aid fact-checkers in identifying debunked narratives across multiple languages. Our main contributions are:

- The **M**ultilingual **M**isinformation **Tweets** (**MMTweets**): a novel benchmark that stands out, featuring cross-lingual pairs, images, and fine-grained human annotations, making it a comprehensive resource compared to its counterparts (see Section 3.4). In total, it comprises $1,600$ query tweet claims (in Hindi, English, Portuguese & Spanish) and $30,452$ debunk corpus (in 11 different languages) for retrieval. Table 1 shows dataset examples.
- An extensive evaluation of state-of-the-art (SOTA) cross-lingual retrieval models on the MMTweets dataset. We also introduce two multistage retrieval methods (*BE+CE* and *BE+GPT3.5*) adapting earlier approaches to effectively address the cross-lingual nature of the X-DNR task. Nevertheless, the results suggest that dealing with multiple languages in the MMTweets dataset poses a challenge, and there is still room for improvement in models.
- A comprehensive evaluation aims to investigate: 1) cross-lingual transfer and generalisation across languages within MMTweets; 2) how challenging it is for models trained on existing datasets to transfer knowledge to the MMTweets test set; and 3) insights into the retrieval latency of different models (see Section 5).

In the following section, we discuss the related work. Section 3 details the MMTweets dataset. Section 4 presents the various experimental details related to the X-DNR task. The results are presented in Section 5 and we conclude the paper in Section 6.

2. Related Work

In order to minimise the spread of misinformation and speed up professional fact-checking, the initial verification step often involves searching for fact-checking articles that have already debunked similar narratives [14, 15, 16, 17]. Several benchmark datasets have been created for this task [7, 8, 18, 19, 20, 21, 22]. For instance, Shaar et al. [17] release a dataset of English claims and fact-checking articles from Snopes[23] and PolitiFact [24]. On the other hand, Vo and Lee [25] release a multimodal English dataset of tweet claims collected from Snopes and PolitiFact and investigates the use of images in tweets to retrieve previously fact-checked content. The CLEF CheckThat! Lab evaluations [9, 8, 7, 26] focus on a fully automated pipeline of fact-checking claims, where fact-checked claim retrieval is one of the steps in the claim verification workflow. They release a dataset of claims collected from Snopes, PolitiFact and AraFacts [27] and ClaimsKG [28]. However, the aforementioned work only focuses on monolingual scenarios where the claim and debunk share the same language. In contrast, our MMTweets dataset includes cross-lingual cases, making it more challenging. For a detailed comparison of different datasets with our MMTweets, please refer to Section 3.4. We also test domain overlap between MMTweets and other datasets in Section 5.3.

Prior work on claim matching [10] release a dataset of claims collected from tiplines on WhatsApp [10] and conduct retrieval experiments. Although they present results for multiple languages, their dataset only includes monolingual pairs [10], thereby hindering the development of retrieval models capable of detecting debunked narratives in multiple languages. Finally, the closest match to our work [29] focuses on cross-lingual claim matching. They release a dataset of debunked tweets sourced from the International Fact-Checking Network (IFCN) [30] and some other fact-checking aggregators [29]. However, their dataset lacks diverse cross-lingual pairs (see Section 3.4), and tweet claims are automatically extracted from debunk articles [29], which can result in false positives. In contrast, our dataset has diverse cross-lingual pairs, and each tweet in MMTweets undergoes manual annotation to ensure high-quality data (see Section 3). Moreover, prior work [29] does not train custom debunked narrative retrieval models or perform cross-lingual and cross-dataset transfer testing, a gap that we address in this paper with a specific focus on the MMTweets dataset (Section 5).

Furthermore, Kazemi et al. [10] found that multistage retrieval [31] using BM25 and XLM-RoBERTa transformer [32] re-ranking can beat the competitive BM25 baseline for debunked narrative retrieval. However, the use of multistage retrieval with BM25 and transformer model re-ranking, as demonstrated in prior work [10, 17, 31, 33], introduces translation overhead for BM25 in cross-lingual scenarios where the query claim and document languages differ. To address this, this paper introduces translation-free multistage retrieval methods, employing both bi-encoders and cross-encoders for the X-DNR task (Section 4.1). Additionally, due to dataset limitation, much of the prior research [17, 8, 7] trains retrieval models using debunks available from a single fact-checking organisation. In contrast, our MMTweets dataset involves debunks from multiple fact-checking organisations (Section 3). This enables the development of retrieval models that are agnostic to debunk structure, a crucial aspect for X-DNR, as relevant debunks can originate from any fact-checking organisation.

3. MMTweets Dataset

MMTweets is a new dataset of misinformation tweets annotated with their corresponding debunks (or fact-checks), both available in multiple languages. MMTweets primarily comprises of tweets related to COVID-19 misinformation in English, Hindi, Portuguese and Spanish. The languages of tweets were selected based on two criteria: 1) these are the most frequent languages in previous publicly available COVID-19 misinformation datasets [34, 12]; 2) the chosen languages are among some of the most widely spoken ones worldwide. The dataset was built in two steps: first, the raw data was collected, followed by manual data annotation.

3.1. Raw Data Collection

First, we collect debunk narratives published by different fact-checking organisations covering our target languages. For this, we collect a total of $30,452$ debunk articles from the following organisations (language in brackets): Boomlive (English) [35], Agence France-Presse (AFP) (German, English, Arabic, French, Spanish, Portuguese, Indonesian, Catalan, Polish, Slovak and Czech) [36], Agencia EFE (Spanish) [37] and Politifact (English) [24]. For each debunk article, we collect the following information fields: the article title, the debunked claim statement and the article body.

Next, we select a sample of $1,600$ debunk articles from the corpus of $30,452$ debunk articles based on two specific criteria. Firstly, we focus on debunks published between January 2020 and March 2021, allowing for temporal and topical diversity as the COVID-19 pandemic unfolded. This approach, given the global nature of the pandemic, maximises the chance of including similar narratives spreading in multiple languages. Secondly, our aim is to maximise instances where the language of the potential misinformation tweets mentioned in the debunk articles differs from that of the debunk article itself. For example, while Boomlive publishes debunk articles in English, the associated tweets may be in Hindi. Overall, this careful selection of debunks ensures comprehensive cross-lingual coverage within the MMTweets dataset (Section 3.5).

Finally, following the previous work [17, 29], we extract all the tweets mentioned in the debunk article body. We use Twitter API [38] to get tweet details including tweet text and attached media (if any). We chose Twitter because of its easy open access as compared to other social media platforms at the time of this study.

3.2. Data Annotation – Tweet Classification

Table 2
Details of the MMTweets dataset: class count, Fleiss Kappa and textual misinformation ratio. Please note that the class count does not sum up to the total tweet count due to the overlap between textual and non-textual misinformation cases.

Language	Tweet Count	Class Count				Fleiss Kappa	Textual Misinformation Ratio
		Textual Misinformation	Non-textual Misinformation	Debunk	Other		
Hindi	400	328	254	11	27	0.53	0.86
Portuguese	400	310	200	5	30	0.59	0.77
English	400	247	166	68	82	0.79	0.61
Spanish	400	291	233	14	62	0.57	0.70
Total	1600	1176	853	98	201	Average: 0.62	Average: 0.74

The approach described in Section 3.1 does not guarantee that the extracted tweets from debunk articles contain text-based misinformation. We found that some contained only images or videos, while others made general comments or debunked the misinformation itself. Therefore, the extracted tweets were classified manually to create gold-standard data for evaluation. In particular, we recruited 12 student volunteers[1] who were native speakers of either English, Hindi, Portuguese or Spanish (three native speakers per language). The annotators were shown all debunk information fields and asked to annotate the tweets as belonging to one of three classes:

- **Misinformation tweets:** with two sub-classes – **A) Textual misinformation**, if the textual part of a tweet expresses the false claim which is being debunked by the fact-checking article. **B) Non-textual misinformation**, if a tweet contains misinformation in image or video only. Please note that a tweet can have both text and non-textual misinformation. For such cases, annotators were asked to label the tweet as having both "textual misinformation" and "non-textual misinformation".

[1]The dataset annotation received ethical approval from the University of Sheffield Ethics Board (Application ID 040156). This paper only discusses analysis results in aggregate, without providing examples or information about individual users.

- **Debunk tweets:** If the tweet does not express misinformation uncritically, but instead exposes the falsehood of the claim.
- **Other tweets:** If the tweet is neither "misinformation" nor "debunk", then it is classified as "other". For instance, this can be a general comment or a general enquiry relevant to the false claim that is being debunked.

Please refer to the annotation codebook[2] for examples of misinformation, debunk, and other tweets. To ensure data quality, we first conducted training sessions with the annotators and went through several examples to familiarise them with the task. We also had a final adjudication step, where problems and disagreements flagged by the annotators were resolved by domain experts. For instance, there were some tweets which agreed with the misinformation but did not state it directly or the annotator was unsure about the claim's veracity. All such cases were considered "other" due to the chosen narrower definition of misinformation tweets.

A total of $1,600$ tweets were annotated, resulting in approximately 400 tweets per language (see Table 2). Following previous methodology [39, 29], a total of 400 tweets (100 per language) were triple annotated to compute inter-annotator agreement (IAA) and the final category was chosen by majority voting. Table 2 reports Fleiss Kappa scores which indicate moderate to substantial IAA for all languages. Table 2 also shows the textual misinformation ratio (i.e. the proportion of tweets annotated as "textual misinformation" out of all annotated tweets) for each language. The ratio is variable due to the varied nature of the debunks in each language and the different ways in which fact-checkers refer to misinformation-bearing tweets. On average, textual misinformation comprised 74% of all the classified tweets in the dataset.

3.3. Data Annotation – Claim Matching

The annotations gathered in Section 3.2 only pertain to tweets mentioned in the debunk articles, indicating a one-to-one relationship between tweets and debunks. However, prior research [12, 11] demonstrates that there can be various potential debunks for the same misinformation. To address this and establish a one-to-many relationship between misinformation tweets and debunks, we conduct a subsequent round of annotations to identify comparable debunks. However, annotating relevance judgments between tweets and all the previously collected 30,452 debunks is not feasible. Therefore, we take debunked claim statements linked to each tweet and compute cosine similarity[3] with all 30,452 debunked claim statements in the hope of finding similar debunked claim statements. To ensure this, we select the *top-k* matching claim statements for annotation, with a depth of seven as per previous work [40]. We also retain only those claim pairs with a similarity score exceeding the 0.6 threshold to exclude irrelevant claim pairs from the annotations. Finally, annotators classified 4,594 pairs of debunked claim statements into exact match, partial match, or irrelevant (3-level) using previously published annotation guidelines [10]. Examples for each class can be found in the annotation codebook[4].

The annotations were conducted on the GATE Teamware annotation tool [41] – refer to the annotation codebook for examples of the tool's user interface. A total of 14 PhD researchers were recruited to manually annotate pairs of debunked claim statements. To ensure high-quality annotations, we conducted a pre-annotation phase. An initial annotator training session familiarises them with the instructions. Subsequently, annotators were asked to annotate a certain number of test samples. We then review these annotations and only those annotators who correctly classify at least 80% of the samples proceed with further annotations. Based on prior research [42], we also ask annotators to provide a confidence score for each annotation, and we further discard annotations with low confidence scores to maintain data quality. Finally, following prior works [40, 43, 44], we find the IAA Kappa to be 0.5 on a subset of the data using triple annotations, suggesting a moderate level of agreement among the annotators. All annotators were paid at a standard rate of 15 GBP per hour for their work.

[2] https://doi.org/10.5281/zenodo.10637161

[3] We use the best performing Sentence-transformer model *all-mpnet-base-v2* on English-translated statements (Ref. https://www.sbert.net/docs/pretrained_models.html)

[4] The annotation codebook is available in Supplementary files.

Table 3
Complete summary of the MMTweets dataset.

Language	Hindi	Portuguese	English	Spanish	Total
Query Tweets	400	400	400	400	1600
Exact Match	518	742	812	644	2716
Partial Match	417	409	342	374	1542
Irrelevant	475	656	337	468	1936

Table 3 presents a summary of the complete MMTweets dataset, including the number of query tweets and the count of query tweet and debunk pairs for 3-level relevance annotations. Specifically, it includes 2,716 exact matches, 1,542 partial matches, and the remaining are categorised as irrelevant (see Section 3.5 for count in different language pairs). The average word count in query tweets is 28 ± 14.3 (1 std). There are a total of 1,600 tweets in MMTweets, and on average, each tweet is linked with 2.7 ± 2.0 (1 std) debunks, either exact or partial match. Please note that the one-to-many relation between query tweets and the debunks enriches our dataset to include cases beyond the tweets mentioned in the debunk articles. Additionally, the fine-grained classification of debunks into exact and partial matches serves as fine-grained labels for our subsequent information retrieval experiments (see Section 4.1).

3.4. Comparison to Existing Datasets

Table 4 provides a comparison between MMTweets and the existing datasets, revealing favourable query claim counts in our dataset compared to others. Notably, MMTweets stands out with 43% cross-lingual instances across various language pairs (see Section 3.5). This is in stark contrast to the sole existing cross-lingual dataset [29], which only comprises 10% of Hindi-English pairs, where the claim is in Hindi and the debunk is in English. Additionally, all tweets in MMTweets undergo manual annotation, unlike other existing datasets [45, 29, 17], where tweets are automatically extracted from fact-check articles, potentially leading to false positives. Moreover, automated extraction of tweets also leads to missing one-to-many connections between claims and debunks as shown in prior work [14]. Furthermore, MMTweets provides 3-level graded relevance scores (fine-grained) for query-passage pairs, unlike prior work which use binary relevance scores (coarse-grained) [17, 7, 8].

Among other datasets, Shaar et al. [17] and CLEF variants lack cross-lingual pairs, images, and fine-grained labels. The larger Vo and Lee [25] dataset incorporates images and human annotations but lacks fine-grained labels. CrowdChecked [45] contains a massive volume of claims but lacks crucial features like manual annotations and cross-lingual pairs. Although prior work [29, 10] provide multilingual support, it's impossible to replicate or conduct comparative experiments on their datasets because they do not release the corpora of debunks used in the retrieval experiments – only the query claims are released. Moreover, it lacks images, manual annotations and fine-grained labels [29]. In contrast, our MMTweets dataset stands out, featuring cross-lingual pairs, images, human annotations, and fine-grained labels, making it a comprehensive resource compared to its counterparts. Additionally, we examine the domain overlap between MMTweets and other datasets, revealing a low degree of overlap (refer to Section 5.3).

3.5. Data Analysis

To assess the linguistic diversity, Table 5 shows the count of query tweet and debunk pairs for different languages[5]. In particular, there are a total of 4,258 positive pairs (exact and partial matches) of tweets and their corresponding debunks. Among these, 1,809 instances (43%) are pairs where the language of tweets and debunks is different (cross-lingual). This makes our dataset the one with the highest proportion of cross-lingual instances when compared to existing datasets (see Section 3.4). The majority

[5]We use *langdetect* (https://pypi.org/project/langdetect/) for detecting the language.

Table 4
Comparison of debunked narrative retrieval datasets: "Lang" denotes the count of different languages of claims; "Cross" indicates the presence of cross-lingual pairs; "Img" indicates whether the dataset is multi-modal and includes images; "Ant" indicates whether the dataset is human-annotated or automatically extracted from articles; "Fine" indicates the availability of fine-grained labels.

Dataset	Items	Lang	Cross	Img	Ant	Fine
Shaar et al. [17]	1,768	1	×	×	×	×
CLEF20-EN	1,197	1	×	×	✓	×
CLEF21 2A-EN	2,070	1	×	×	×	×
CLEF21 2A-AR	858	1	×	×	✓	×
CLEF22 2A-EN	2,362	1	×	×	×	×
CLEF22 2A-AR	908	1	×	×	✓	×
Vo and Lee [25]	13,239	1	×	✓	✓	×
CrowdChecked [45]	330,000	1	×	×	×	×
Kazemi et al. [10]	382	5	×	×	✓	✓
Kazemi et al. [29]	6,533	4	✓	×	×	×
MMTweets (ours)	1,600	4	✓	✓	✓	✓

Table 5
Language of tweet and debunk pairs in MMTweets. Language codes are ISO 639-1 representations for Portuguese (PT), Spanish (ES), Hindi (HI), English (EN), Indonesian (ID), Slovak (SK), Catalan (CA), Polish (PL), Czech (CS), and French (FR).

Tweet Language	PT	ES	HI	EN	EN	EN	PT	EN	EN	ES	EN	EN	EN	PT	EN	ES	HI	PT	HI	ES	Total
Debunk Language	PT	ES	EN	EN	ES	ID	ES	PT	SK	CA	PL	CA	CS	ID	FR	ID	PT	EN	FR	EN	
Count	1045	954	925	450	332	158	80	65	53	50	30	27	22	22	17	11	7	4	3	3	4,258

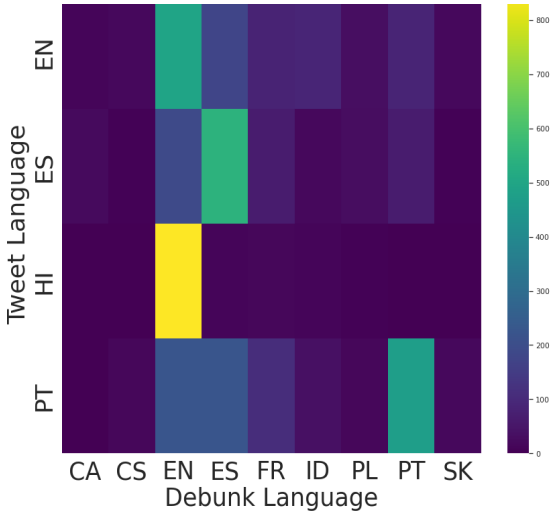


Figure 2: Cross-language analysis: tweet vs. debunk.

of these cross-lingual pairs have tweets in Hindi and corresponding debunks in English, followed by instances with tweets in English and debunks in Spanish.

Figure 2 displays the heatmap illustrating language dynamics of tweets and its related debunks in MMTweets. Notably, multiple languages exhibit near-zero associated debunks in languages besides English (e.g., Hindi), suggesting a potential gap in fact-checking coverage for specific languages. This emphasises the need to address disparities in debunk distribution and highlights opportunities for automated cross-language fact-checking methods like X-DNR.

Please refer to Appendix A.1, which presents the temporal characteristics of tweets, providing a month-by-month breakdown of tweet counts for each language in MMTweets. Appendix A.2 presents

the results of topic modelling using Latent Dirichlet Allocation.

4. Cross-lingual Debunked Narrative Retrieval (X-DNR)

In this section, we formally define the X-DNR task. Given a tweet claim as a query t, the X-DNR system employs a retrieval model to obtain a candidate set of debunked narratives from a larger corpus of debunks $D = \{d_i\}_{i=1}^{D}$ in multiple languages. The final trained model can be expressed as X-DNR(t, D), whose ultimate goal is to provide the most accurate fact-checking information to users in response to potential misinformation claims in any language.

In this paper, we exclusively focus on textual misinformation cases (totalling 1,176, as shown in Table 2). For the retrieval corpus, we utilise a collection of $30,452$ previously gathered debunks in multiple languages (refer to Section 3.1). Each debunk comprises a concatenated debunked claim and article title field (Section 3.1).

4.1. Cross-lingual Retreival Models

We test the following cross-lingual retrieval models on MMTweets.

Okapi BM25. We utilise the ElasticSearch [46] implementation of BM25 [46] with default parameters ($k = 1.2$ and $b = 0.75$). Since BM25 is designed for monolingual retrieval, we employ machine translation using the Fairseq's `m2m100_418M` model [47] to make it applicable to cross-lingual query and document pairs. All non-English tweets and debunks are translated into English, and the complete corpus of debunks is indexed in ElasticSearch [46]. We then use the English-translated tweets as queries over the debunks.

xDPR Dense Passage Retrieval (DPR) [48], an early dense retrieval model, uses BERT-based encoders for queries and documents to assess relevance based on their similarity. To expand its support beyond English, we use a multilingual variant, xDPR [49, 50], which is an XLM-RoBERTa [32] model fine-tuned on the MSMARCO dataset [51]. We further fine-tune xDPR on our MMTweets [49].

mContriever Izacard et al. [52] introduced mContriever, which employs contrastive loss for unsupervised pretraining of mBERT [53], showing enhanced performance on various IR tasks. We use the provided multilingual checkpoint [54], already fine-tuned on MSMARCO [51]. We further fine-tune this model on MMTweets, employing the same methodology as described in Izacard et al. [52].

Bi-Encoder (BE) We fine-tune different Multilingual Pretrained Transformer (MPT) models as bi-encoders [55, 48] on pairs of query tweets and their corresponding debunks. The objective function employed is the mean squared error, measuring the disparity between the true label and the model-calculated relevance score for the tweet-debunk pair. This adjusts model parameters, aligning the embedding of a query tweet closer to its relevant debunks in the vector space. The loss equation is as follows,

$$\mathcal{L}(\theta) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \left(\mathcal{Y}_i - \left(\frac{f_\theta(t_i) \cdot f_\theta(d_i)}{\|f_\theta(t_i)\|_2 \|f_\theta(d_i)\|_2} \right) \right)^2 \tag{1}$$

where f_θ is the shared MPT encoder for tweet t_i and debunk d_i, $\mathcal{Y}_i$ represents the true label of the i-th sample. The relevance score between tweet and debunk is computed using cosine similarity. We employ cosine similarity with the mean-pooling technique due to its proven effectiveness in prior research [55].

We fine-tune bi-encoder using five different MPT models, namely multilingual BERT (mBERT) [53], XLM-RoBERTa (XLMR) [32] and Language-Agnostic BERT Sentence Embedding (LaBSE) [56]. Additionally, we also fine-tune two Sentence-Transformer model variants i.e. Universal Sentence

Encoder (USE) [57, 58] and Masked and Permuted Pretraining for Language Understanding (MPNet) [59, 60]. These bi-encoder models are denoted by the prefix "BE-" in subsequent experiments (Section 5.1).

Multistage Retrieval Drawing inspiration from the success of multistage retrieval methods in IR tasks [31, 33, 61], we apply these techniques to the X-DNR task. Within this context, we introduce two methods that adapt earlier approaches, specifically tailored for the X-DNR task. These methods are as follows:

- **Bi-Encoder+Cross-Encoder (*BE+CE*):** In the first retrieval stage, we fine-tune an MPT model as a bi-encoder instead of the standard BM25-based lexical retrieval approach adopted in prior work [17, 10]. This choice is motivated by the MPT model's suitability for the cross-lingual nature of the task, eliminating the need for translation. In the second stage, we fine-tune an MPT model as a cross-encoder [31] to re-rank the top-K retrieved debunks from the first stage. Here, the model employs self-attention mechanisms on the given tweet and debunk pair to get the final relevance score. The input to the model follows the structure: $[CLS]\ [T_1]...[T_n]\ [SEP]\ [DC_1]...[DC_i][DT_1]...[DT_j]$, where T_n are the tweet subword tokens and DC_i and DT_j are the debunked claim and title subword tokens, respectively. $[CLS]$ and $[SEP]$ are the default tokens to indicate "start of input" and "separator", respectively, in the Next Sentence Prediction task [53].
- **Bi-Encoder+ChatGPT (*BE+GPT3.5*):** Large language models like ChatGPT (*gpt-3.5-turbo*) have consistently showcased impressive capabilities across a broad spectrum of natural language processing tasks [62]. Therefore, to evaluate ChatGPT's performance, we implement a Listwise Re-ranker with a Large Language Model (LRL) [63] to re-rank documents retrieved by the first stage ranker. The main distinctions in our approach compared to prior work [63] are: 1) we employ multilingual bi-encoders described in Section 4.1 as the first-stage ranker 2) each re-ranked document consists of concatenated debunk claim and title fields. All parameters are kept the same as used by [63].

4.2. Experimental Details

4.2.1. Train and test sets

We divide 1,176 textual misinformation tweet queries into train and test sets. The test set consists of 400 tweet queries (100 queries per language), comprising the same triple-annotated tweets used for calculating IAA. The remaining 776 tweet queries are used as training data, with a 10% subset used as a validation set. Please note that during test time, we do not know if a tweet has been debunked, because tweets linked with debunks in the test set do not occur in the train set. This ensures a realistic test scenario by preventing tweets linked to the same debunk from appearing in both the train and test sets.

Now, since each query tweet in the training set is linked to multiple debunks (Section 3.3), therefore, the final training set comprises 2,360 positive (1,420 exact matches and 940 partial matches) tweet and debunk pairs. For negative pairs, ten debunks are randomly sampled for each tweet, resulting in total 7,760 negative tweet and debunk pairs. We also experimented with hard negative mining and higher counts of negatives, but did not observe any significant improvements. In total, the training set consists of 10,120 fine-grained tweet and debunk pairs for training different retrieval models.

4.2.2. Hyperparameters

The bi-encoder is trained for four epochs with a batch size of 32, a learning rate of $4e-5$ and maximal input sequence length of 256. The cross-encoder, trained for two epochs, uses a batch size of 16, $4e-5$ learning rate, with truncation of subword tokens beyond 512. Both models employ linear warmup, AdamW optimiser, and manual hyperparameter tuning on a validation set. Hyperparameter bounds are set as: 1) 1 to 5 epoch 2) $1e-5$ to $5e-5$ learning rate 3) 8 to 64 batch size on NVIDIA RTX 3090.

Table 6
Results for different cross-lingual retrieval models on the test set of MMTweets. The best scores are in bold.

Language	Metric	BM25	xDPR	mCont	BE-mBERT	BE-XLMR	BE-USE	BE-LaBSE	BE-MPNet	BE+CE	BE+GPT3.5
MMTweets-HI	**nDCG@1**	0.263	0.435	0.240	0.135	0.160	0.210	0.525	0.320	**0.610**	0.575
	nDCG@5	0.267	0.421	0.304	0.149	0.188	0.246	0.514	0.366	**0.569**	0.527
	MRR	0.320	0.503	0.352	0.199	0.250	0.310	0.623	0.439	**0.674**	0.637
MMTweets-PT	**nDCG@1**	0.625	0.695	0.770	0.540	0.685	0.730	0.755	0.755	**0.845**	0.840
	nDCG@5	0.598	0.690	0.761	0.514	0.595	0.672	0.726	0.720	**0.765**	0.757
	MRR	0.723	0.781	0.849	0.627	0.737	0.782	0.822	0.821	**0.887**	0.880
MMTweets-EN	**nDCG@1**	0.591	0.635	0.705	0.515	0.465	0.675	0.680	0.710	**0.720**	0.715
	nDCG@5	0.572	0.625	0.670	0.475	0.472	0.638	0.650	**0.696**	0.682	0.662
	MRR	0.706	0.759	0.801	0.603	0.590	0.760	0.780	**0.814**	**0.814**	0.807
MMTweets-ES	**nDCG@1**	0.560	0.620	0.610	0.405	0.435	0.500	0.585	0.615	**0.735**	0.660
	nDCG@5	0.525	0.621	0.646	0.394	0.428	0.497	0.582	0.582	**0.662**	0.632
	MRR	0.648	0.707	0.730	0.491	0.536	0.591	0.670	0.678	**0.804**	0.741
Average	**nDCG@1**	0.510	0.596	0.581	0.399	0.436	0.529	0.636	0.600	**0.728**	0.698
	nDCG@5	0.490	0.589	0.595	0.383	0.421	0.513	0.618	0.591	**0.669**	0.644
	MRR	0.599	0.687	0.683	0.480	0.528	0.611	0.724	0.688	**0.795**	0.766

5. Results and Discussion

In this section, we present the results of retrieval experiments that aim to address the following five research questions:

RQ1 To what extent do the current SOTA cross-lingual retrieval models perform in addressing the specific challenges posed by the MMTweets dataset? (Section 5.1)
RQ2 How challenging is it for models to transfer and generalise across languages within MMTweets? (Section 5.2)
RQ3 Can models trained on existing datasets transfer knowledge and generalise on the MMTweets test set? (Section 5.3)
RQ4 What insights can be gained into the retrieval latency of various cross-lingual retrieval models? (Section 5.4)

5.1. Model Performance

Table 6 shows Mean Reciprocal Rank (MRR) and Normalised Discounted Cumulative Gain (nDCG@1 & nDCG@5) on the test set of MMTweets (HI, PT, EN & ES). The results suggest that BE-mBERT and BE-XLMR consistently show lower scores, with occasional lower performance when compared to BM25. BM25's strength lies in lexical overlap with machine-translated text, giving it an advantage over other models. However, other retrieval models outperform BM25 on several metrics. Notably, BE-LaBSE performs better than BE-MPNet, BE-USE, BE-mBERT, and BE-XLMR, even outperforming state-of-the-art models like xDPR and mContriever in average metric scores. This is attributed to LaBSE's sentence-level objective, combined with pretraining techniques involving translation and masked language modelling, as discussed in Feng et al. [56].

The last two columns of Table 6 report the scores of multistage retrieval methods (*BE+CE* & *BE+GPT3.5*). In multistage retrieval, we employ LaBSE for the first stage due to its superior performance over other models (see Table 6). Similarly, the second stage in *BE+CE* also utilises LaBSE, with the number of re-ranked documents set to 20. Although we experimented with various MPT models and different counts of re-ranked documents in the second stage, no significant improvements were observed. The results show that *BE+CE* consistently emerges as the top performer across all datasets and metrics, achieving an average nDCG@1 score of 0.728, an average nDCG@5 score of 0.669, and an average MRR score of 0.795 (Table 6 – second last column). On the other hand, while *BE+GPT3.5* outperforms other models in average metric scores, its retrieval latency is the highest (see section 5.4). Although other models like BE-LaBSE, BE-MPNet, xDPR, and mContriever showcase competitive performance, none consistently match the performance demonstrated by multistage retrieval methods. Additionally, despite being trained on the extensive MSMARCO training dataset (Section 4.1), models

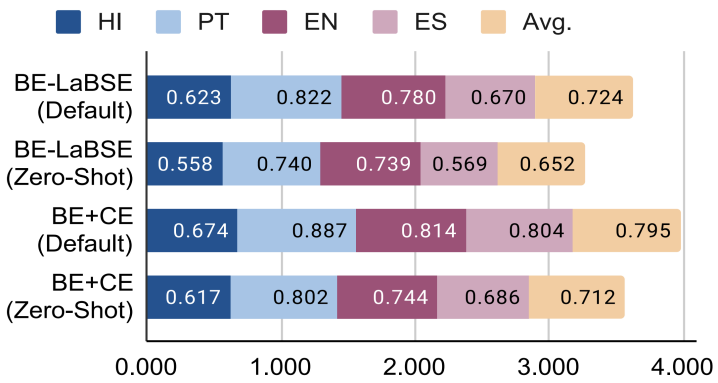


Figure 3: Stacked bar plot for MRR scores for zero-shot cross-lingual transfer and the default results (from Table 6).

such as xDPR and mContriever do not notably enhance performance, suggesting distinctive challenges presented by MMTweets.

For *BE+CE*, the extent of improvement varies across languages. For example, in the case of Portuguese, *BE+CE* outperforms BM25 with increases of 132% for nDCG@1, 112% for nDCG@5, and 110% for MRR. Conversely, the improvement is relatively low for English, with increases of only 22%, 19%, and 15% for nDCG@1, nDCG@5, and MRR scores, respectively. We hypothesise that this disparity in performance across different languages may be attributed to noisy translations in the case of BM25, while *BE+CE* doesn't rely on translation. Additionally, the scores on different languages are reflective of how topics found in each language impact a model's performance. For example, the Hindi tweets have the lowest performance across all models and evaluation metrics, which suggests that the topics found in these languages (Appendix A.2) are quite challenging for the model. Another reason for poor Hindi performance could be the change in the language script to Devanagari. This suggests that dealing with various languages in MMTweets poses a challenge, and there is still potential for improvement in retrieval models.

Furthermore, we also observe the challenge of distinguishing closely related debunks by the model. This occurs when the retrieved debunk is not entirely relevant, but still shares some degree of relevance with the query claim. For instance, consider the query claim about the sighting of crocodiles in the flooded streets of Hyderabad; the top-retrieved debunks are closely related, involving sightings of crocodiles in Mumbai, Bengaluru, Florida, etc. This highlights the need for continued refinement in retrieval models to enhance the relevance of top-ranked debunks for the X-DNR task.

In summary, these evaluations highlight performance differences among models, emphasising the consistent superiority of multistage retrieval methods across various languages and metrics. While BM25 is faster (see Section 5.4), the necessity of machine translation for BM25 incurs additional costs and time overheads.

5.2. Cross-lingual Transfer

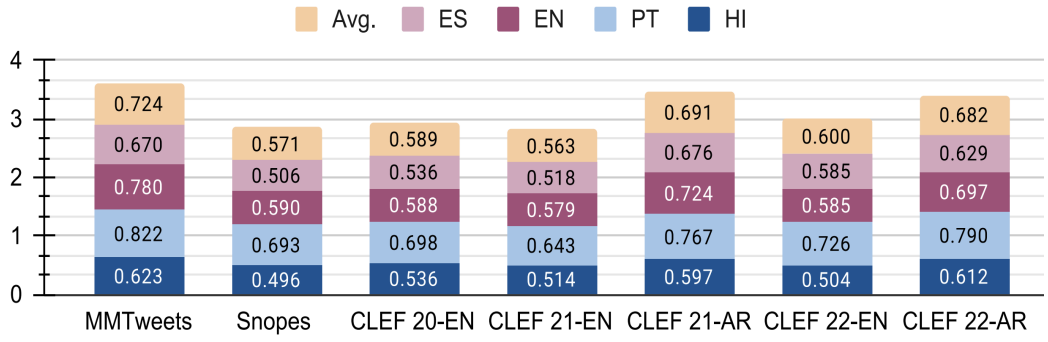
To test the zero-shot transfer capabilities, the model is trained on languages other than the one it is tested on. For instance, to test zero-shot transfer for Hindi, the models are trained on only those tweet and debunk pairs that are not in Hindi. Hence, in total four models are trained for four different languages in the MMTweets.

We evaluate the cross-lingual transfer capability of BE-LaBSE and *BE+CE*, which yield the highest average scores (Table 6). Figure 3 shows a stacked bar plot illustrating MRR scores for zero-shot cross-lingual transfer and the default results sourced from Table 6.

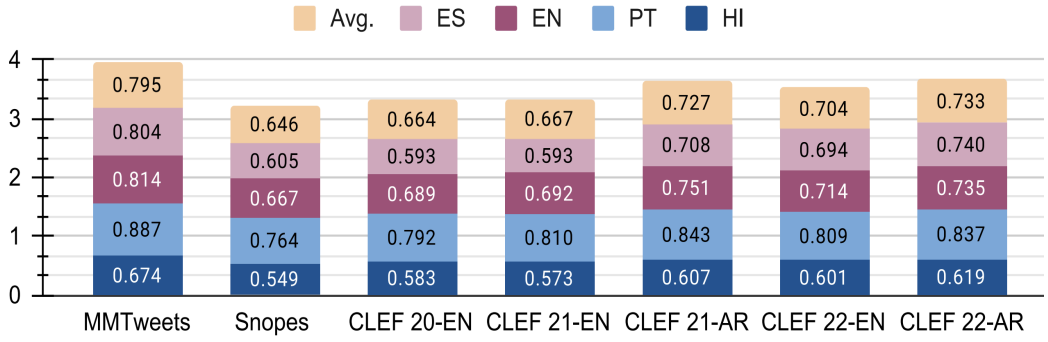
When comparing the zero-shot results with the default results, the default results consistently outperform zero-shot results for both models (BE-LaBSE and *BE+CE*) across all languages, as expected due to training on the complete dataset. Nevertheless, zero-shot models surpass several baselines, including BM25 (from Table 6) in this challenging setting. The results suggest that models have the

Table 7
Domain overlap between the test set of MMTweets and the train set of other datasets.

Train Set	MMTweets	Snopes	CLEF 20-EN	CLEF 21-EN	CLEF 21-AR	CLEF 22-EN	CLEF 22-AR
Overalp	0.29	0.15	0.14	0.12	0.16	0.11	0.13



(a) BE-LaBSE



(b) BE+CE

Figure 4: Stacked bar plot for MRR scores for zero-shot cross-dataset transfer using BE-LaBSE (a) and BE+CE (b).

potential to transfer knowledge between languages without the need for language-specific training. This also supports prior observations that MPT models, when fine-tuned on monolingual data, exhibit strong performance on a different language [52, 64]. Despite these promising outcomes, there is still room for improvement for zero-shot models to match the performance of default models.

5.3. Cross-dataset Transfer

To test the zero-shot cross-dataset transfer capabilities of the models, we train them on the training set of previously published datasets and subsequently evaluate their performance on the test set of MMTweets. This ensures real-life testing to assess the generalisability of the models. The previously published datasets include Snopes [17] and CLEF CheckThat! Lab task datasets which include CLEF 22-EN and CLEF 22-AR [7], CLEF 21-EN and CLEF 21-AR, [8] and CLEF 20-EN [9]. Please note that CLEF 22-AR and CLEF 21-AR are Arabic datasets while other datasets are in English.

First, we assess the domain overlap to see how challenging it is for models trained on existing datasets to transfer knowledge to the MMTweets test set. For this, we use weighted Jaccard similarity [65] to compute the domain overlap between the test set of MMTweets and the train set of other datasets used for cross-dataset analysis (Table 7). We also report the overlap between the train and test set of MMTweets for reference. We find low domain overlap (ranging from 11-16%) with other datasets' train sets compared to MMTweets' train set (which has a 29% overlap) indicating distinct or less common instances between MMTweets and other datasets. We also conducted this analysis for each language but didn't find much variation in the results. Overall, MMTweets stands out as a unique dataset, showing

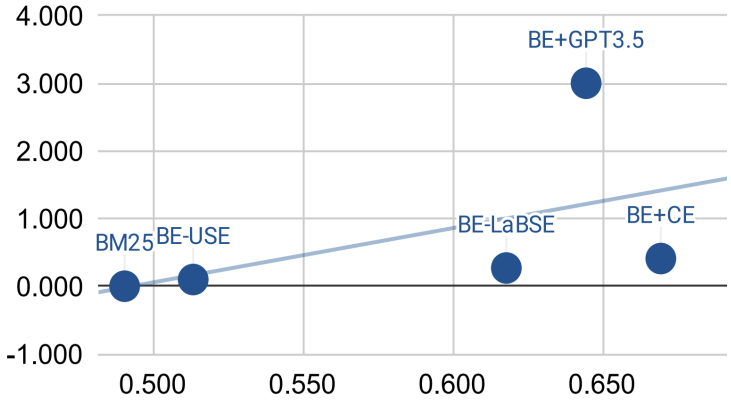


Figure 5: Retrieval latency (in seconds) and MRR scores.

low domain overlap with existing datasets.

Figure 4 shows MRR scores for zero-shot cross-dataset transfer using BE-LaBSE (a) and *BE+CE* (b), alongside default MMTweets trained results (from Table 6). Notably, models trained on CLEF 21-AR and CLEF 22-AR, despite being in Arabic, achieve the highest scores across all languages after the default MMTweets trained models. Additionally, models fine-tuned on CLEF 22-EN and 20-EN closely compete with other retrieval models (Table 6). Notably, while all claims in other datasets are either in English or Arabic, the MMTweets test set encompasses multiple other languages, making it even more challenging to retrieve the best matching debunk.

Overall, the findings suggest some knowledge transfer between datasets, which is especially valuable when obtaining a domain-specific dataset for training a dedicated model is challenging. However, despite these positive outcomes, there remains potential for models to match or surpass default MMTweets trained results.

5.4. Retrieval Latency

Figure 5 shows scatter plot for the average MRR scores and retrieval latency for different models. Lower values in the retrieval latency indicate faster query processing by the IR model. When comparing models, *BE+CE* achieves the highest MRR score (0.669) but exhibits a latency of 0.41 seconds, indicating a comparatively longer retrieval time. BE-LaBSE follows closely with an MRR score of 0.618 and a moderate retrieval latency of 0.27 seconds, striking a balance between performance and retrieval speed. While *BE+GPT3.5* displays a competitive MRR score (0.644), its retrieval latency increases to 3 seconds, impacting its practical application in real-time scenarios. BM25, although has the fastest retrieval latency at 0.001 seconds, it compromises ranking quality with the lowest MRR score of 0.490.

Overall, BE-LaBSE provides a balanced option with reasonable performance and moderate retrieval latency, while *BE+CE* excels in ranking quality, albeit with a slightly longer retrieval latency.

6. Conclusion and Future Work

This paper focuses on cross-lingual debunked narrative retrieval (X-DNR) for automated fact-checking. It introduces MMTweets, a novel benchmark dataset that stands out, featuring cross-lingual pairs, human annotations, fine-grained labels, and images, making it a comprehensive resource compared to other datasets. Furthermore, initial tests benchmarking SOTA cross-lingual retrieval models reveal that dealing with multiple languages in the MMTweets dataset poses a challenge, indicating a need for further improvement in retrieval models. Nevertheless, the introduction of tailored multistage retrieval methods demonstrates superior performance over other SOTA models, achieving an average nDCG@5 of 0.669. However, it's crucial to note the trade-offs between model performance and retrieval latency, with *BE+CE* offering better ranking quality at the expense of longer retrieval times. Finally, the findings

also suggest some knowledge transfer across languages and datasets, which is especially valuable in scenarios where language-specific models are not available or feasible to train. However, despite these positive outcomes, there is still room for models to match or even surpass the performance of default MMTweets trained models. To achieve this objective, the model needs an in-depth understanding of both language and context, along with the capability to differentiate among closely related debunked narratives. More sophisticated models could potentially introduce these capabilities in the future.

In future, we plan to extend the dataset to include claims from other social media platforms and domains to enhance its generalisability. Additionally, we aim to explore multimodal debunked narrative retrieval, leveraging information from various modalities.

Acknowledgments

This research is supported by a UKRI grant EP/W011212/1, an EU Horizon 2020 grant (agreement no.871042) ("SoBigData++: European Integrated Infrastructure for Social Mining and BigData Analytics" (http://www.sobigdata.eu)) and a European Union grant INEA/CEF/ICT/A2020/2381686 (European Digital Media Observatory (EDMO) Ireland, https://edmohub.ie/).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. S. Nielsen, R. McConville, Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3141–3153.

[2] X. Shang, Y. Chen, Y. Fang, Y. Liu, S. Vincent, Amica: Alleviating misinformation for chinese americans, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 3145–3149.

[3] J. Wu, Q. Liu, W. Xu, S. Wu, Bias mitigation for evidence-aware fake news detection by causal intervention, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2308–2313.

[4] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[5] D. Zhang, A. Vakili Tahami, M. Abualsaud, M. D. Smucker, Learning trustworthy web sources to derive correct answers and reduce health misinformation in search, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2099–2104.

[6] X. Zeng, A. S. Abumansour, A. Zubiaga, Automated fact-checking: A survey, Language and Linguistics Compass 15 (2021) e12438.

[7] P. Nakov, G. Da San Martino, F. Alam, S. Shaar, H. Mubarak, N. Babulkov, Overview of the clef-2022 checkthat! lab task 2 on detecting previously fact-checked claims (2022).

[8] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeno, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, et al., The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: ECIR (2), 2021.

[9] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeno, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, et al., Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media, in: CLEF (Working Notes), 2020.

[10] A. Kazemi, K. Garimella, D. Gaffney, S. Hale, Claim matching beyond English to scale global fact-checking, in: Proc. 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, Online, 2021, pp. 4504–4517. URL: https://aclanthology.org/2021.acl-long.347. doi:10.18653/v1/2021.acl-long.347.
[11] I. Singh, K. Bontcheva, X. Song, C. Scarton, Comparative analysis of engagement, themes, and causality of ukraine-related debunks and disinformation, in: International Conference on Social Informatics, Springer, 2022, pp. 128–143.
[12] I. Singh, K. Bontcheva, C. Scarton, The false covid-19 narratives that keep being debunked: A spatiotemporal analysis, arXiv preprint arXiv:2107.12303 (2021). URL: https://arxiv.org/abs/2107.12303.
[13] J. Reis, P. d. F. Melo, K. Garimella, F. Benevenuto, Can whatsapp benefit from debunked fact-checked stories to reduce misinformation?, arXiv preprint arXiv:2006.02471 (2020). URL: https://arxiv.org/abs/2006.02471.
[14] I. Singh, C. Scarton, K. Bontcheva, Utdrm: unsupervised method for training debunked-narrative retrieval models, EPJ Data Science 12 (2023) 59.
[15] V. La Gatta, C. Wei, L. Luceri, F. Pierri, E. Ferrara, Retrieving false claims on twitter during the russia-ukraine conflict, arXiv preprint arXiv:2303.10121 (2023).
[16] S. Shaar, F. Alam, G. D. S. Martino, P. Nakov, The role of context in detecting previously fact-checked claims, arXiv preprint arXiv:2104.07423 (2021). URL: https://arxiv.org/abs/2104.07423.
[17] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3607–3618. URL: https://aclanthology.org/2020.acl-main.332. doi:10.18653/v1/2020.acl-main.332.
[18] S. Shaar, F. Alam, G. D. S. Martino, P. Nakov, Assisting the human fact-checkers: detecting all previously fact-checked claims in a document, arXiv preprint arXiv:2109.07410 (2021).
[19] W. Mansour, T. Elsayed, A. Al-Ali, This is not new! spotting previously-verified claims over twitter, Information Processing & Management 60 (2023) 103414.
[20] W. Mansour, T. Elsayed, A. Al-Ali, Did i see it before? detecting previously-checked claims over twitter, in: European Conference on Information Retrieval, Springer, 2022, pp. 367–381.
[21] Q. Sheng, J. Cao, X. Zhang, X. Li, L. Zhong, 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), 2021, pp. 5468–5481.
[22] T. Chakraborty, V. La Gatta, V. Moscato, G. Sperlì, Information retrieval algorithms and neural ranking models to detect previously fact-checked information, Neurocomputing 557 (2023) 126680.
[23] Snopes, Snopes.com — snopes.com, https://www.snopes.com/, 2024. [Accessed 03-02-2024].
[24] PolitiFact, PolitiFact — politifact.com, https://www.politifact.com/, 2024. [Accessed 03-02-2024].
[25] N. Vo, K. Lee, 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), Association for Computational Linguistics, Online, 2020, pp. 7717–7731. URL: https://aclanthology.org/2020.emnlp-main.621. doi:10.18653/v1/2020.emnlp-main.621.
[26] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, et al., The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: European Conference on Information Retrieval, Springer, 2023, pp. 506–517.
[27] Z. S. Ali, W. Mansour, T. Elsayed, A. Al-Ali, Arafacts: the first large arabic dataset of naturally occurring claims, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021, pp. 231–236.
[28] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, K. Todorov, Claimskg: A knowledge graph of fact-checked claims, in: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18, Springer, 2019, pp. 309–324.
[29] A. Kazemi, Z. Li, V. Pérez-Rosas, S. A. Hale, R. Mihalcea, Matching tweets with applicable fact-checks across languages, arXiv preprint arXiv:2202.07094 (2022).

[30] I. F.-C. N. (IFCN), International fact-checking network — poynter.org/, https://www.poynter.org/ifcn/, 2024. [Accessed 03-02-2024].
[31] R. Nogueira, K. Cho, Passage re-ranking with bert, arXiv preprint arXiv:1901.04085 (2019). URL: https://arxiv.org/abs/1901.04085.
[32] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proc. of the 58th ACL, Assoc. for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.
[33] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
[34] Y. Li, B. Jiang, K. Shu, H. Liu, Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation, arXiv preprint arXiv:2011.04088 (2020). URL: https://arxiv.org/abs/2011.04088.
[35] Boomlive, Boomlive — boomlive.in, https://www.boomlive.in/, 2024. [Accessed 03-02-2024].
[36] A. F.-P. (AFP), Agence france-presse (AFP) — afp.com/, https://www.afp.com/, 2024. [Accessed 03-02-2024].
[37] A. EFE, Agencia EFE — efe.com, https://www.efe.com/, 2024. [Accessed 03-02-2024].
[38] T. API, Twitter api — twitter.com/, https://developer.twitter.com/en/docs/twitter-api, 2024. [Accessed 03-02-2024].
[39] Q. Sheng, J. Cao, H. R. Bernard, K. Shu, J. Li, H. Liu, Characterizing multi-domain false news and underlying user effects on chinese weibo, Information Processing & Management 59 (2022) 102959.
[40] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, Trec-covid: constructing a pandemic information retrieval test collection, in: ACM SIGIR Forum, volume 54, ACM New York, NY, USA, 2021, pp. 1–12.
[41] K. Bontcheva, H. Cunningham, I. Roberts, A. Roberts, V. Tablan, N. Aswani, G. Gorrell, Gate teamware: a web-based, collaborative text annotation framework, Language Resources and Evaluation 47 (2013) 1007–1029.
[42] Y. Mu, M. Jin, C. Grimshaw, C. Scarton, K. Bontcheva, X. Song, Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 1052–1062.
[43] X. Hu, Z. Guo, J. Chen, L. Wen, P. S. Yu, Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2901–2912.
[44] G. Bonisoli, M. P. Di Buono, L. Po, F. Rollo, Dice: a dataset of italian crime event news, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2985–2995.
[45] M. Hardalov, A. Chernyavskiy, I. Koychev, D. Ilvovsky, P. Nakov, Crowdchecked: 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, 2022, pp. 266–285.
[46] C. Gormley, Z. Tong, Elasticsearch: the definitive guide: a distributed real-time search and analytics engine, " O'Reilly Media, Inc.", 2015.
[47] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al., Beyond english-centric multilingual machine translation, Journal of Machine Learning Research 22 (2021) 1–48.
[48] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proc. of EMNLP'2020, ACL, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.
[49] E. Yang, S. Nair, R. Chandradevan, R. Iglesias-Flores, D. W. Oard, C3: Continued pretraining with

contrastive weak supervision for cross language ad-hoc retrieval, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2507–2512.
[50] Huggingface, Huggingface — huggingface.co/, https://huggingface.co/eugene-yang/dpr-xlm-align-engtrained, 2024. [Accessed 03-02-2024].
[51] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, in: T. R. Besold, A. Bordes, A. S. d'Avila Garcez, G. Wayne (Eds.), Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016.
[52] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, Transactions on Machine Learning Research (2022).
[53] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. of NAACL'2019, ACL, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[54] Huggingface, Huggingface — huggingface.co/, https://huggingface.co/facebook/mcontriever-msmarco, 2024. [Accessed 03-02-2024].
[55] N. Reimers, I. Gurevych, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.
[56] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020). URL: https://arxiv.org/abs/2007.01852.
[57] Huggingface, Huggingface — use variant which supports around 50 languages., https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2, 2024. [Accessed 03-02-2024].
[58] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, R. Kurzweil, Multilingual universal sentence encoder for semantic retrieval, in: Proc. of the 58th ACL: System Demonstrations, ACL, Online, 2020, pp. 87–94. URL: https://aclanthology.org/2020.acl-demos.12. doi:10.18653/v1/2020.acl-demos.12.
[59] Huggingface, Huggingface — huggingface.co/, https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2, 2024. [Accessed 03-02-2024].
[60] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Advances in Neural Information Processing Systems 33 (2020) 16857–16867.
[61] I. Singh, C. Scarton, K. Bontcheva, Multistage bicross encoder for multilingual access to covid-19 health information, PloS one 16 (2021) e0256874.
[62] O. AI, Open ai — openai.com/, https://platform.openai.com/docs/models, 2024. [Accessed 03-02-2024].
[63] X. Ma, X. Zhang, R. Pradeep, J. Lin, Zero-shot listwise document reranking with a large language model, arXiv preprint arXiv:2305.02156 (2023).
[64] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, V. Stoyanov, Xnli: Evaluating cross-lingual sentence representations, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2475–2485.
[65] S. Ioffe, Improved consistent sampling, weighted minhash and l1 sketching, in: 2010 IEEE international conference on data mining, IEEE, 2010, pp. 246–255.
[66] A. F.-P. (AFP), Agence france-presse (AFP) — afp.com/, https://factuel.afp.com/non-bill-gates-na-pas-propose-dimplanter-une-puce-electronique-la-population, 2024. [Accessed 03-02-2024].
[67] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[68] Wikipedia, Wikipedia — wikipedia.org/, https://en.wikipedia.org/wiki/2020%E2%80%932021_Indian_farmers%27_protest, 2024. [Accessed 03-02-2024].
[69] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, S. Sharif, An analysis of covid-19 vaccine sentiments and opinions on twitter, International Journal of Infectious Diseases 108 (2021) 256–262.

A. Appendix

A.1. Temporal Diversity

To assess dataset diversity, we analyse the temporal characteristics of tweets in Figure 6, presenting a month-by-month breakdown of tweet counts for each language in MMTweets. We observe that Hindi and English tweets exhibit a relatively even distribution from Jan 2020 to Mar 2021. Conversely, Portuguese and Spanish tweets show a more concentrated presence, primarily emerging in late 2020 and early 2021. It's important to note that the MMTweets dataset encompasses at least one tweet for each month from Jan 2020 to Mar 2021 (spanning 15 months). Overall, we find the tweets to be temporally diverse across languages.

In examining cases where debunking precedes misinformation tweets (22.3% of cases), Figure 7 illustrates publication date gaps. With a median gap of 76 days, the findings reveal misinformation can persist even after relevant debunks are available. For instance, one of the false tweets about "Bill Gates launching implantable chips to track COVID-19," appeared in English on Twitter on 3 July 2020, while the earliest related debunk available was published on 13 May 2020 [66] in the French language (49 days gap). This emphasises the need for effective methods, such as X-DNR, to detect the spread of already debunked narratives in multiple languages.

A.2. Domain Diversity

Table 8 presents the results of topic modelling using Latent Dirichlet Allocation (LDA) [67], showcasing the top three topics for each tweet language in MMTweets. As expected, the topics related to coronavirus are apparent in all four languages. However, some topics are specific to events in the country where the language is spoken. In Hindi, the first topic appears to focus on a combination of religious and political elements. For instance, words such as "farmer" and "Delhi" are related to the misinformation that spread during the farmers' protest in Delhi, India [68]. Similarly, in Portuguese, the dominant topics revolve around President Bolsonaro and vaccines. The topics related to "vaccine" are dominant in both Portuguese and Spanish tweets which is likely because the tweets for these languages are mainly from the end of 2020 (see Figure 6), when vaccine-related information was at its peak [69]. English topics cover diverse aspects, including misinformation related to the origin of COVID-19 and its impact on people and hospitals. Overall, the table provides insights into the diverse and multifaceted nature of claims related to COVID-19 in MMTweets.

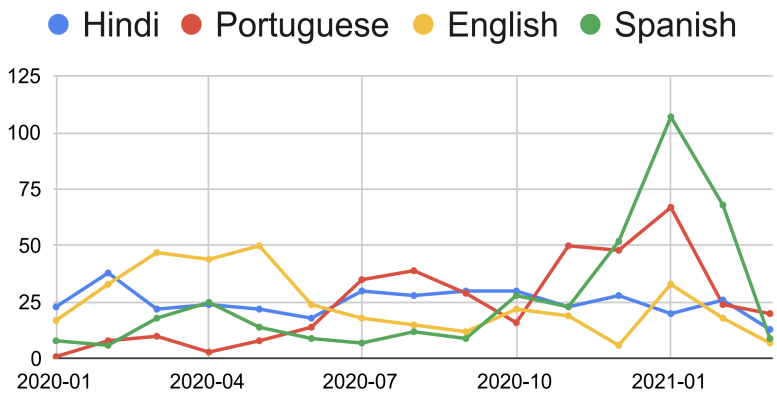


Figure 6: Line plot for month-by-month breakdown of tweet counts for each language in the MMTweets dataset.

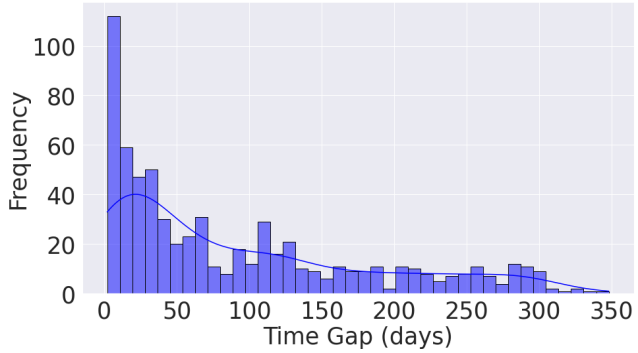


Figure 7: Time gap between tweet and debunk.

Table 8
Topics captured by Latent Dirichlet Allocation.

Language	Latent Dirichlet Allocation Topics
Hindi	Topic 1: hindu, delhi, corona, farmer, government
Hindi	Topic 2: going, people, world, temple, muslim
Hindi	Topic 3: massive, please, wipe, foreign, affair
Portuguese	Topic 1: people, bolsonaro, vaccine, world, covid
Portuguese	Topic 2: vaccine, work, mask, vote, without
Portuguese	Topic 3: vaccine, world, minister, took, abortion
English	Topic 1: deployed, mask, time, corona, epidemic
English	Topic 2: coronavirus, wuhan, like, china, year
English	Topic 3: people, work, coronavirus, hospital, covid
Spanish	Topic 1 : people, without, vaccine, go, mask
Spanish	Topic 2: day, say, first, government, usa
Spanish	Topic 3: vaccine, died, nurse, covid, netherlands