Finding Already Debunked Narratives via Multistage Retrieval: Enabling Cross-Lingual, Cross-Dataset and Zero-Shot Learning

Iknoor Singh, Carolina Scarton, Xingyi Song, Kalina Bontcheva
University of Sheffield
Sheffield, United Kingdom
{i.singh,c.scarton,x.song,k.bontcheva}@sheffield.ac.uk

ABSTRACT

The task of retrieving already debunked narratives aims to detect stories that have already been fact-checked. The successful detection of claims that have already been debunked not only reduces the manual efforts of professional fact-checkers but can also contribute to slowing the spread of misinformation. Mainly due to the lack of readily available data, this is an understudied problem, particularly when considering the cross-lingual task, i.e. the retrieval of fact-checking articles in a language different from the language of the online post being checked. This paper fills this gap by (i) creating a novel dataset to enable research on cross-lingual retrieval of already debunked narratives, using tweets as queries to a database of fact-checking articles; (ii) presenting an extensive experiment to benchmark fine-tuned and off-the-shelf multilingual pre-trained Transformer models for this task; and (iii) proposing a novel multistage framework that divides this cross-lingual debunk retrieval task into refinement and re-ranking stages. Results show that the task of cross-lingual retrieval of already debunked narratives is challenging and off-the-shelf Transformer models fail to outperform a strong lexical-based baseline (BM25). Nevertheless, our multistage retrieval framework is robust, outperforming BM25 in most scenarios and enabling cross-domain and zero-shot learning, without significantly harming the model's performance.

KEYWORDS

Misinformation 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 [8]. A key functionality of these systems is the retrieval of previously fact-checked similar claims, which essentially means the retrieval of already debunked narratives [18, 19, 25]. To achieve this, prior work involves training retrieval models, primarily focusing on monolingual retrieval [13, 18, 19, 25]. In this, the underlying assumption is that these monolingual retrieval models consider previously fact-checked claims to be exclusively present in one language. However, previous studies [22, 29, 30] demonstrate that similar false narratives continue to spread in multiple languages, despite the availability of fact-checks for several months in another language. Hence, automatically finding debunked narratives in multiple languages is crucial to make the best use of scarce fact-checkers resources. For this research, "debunked narratives" are defined as false narratives that spread even after they have already been debunked by at least one professional fact-checker.

In this study, we define the task of **cross-lingual retrieval of already debunked narratives** (CLRADN) as a cross-lingual information retrieval problem. Here, we use a misinformation tweet as a query and retrieve from a corpus of fact-checking articles in multiple languages (Figure 1). The primary goal of CLRADN is to assist fact-checkers in identifying false narratives that continue to spread in one language, even after being debunked by some fact-checker in another language. Our main contributions are:

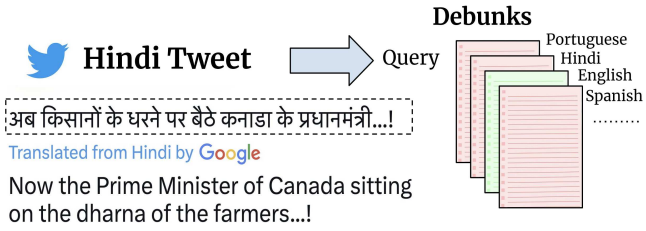

Figure 1: Cross-lingual retrieval of already debunked narratives: tweet is in Hindi and the relevant debunk is in English.

- The **M**ultilingual **M**isinformation **Tweets** (**MMTweets**) dataset: a **novel corpus of annotated tweets** containing **false narratives** in **English, Portuguese, Spanish and Hindi**, together with their corresponding **fact-checking articles** in different languages. In total, it comprises 1,600 query tweets and 30,452 fact-checking articles for retrieval[1] (see Table 1 for examples).
- A **multistage retrieval framework** that effectively tackles the cross-lingual nature of the CLRADN task. An extensive comparative evaluation with lexical, off-the-shelf and fine-tuned state-of-the-art Multilingual Pretrained Transformer(MPT) models shows that our method achieves highest scores on five different languages, demonstrating its effectiveness and broad applicability.
- **Cross-lingual and cross-dataset evaluations** using different datasets for training and testing. With these evaluations, we demonstrate the challenge of building a generalisable CLRADN model that can be used across different languages and domains.

In the following section, we discuss the related work. Section 3 details the MMTweets dataset. Section 4 presents the various experimental methods and the results are presented in Section 5. We conclude the paper in Section 7.

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 [18, 19]. This task is accomplished by

[1]The dataset and code are available at https://doi.org/10.5281/zenodo.7144808.

Table 1: Sample Hindi and Spanish query tweets and their corresponding debunks from the MMTweets dataset.

Fields	Hindi Query Tweet - English Debunk	Spanish Query Tweet - English Debunk
Tweet	अब किसानों के धरने पर बैठे कनाडा के प्रधानमंत्री...! (English translation: Now the Prime Minister of Canada sitting on the farmers' dharna..!)	Covid19 es el arma biológica que se cree q fue creado en el Lab. d Wuhan [...] (English translation: COVID-19 is the biological weapon that is created in the Wuhan Lab [...])
Debunk title	Old Photo Passed Off As Justin Trudeau Sitting In An Anti-Farm Laws Protest	It is impossible to implant a chip in the vaccine against COVID-19 to control the population
Debunk claim	Justin Trudeau sits in protest in support of the protesting farmers.	They will implant a microchip in the coronavirus vaccine to control the population for political and economic purposes.
Debunk article	A photo from 2015 of Canadian Prime Minisiter Justin Trudeau attending a Diwali celebration [...]	Current technologies and international legal and health controls would not allow the introduction of a chip in future [...]

training of debunked-narrative retrieval models, which use misinformation claims as queries to find relevant debunked narratives [16, 17, 24, 28]. Shaar et al. [24] propose the task of detecting previously fact-checked claims and released a dataset of claims and fact-checking articles from Snopes and PolitiFact. Shaar et al. [23] study the role of context in claims made in a political debate, while Vo and Lee [33] investigate the use of multimodal information in tweets to retrieve previously fact-checked content. The CLEF CheckThat! Lab evaluations [2, 18, 19, 25] 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. Nevertheless, all the above mentioned work only focuses on the monolingual scenario.

Kazemi et al. [13] addressed the task of claim matching to identify pairs of similar claims. They conduct retrieval experiments with BM25 and different off-the-shelf Transformer models such as LaBSE [6] and XLM-RoBERTa [3]. They found that multistage retrieval [20] using BM25 and XLM-RoBERTa [3] re-ranking beats the competitive BM25 baseline in some cases. Although they present results for multiple languages, they only experiment with monolingual settings and do not perform cross-lingual claim retrieval. Next, Kazemi et al. [14] work on cross-lingual claim retrieval. They find that the BM25 outperforms or is on par with MPT models on monolingual claim retrieval and LaBSE to be performing best on cross-lingual claim retrieval. However, they do not train custom claim retrieval models or perform cross-lingual zero-shot testing, which we do in this paper (see Section 4.2).

Furthermore, this paper investigates how MPT models can be exploited in a multistage retrieval setting to encode useful semantic information from fact-check pages in a way that effectively tackles the cross-lingual nature of the CLRADN task. For this, we propose a retrieval framework which trains powerful bi-encoder and cross-encoder models for the task of CLRADN (see Section 4.2). Moreover, the majority of previous work [19, 24] builds models that retrieve claims from corpora of fact-checks produced by a single fact-checking organisation. In order to explore method generalisability, we experiment with fact-checking articles published by multiple fact-checking organisations (Section 3). This allows us to build retrieval models that are agnostic to fact-check article structure which is particularly crucial for cross-lingual retrieval.

In summary, our research is novel because it differs from previous work by (i) presenting a new dataset for the task of debunked-narrative retrieval that enables cross-lingual and multilingual research; (ii) proposing a multistage retrieval framework for CLRADN that shows competitive results in domain adaptation and zero-shot learning settings; and (iii) providing the first cross-lingual study to compare changes in retrieval performance using different off-the-shelf and fine-tuned MPT models.

3 MMTWEETS DATASET

MMTweets is a new dataset of misinformation tweets annotated with their corresponding fact-checked articles, both available in multiple languages. MMTweets primarily comprises COVID-19-related misinformation tweets in English, Hindi, Portuguese and Spanish. The languages were selected based on two criteria: 1) these are the most frequent languages in previous publicly available COVID-19 misinformation datasets [15, 29]; 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 total of 30, 452 fact-checking articles published by different fact-checking organisations covering our target languages, namely Boomlive[2] (English), Agence France-Presse (AFP)[3] (German, English, Arabic, French, Spanish, Portuguese, Indonesian, Catalan, Polish, Slovak and Czech), Agencia EFE[4] (Spanish) and Politifact[5] (English). For each fact-checking article, we collect the following information fields: article title, the debunked claim statement and the article body full-text. Next, we select a random sample of 1, 600 fact-checking articles with a focus on COVID-19 misinformation published between January 2020 and March 2021, so as to allow for temporal and topical variety as the pandemic unfolded. Due to the global nature of the COVID-19 pandemic, this sampling approach also maximises the chance of including similar narratives spreading in multiple languages.

[2]https://www.boomlive.in/
[3]https://www.afp.com/
[4]https://www.efe.com/
[5]https://www.politifact.com/

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

Finally, following the previous work [14, 24], we collect the misinformation tweets that were debunked in the sample of fact-checking articles. We use Twitter API [6] to extract the tweet text from URLs. We chose Twitter because of its easy open access as compared to other social media platforms. In this, we aim to maximise cases where the language of the tweet differs from that of the relevant fact-checking article. For instance, Boomlive publishes fact-checking articles in English, but the associated tweets may be in Hindi. This ensures the cross-lingual coverage of the MMTweets dataset.

3.2 Data Annotation

The approach described in Section 3.1 does not guarantee that retrieved tweets 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 retrieved tweets were classified manually to create gold-standard data for evaluation. In particular, we recruited 12 student volunteers who were native speakers of either English, Hindi, Portuguese or Spanish (three native speakers per language).[7] The annotators were shown all information fields from the fact-checking articles and asked to annotate the tweets as belonging to one of three classes:

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

We also implemented 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 author was unsure about the claim's veracity. All such cases were considered "other" due to the chosen narrower definition of misinformation tweets.

[6] https://developer.twitter.com/en/docs/twitter-api

[7] The dataset annotation received ethical approval by the Sheffield Ethics Board.

Table 3: MMTweets language diversity.

Query Tweet Language	Fact-check Language	Count
Portuguese	Portuguese	287
Portuguese	Spanish	18
Portuguese	Indonesian	4
Portuguese	English	1
Hindi	English	326
Hindi	Portuguese	1
Hindi	French	1
Spanish	Spanish	278
Spanish	Catalan	9
Spanish	English	2
Spanish	Indonesian	2
English	English	126
English	Spanish	62
English	Indonesian	25
English	Polish	9
English	Portuguese	9
English	Slovak	5
English	French	4
English	Catalan	4
English	Czech	3

3.3 Data Statistics

A total of 1,600 tweets were annotated, resulting in approximately 400 tweets per language (see Table 2). Following previous methodology [14, 27], we randomly selected 100 tweets from each language to be annotated by three annotators so as to compute inter-annotator agreement (IAA). The Fleiss Kappa scores 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 tweets in the dataset.

Table 4: Volume of fact-checking articles by month in the MMTweets dataset; darker colour means higher volume.

Year	2020												2021		
Month	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3
Hindi	20	27	18	22	24	11	21	29	24	31	23	24	18	19	17
Portuguese	0	3	1	4	0	3	24	30	22	33	32	42	48	42	26
English	13	13	25	34	28	16	22	5	7	17	15	5	19	17	11
Spanish	14	6	2	17	9	10	7	1	8	24	10	34	82	52	15

Table 5: English translations of the top most frequent words in misinformation tweets in our dataset.

Language	Frequent words
Hindi	delhi, corona, government, farmers, ram temple
Portuguese	vaccine, covid, trump, bolsonaro, virus, minister
English	coronavirus, people, china, wuhan, virus
Spanish	vaccine, covid, government, nurse, spain

Table 3 shows the count of query tweet and fact-check pairs in different languages[8] for textual misinformation cases. In particular, in 40% of instances, the language of the tweets and their corresponding fact-checks is different, which makes our dataset the one with the highest proportion of cross-lingual instances (Table 6). The majority of these cross-lingual pairs have tweets in Hindi and corresponding fact-checks in English, followed by instances with misinformation tweets in English and fact-checks in Spanish.

3.4 Dataset Diversity

In order to check the dataset diversity, we analyse the temporal characteristics of the fact-checking articles. Table 4 shows that fact-checking articles for Hindi and English are more uniformly spread between Jan 2020 and Mar 2021, whereas Portuguese and Spanish articles are concentrated in the second half of 2020. It is important to note that the MMTweets dataset covers at least one fact-checking article for each month starting from Jan 2020.

Table 5 depicts the top five most frequent words in misinformation tweets for each language. These words represent events in the country where the language is spoken. The words related to coronavirus are apparent in all four languages. Some distinct words are specific to a given language. For instance, in Hindi tweets, words such as "farmers" and "Delhi" are related to the misinformation that spread during the farmers' protest in Delhi, India[9]. The word "vaccine" is dominant in both Portuguese and Spanish misinformation tweets which is likely because the tweets for these languages are mainly from the end of 2020 (Table 4), when vaccine-related information was at its peak [35]. Appendix A.1 shows a more detailed topic analysis per month for each language. Overall, we find that the misinformation tweets in the MMTweets dataset are topically diverse for each language (Table 9).

[8]We use *langdetect* (https://pypi.org/project/langdetect/) for detecting the language.
[9]https://en.wikipedia.org/wiki/2020%E2%80%932021_Indian_farmers%27_protest

Table 6: Count and language of query claims in MMTweets and the existing debunked-narrative retrieval datasets; where Mono - monolingual, Cross - cross-lingual pairs.

Dataset	Items	Language
Snopes [24]	768	Mono (EN)
Politifact [24]	1,000	Mono (EN)
CLEF 20 - Snopes	1,197	Mono (EN)
CLEF 21 2A - Snopes	1,401	Mono (EN)
CLEF 21 2A-AraFacts	858	Mono (AR)
CLEF 21 2B-PolitiFact	669	Mono (EN)
CLEF 22 2A-Snopes	1,610	Mono (EN)
CLEF 22 2A-AraFacts	908	Mono (AR)
CLEF 22 2B-PolitiFact	752	Mono (EN)
Snopes [33]	11,202	Mono (EN)
Politifact [33]	2,037	Mono (EN)
CrowdChecked [9]	330,000	Mono (EN)
Kazemi et al. [13]	382	Mono (EN, HI, BN, ML, TA)
Kazemi et al. [14]	6,533	Cross (10% - HI-EN lang pairs)
MMTweets	1,600	Cross (40% - multiple lang pairs)

3.5 Comparison to Existing Datasets

Table 6 shows the count and the language of claim queries in previously published datasets for debunked-narrative retrieval task. Generally, the number of claims in our dataset compares favourably to other existing datasets, with most of them being smaller in size (see Table 6). Moreover, all tweets in our dataset are manually classified as misinformation, unlike other existing datasets [14, 26] where tweets are automatically extracted from the fact-check articles and are considered to be misinformation even though our dataset clearly shows that's not always the case (see Table 2). Finally, among the existing datasets, Kazemi et al. [14] is the only one that offers cross-lingual instances, but it comprises only 10% of Hindi-English pairs (where the claim is in Hindi and the fact-checking article is in English), while our MMTweets dataset contains 40% cross-lingual instances across multiple language pairs (see Table 3). Moreover, it is not possible to run replication or comparative experiments on the Kazemi et al. [13, 14] datasets because the papers do not release the corpora of fact-checked articles used in the retrieval experiments - only the claims are released.

4 CROSS-LINGUAL RETRIEVAL OF ALREADY DEBUNKED NARRATIVES

CLRADN is formulated as an information retrieval task where tweets are used as queries to search for relevant fact-checking articles. To accomplish this, we conduct experiments utilising three

distinct fields of fact-checking articles, which we refer to as *debunk fields*. These fields include debunk title (`debunkTitle`), debunk claim (`debunkClaim`), and debunk article full-text (`debunkArticle`), as outlined in Section 3.1. The primary objective is to retrieve the most appropriate fact-checking article based on a query tweet that contains misinformation, ultimately aiming to provide the most relevant and accurate fact-checking information to the user.

4.1 Data Pre-processing

In this paper, we only consider tweets classified as textual misinformation (1,176 in total – Table 2) because we aim to find relevant debunks for tweets that spread misinformation. We divide this dataset into training and test sets. The test set is composed of 100 tweet queries per language, i.e. the same triple annotated tweets used for calculating IAA (Section 3.3). The remaining tweet queries in each language are used as training data. It is important to note that during test time, we do not know if a tweet has been debunked, because tweets linked with fact-checking articles in the test set do not occur in the training set.

In our evaluation experiments, for retrieval, we use previously collected 30,452 fact-checking articles in multiple languages (see Section 3.1). We remove the occurrences of misinformation tweets in the MMTweets dataset that appear on the fact-checking article body to prevent lexical overlap. We also remove Hindi characters (if any) from the English fact-checking articles linked with Hindi misinformation tweets in our dataset.

4.2 Multistage Retrieval Framework

In this study, we introduce a multistage retrieval framework for CLRADN, inspired by the success of similar approaches in other information retrieval (IR) tasks [20, 31, 32]. Our proposed framework divides the CLRADN task into two retrieval stages: a first refinement stage and a second re-ranking stage.

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 [13, 24], since an MPT model is more suitable for the cross-lingual nature of the task. In the second re-ranking stage, we fine-tune an MPT model as a cross-encoder to re-rank the top candidate debunks. Although the utilisation of MPT models as a bi-encoder or cross-encoder is not new, our contribution lies in the novel way we train and provide input to these models that is specifically tailored for the CLRADN task. The details are:

Refinement stage. MPT models are fine-tuned as bi-encoders on misinformation tweet and debunk pairs[10] using in-batch multiple negatives loss [10, 21]. We use in-batch negative training because it is a memory-efficient approach to utilise the negative instances already in the batch rather than creating new ones [10, 12]. Consider a dataset of misinformation tweets $t = (t_1, ..., t_N)$ and their corresponding debunks $d = (d_1, ..., d_N)$. During training, each batch of size K contains one tweet t_i, one relevant debunk d_i, and $K-1$ irrelevant (negative) debunks. Every debunk d_j is essentially treated as a negative candidate debunk for tweet t_i if $i \neq j$. The MPT model is trained to minimise the negative log-likelihood of the data using softmax normalised scores. This updates the model parameters such that the embedding of a misinformation tweet lie in proximity of its relevant debunks as compared to the other irrelevant debunks in the high dimensional vector space. The loss for a single batch of size K is defined as

$$Loss(\theta) = -\frac{1}{K}\sum_{i=1}^{K}\log\frac{exp(DS(f_\theta(t_i), f_\theta(d_i)))}{\sum_{j=1}^{K} exp(DS(f_\theta(t_i), f_\theta(d_j)))} \quad (1)$$

where f_θ is the sentence encoder using the MPT model and DS is the cosine similarity score between misinformation tweet t_i and debunk d_i. We refer to this score as "debunk score" (DS) hereafter. It is defined as

$$DS(u_{tweet}, v_{debunk}) = \frac{u_{tweet} \cdot v_{debunk}}{\|u_{tweet}\|_2 \|v_{debunk}\|_2} \quad (2)$$

where u_{tweet} is the embedding of a tweet and v_{debunk} is the embedding of a debunk, both obtained via mean-pooling (i.e., by averaging embeddings of the constituent subwords of the input text). We employ cosine similarity with mean-pooling technique due to its proven effectiveness in prior research [21]. The L2 norm of u_{tweet} and v_{debunk} are denoted as $\|u_{tweet}\|2$ and $\|v_{debunk}\|_2$, respectively. The fine-tuned MPT model in the first stage is used to rank all debunks based on the DS between the tweet and the debunk. Please refer to Appendix A.2 for hyperparameter details.

Re-ranking stage. MPT models are fine-tuned as cross-encoders and applied to re-rank the top-K retrieved debunks from the first stage. Here, the output of the MPT model is a single relevance score between 0 and 1 representing how relevant a debunk is to a misinformation tweet. The input to the model follows the structure: $[CLS]$ $[T_1]$....$[T_n]$ $[SEP]$ $[DebunkClaim][DC_1]$.....$[DC_i][Debunk\ Title]$ $[DT_1]$.....$[DT_j][DebunkArticle]$ $[DA_1]$.....$[DA_k]$, where T_n represents the tweet subword tokens and DC_i, DT_j and DA_k are the debunk claim, debunk title and debunk article subword tokens, respectively. The explicitly added *DebunkClaim*, *DebunkTitle* and *DebunkArticle* tokens in between the input are placed to guide the model to learn to utilise the different types of available information. $[CLS]$ and $[SEP]$ are the default tokens to indicate "start of input" and "separator", respectively, used in the Next Sentence Prediction task [4]. Although cross-encoder is compute intensive [32], if the number of debunks to be re-ranked is small, it can perform self-attention over the given misinformation tweet and all the *debunk fields* in order to get the final relevance score. Please refer to Appendix A.2 for hyperparameter details.

4.3 Evaluation Setting

MMTweets. The multistage retrieval framework is trained on three different settings and then tested on MMTweets test set:

- `MMTweets-default`: Training on the complete training set of the MMTweets dataset and then evaluation on the respective test sets for each language.
- `Cross-dataset`: In order to test the domain adaptation capabilities of the multistage retrieval framework, we train on the previously published dataset [24], which contains English-only fact-checking articles from Politifact and Snopes (referred to as "SnopesPolitifact" hereafter) and then evaluate the model on the test set of MMTweets. In addition, we only use debunks that are published before January 2020 in SnopesPolitifact in order to

[10] For debunks, we use concatenated `debunkClaim`, `debunkTitle` and `debunkArticle` fields (Section 4).

achieve as little overlap as possible with MMTweets. This ensures a real-life test methodology which checks the generalisability of the model to different datasets.
- `Zero-shot`: We also test the zero-shot capabilities of our multi-stage approach by experimenting with cross-lingual zero-shot learning. In this case, the model is trained on three languages and tested on the unseen fourth language in MMTweets. For instance, in order to test zero-shot for Hindi, the MPT models are trained only on English, Spanish and Portuguese data and then applied to Hindi data only. Hence, in total four models are trained for four different languages in the MMTweets dataset.

CLEF Arabic Datasets. Apart from evaluating on the MMTweets dataset, we also test the performance of our multistage retrieval framework on two additional monolingual Arabic datasets: the CLEF-21 Subtask 2A Arabic dataset [19] and the CLEF-22 Subtask 2A Arabic dataset [18] (see Table 6). We only use the title and claim field since the article body is not available [19]. By testing our framework on these different datasets, we aim to demonstrate its generalisability and robustness in different scenarios.

4.4 Baseline models

Okapi BM25. We use the ElasticSearch[11] [7] implementation of BM25 [11], with default parameters in ElasticSearch ($k = 1.2$ and $b = 0.75$). BM25 is a lexical-based retrieval method and only works for monolingual retrieval. Therefore we employ machine translation as a way of applying BM25 to cross-lingual query - document pairs. To this end, we translated all non-English tweets and *debunk fields* to English using the Fairseq's `m2m100_418M` model [5]. Then, we index the complete corpus of fact-checking articles in elasticsearch [7] and use the translated tweets as queries over the different *debunk fields* (title, claim or article full-text).

MPT Models. We use off-the-shelf state-of-the-art MPT models, namely mulitlingual BERT (mBERT) [4], XLM-RoBERTa [3] and Language-Agnostic BERT Sentence Embedding (LaBSE) [6]. Additionally, we also test the pre-trained Universal Sentence Encoder (USE) [34] variant which supports around 50 languages[12]. We test these models in their default configuration without any supervision from the training dataset to assess their zero-shot performance. In particular, MPT models are applied to extract contextualised embeddings of a tweet and a debunk. Finally, all the debunks in the corpus are ranked based on the debunk score (Equation 2) between tweet and debunk embeddings.

Fine-tuned MPT Models. To ensure a fair comparison with our multistage framework, we also fine-tune the above mentioned MPT models (Section 4.4) using the first stage of our retrieval framework. This involves training the MPT models as bi-encoders on pairs of tweets and debunks with in-batch negatives loss (Section 4.2).

4.5 Evaluation Measures

We employ two widely used ranking evaluation metrics [18, 19] for evaluation : Mean Reciprocal Rank (**MRR**) and Mean Average Precision (**MAP@1** & **MAP@5**).

[11] https://www.elastic.co/elasticsearch/
[12] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

5 RESULTS AND DISCUSSION

5.1 Baseline models

Table 7 shows the evaluation results of BM25 and fine-tuned MPT models on: MMTweets (HI, PT, EN & ES), CLEF-21 Subtask-2A Arabic (CLEF 21-AR) and CLEF-22 Subtask-2A Arabic (CLEF 22-AR). While BM25 is evaluated on four different *debunk fields*, including title, claim, article, and *All* (concatenated claim, title, and article field), MPT models are only evaluated on the *All* field due to the practical limitations of training different models on separate fields.

BM25 and MPT models. The BM25 results show that using combined *debunk fields* (referred to as *All* in Table 7) generally leads to better performance as compared to using only the title, claim, or article fields. In particular, retrieving matching debunks based on titles alone produces the lowest results for all datasets, highlighting its insufficiency. The off-the-shelf MPT models exhibit lower retrieval effectiveness as compared to BM25 results on combined *debunk fields* (see Appendix A.3). Given the impressive results of BM25 on combined *debunk fields*, we retain it as the baseline to compare against other experimental methods.

Fine-tuned MPT models. Fine-tuned LaBSE generally performs better than the BM25, indicating its effectiveness (Table 7 – 9th column). However, the performance of fine-tuned LaBSE varies depending on the dataset and evaluation metric. For example, on the MMTweets-HI dataset, LaBSE achieves the highest MAP@5 score of 0.74, while BM25 achieves a score of only 0.61. On the other hand, on the MMTweets-ES dataset, LaBSE achieves the MAP@5 score of 0.72, while BM25 achieves the highest score of 0.75.

The results also show that the BM25 baseline outperforms both fine-tuned mBERT and XLM-RoBERTa. Surprisingly, the fine-tuned LaBSE model performs comparatively better than fine-tuned mBERT, XLM-RoBERTa, and USE. The multi-task training objective, coupled with standard self-supervised translation and masked language modeling, is likely the reason for LaBSE's superior results [6]. While LaBSE outperforms BM25 in average metric scores, BM25's strong competitive results are likely attributed to the lexical overlap between the translated text, providing BM25 with an extra edge.

Summary of baselines. Overall, the results indicate that BM25 being a lexical retrieval model performs relatively well, especially when evaluated on combined *debunk fields*. However, the other models, particularly LaBSE and USE, outperform BM25 on several metrics and datasets. In terms of datasets, the models perform the best on the CLEF datasets, followed by the MMTweets-PT dataset. The evaluation scores on the different languages are reflective of how topics found in each language impact a model's performance. For example, the English tweets (MMTweets-EN) has the lowest performance across all models and evaluation metrics, which suggests that the topics found in these languages are quite challenging for the model and there is room for improvement (see Table 5 for more details regarding topics found in each language).

5.2 Multistage Retrieval

The last column of Table 7 shows the scores of our multistage retrieval framework. Based on the results of baseline models, we only test LaBSE for training our multistage retrieval framework since it

Table 7: Results for BM25, fine-tuned MPT models and our proposed multistage framework. *All* denotes retrieval based on the concatenated claim, title and full-text of the debunk articles. *NA* means that the article field is not available. * denotes results with statistically significant improvement over the BM25 baseline (with p-value < 0.01). The best scores are in bold.

Dataset	Metrics	BM25				mBERT	XLM-RoBERTa	LaBSE	USE	Multistage Retrieval
		Title	Claim	Article	All	All	All	All	All	All
MMTweets-HI	**MAP@1**	0.09	0.27	0.53	0.55	0.28	0.23	0.67	0.48	**0.76***
	MAP@5	0.12	0.29	0.59	0.61	0.34	0.29	0.74	0.56	**0.81***
	MRR	0.13	0.31	0.60	0.62	0.36	0.31	0.75	0.57	**0.82***
MMTweets-PT	**MAP@1**	0.27	0.36	0.63	0.65	0.72	0.67	0.67	0.61	**0.84***
	MAP@5	0.39	0.47	0.74	0.75	0.77	0.72	0.77	0.71	**0.89***
	MRR	0.43	0.48	0.75	0.76	0.78	0.73	0.77	0.72	**0.89**
MMTweets-EN	**MAP@1**	0.13	0.23	0.34	0.39	0.34	0.24	0.40	0.37	**0.44**
	MAP@5	0.25	0.35	0.49	0.52	0.46	0.33	0.55	0.51	**0.57**
	MRR	0.27	0.37	0.51	0.54	0.48	0.35	0.56	0.53	**0.58**
MMTweets-ES	**MAP@1**	0.32	0.34	0.66	0.66	0.62	0.46	0.61	0.48	**0.73**
	MAP@5	0.40	0.42	0.75	0.75	0.71	0.53	0.72	0.59	**0.80**
	MRR	0.40	0.43	0.75	0.76	0.72	0.55	0.72	0.60	**0.81**
CLEF 21-AR	**MAP@1**	0.54	0.75	NA	0.76	0.67	0.51	**0.82**	0.75	0.81*
	MAP@5	0.64	0.88	NA	0.87	0.76	0.59	**0.95**	0.88	0.94*
	MRR	0.68	0.90	NA	0.90	0.81	0.64	**0.96**	0.90	**0.96***
CLEF 22-AR	**MAP@1**	0.74	0.78	NA	0.80	0.74	0.70	0.88	0.90	**0.94***
	MAP@5	0.78	0.83	NA	0.86	0.78	0.75	0.91	0.92	**0.96***
	MRR	0.79	0.84	NA	0.86	0.79	0.76	0.91	0.93	**0.96***
Average	**MAP@1**	0.35	0.46	0.54	0.64	0.56	0.47	0.67	0.60	**0.75**
	MAP@5	0.43	0.54	0.64	0.73	0.64	0.53	0.77	0.70	**0.83**
	MRR	0.45	0.55	0.65	0.74	0.66	0.56	0.78	0.71	**0.84**

outperforms all other MPT models. The number of documents re-ranked in the second stage is set to 200. Refer to Appendix A.4 and A.5 for the analysis of the count of documents re-ranked and MPT model used in the second stage. Statistical significance is calculated using a pairwise t-test against the BM25 performance on combined *debunk fields*. Table 8 reports the results for our multistage retrieval framework trained on `Cross-dataset` and `Zero-shot` settings along with the `MMTweets-default` setting (from Table 7).

The proposed multistage retrieval framework outperforms all other models across all datasets and metrics, achieving an average MAP@1 score of 0.75, an average MAP@5 score of 0.83, and an average MRR score of 0.84 (Table 7 – last column). This suggests that the proposed framework is more effective for CLRADN as compared to other models. In particular, our multistage retrieval framework significantly outperforms the strong BM25 baseline for MMTweets-HI, MMTweets-PT, CLEF 21-AR and CLEF 22-AR (p-value < 0.01). The multistage retrieval framework even outperforms the previous state-of-the-art model for CLEF 21-AR and CLEF 22-AR shared task [18, 19], demonstrating the generalisability of our approach to different datasets and languages. For instance, the previous best performance on CLEF 21-AR shared task is 0.92, 0.79 and 0.91 for MAP@1, MAP@5, and MRR, respectively[19][13].

Default MMTweets (in-domain). The extent of improvement varies across languages within MMTweets. For example, in the MMTweets-HI dataset, the multistage retrieval approach achieves the highest MAP@1, MAP@5, and MRR scores compared to BM25 on all fields, with an increase of 38%, 33%, and 32%, respectively. Conversely, the improvement is relatively low for the MMTweets-ES dataset, with an increase of only 11%, 7%, and 7% for MAP@1, MAP@5, and MRR scores, respectively. We hypothesise that this difference in performance for the different languages is mostly due to differences in the structure of the fact-checking articles in the different languages and their associated meta-data. Furthermore, it is interesting to note that the effectiveness of the multistage retrieval approach is consistent across metrics. Overall, multistage retrieval improves the results by a significant amount, with an average increase of 18% in MAP@1, 14% in MAP@5, and 13% in MRR compared to BM25. In comparison to fine-tuned LaBSE, the multistage retrieval has an average increase of 12% for MAP@1, 7% for MAP@5, and 7% for MRR metric.

Considering only the monolingual pairs in the MMTweets dataset, i.e. when the language of the misinformation tweet and that of the retrieved fact-checking articles are the same, the MRR scores are 0.84 , 0.94 and 0.82 for English, Portuguese and Spanish, respectively. For Hindi, there are no monolingual pairs (see Table 3). The scores for monolingual pairs are, as expected, higher than for cross-lingual pairs due to the simpler nature of mono-lingual retrieval.

Domain adaptation. Next, we evaluate the usefulness of domain adaptation (Table 8). For that, the `Cross-dataset` setting shows competitive performance, mainly when compared to the baselines in Table 7. The results for MMTweets-HI and MMTweets-PT show statistically significant improvements over BM25 (p-value < 0.01). Nevertheless, the competitive results in this setting indicate that the models can generalise across datasets and they are resilient

[13]There are no reported results for CLEF-22 AR as the scores are zero for the submitted systems [18].

Table 8: Results for multistage retrieval framework using a LaBSE model trained on different datasets. * denotes statistically significant improvement over the BM25 baseline (p-value < 0.01). The best average scores are in bold.

MMTweets Languages	Metrics	MMTweets-default	Cross-dataset (SnopesPolitifact)	Zero-shot (cross-lingual)
HI	**MAP@1**	0.76*	0.66	0.67*
	MAP@5	0.81*	0.71*	0.73*
	MRR	0.82*	0.72*	0.73*
PT	**MAP@1**	0.84*	0.80*	0.43*
	MAP@5	0.89*	0.87	0.59
	MRR	0.89	0.87	0.60
EN	**MAP@1**	0.44	0.43	0.40
	MAP@5	0.57	0.53	0.52
	MRR	0.58	0.54	0.54
ES	**MAP@1**	0.73	0.65	0.71
	MAP@5	0.80	0.75	0.78
	MRR	0.81	0.75	0.79
Average	**MAP@1**	**0.69**	0.64	0.55
	MAP@5	**0.77**	0.71	0.66
	MRR	**0.78**	0.72	0.67

to temporal drift. Moreover, all claim statements in the SnopesPolitifact dataset (used for training of the Cross-dataset model) are in English, whereas our test set contains multiple other languages. This makes it even more challenging for the Cross-dataset model to retrieve the best matching debunk. We hypothesise that training MPT models in a siamese architecture (Section 4.2) makes them learn features that can help bring the misinformation tweet and its relevant debunk close to each other in the representation space even though there is a difference in the domains of the training and testing datasets.

Zero-shot learning. The Zero-shot setting also shows impressive results (last column of Table 8). The zero-shot models significantly outperform BM25 on MMTweets-HI and MMTweets-PT (p-value < 0.01). Although multistage retrieval in MMTweets-default and Cross-dataset setting achieved higher average scores than Zero-shot, it is worth noting that zero-shot models are capable of achieving good results in this challenging setting (Section 4.2).

The competitive results show that transferring knowledge between languages is feasible with our multistage framework and perform cross-language debunk retrieval in low-resource languages without a significant loss in performance. The results also highlight the potential of our approach in scenarios where language-specific models are not available or feasible to train.

Discussion. BM25 demonstrates strong performance, even when debunks require machine translation. In comparison, the multistage framework yields significant improvements. Although BM25 is faster than the multistage approach, the need to machine-translate data for BM25 introduces additional costs and time overheads[14]. Conversely, the multistage framework can be optimised by caching the embeddings. Finally, the findings suggest that the multistage framework facilitates effective knowledge transfer between datasets, especially for low-resource languages where it may be difficult to collect a large-scale dataset for training a dedicated model.

[14]For instance, the machine translation model we used in Section 4.4 operates at 22 sentences per second on a V100 GPU.

6 ERROR ANALYSIS AND FUTURE WORK

We manually inspect the misinformation tweets for which the model is unable to rank the most relevant debunk near the top. For this, we check the top five retrieved debunks for 50 randomly selected cases from the test set. We find that the primary cause of such errors is when the tweet is associated with multiple fact-checking articles debunking similar narratives (12 out of 50). For example, the false narrative "*Cristiano Ronaldo planning to turn his hotels in Portugal into hospitals for people with COVID-19*" has been debunked by AFP and Boomlive. In such cases, the model assigns all relevant fact-checking articles with highly similar high scores, even though in our dataset each tweet is linked to a single best matching fact-checking article. Nevertheless, this does not affect the comparative analysis of our results. Moreover, CLRADN differs from a traditional information retrieval task because it is not aimed at finding all relevant debunks, but instead it is sufficient to find a single relevant debunk, since that would be sufficient for the fact-checkers with establishing that the claim in the tweet is false.

Finally, we also find that CLRADN becomes even more challenging when misinformation spans multiple modalities, i.e. both the text and the image or video of the tweets. In such cases, the retrieval models fail to find a relevant debunk on the basis of tweet text alone (1 out of 50). For instance, a decontextualised video of a protest outside the Ohio Statehouse has been shared as a protest outside the White House following the police killing of George Floyd in Minneapolis [1]. In such cases, the retrieval models would also need to make use of information contained in the video, since the text of the tweet alone is not sufficient. This motivates future work on multimodal debunked-narrative retrieval, where models could exploit joint information from the different modalities.

7 CONCLUSION

This paper focused on cross-lingual retrieval of already debunked narratives (CLRADN) for automated fact-checking. One of the key contributions of the paper is the new MMTweets dataset of annotated misinformation tweets along with their matching fact-checking articles. A second contribution is in the extensive comparative evaluation of the performance of the BM25 lexical model, off-the-shelf and fine-tuned state-of-the-art MPT models on the CLRADN task. We empirically validate that without any task-specific fine-tuning, MPT models fail to outperform a BM25 baseline on machine translated text. However, if our multistage retrieval framework is adopted, then this outperforms the BM25 baselines and fine-tuned MPT models for both MMTweets and CLEF Arabic datasets. In order to assess the zero-shot and domain-adaptation capabilities of our proposed approach, we also perform cross-dataset and cross-lingual evaluations and demonstrate the effectiveness of our approach in transferring knowledge between datasets and languages. This demonstrates the wider practical applicability of our research findings and methods on unseen data. Finally, MMTweets is released as a benchmark dataset and we invite fellow researchers to evaluate their retrieval models on the presented dataset.

8 ETHICAL CONSIDERATIONS

This research has received ethical approval from the Sheffield Ethics board (Application ID 040156). This paper only discusses analysis results in aggregate, without providing examples or information about individual users. Furthermore, any technology, including our models, can be vulnerable to misuse by malicious actors to propagate false information. Hence, we are committed to ensuring our research has a positive impact on society and mitigating any unintended consequences.

ACKNOWLEDGMENTS

This research is supported by a UKRI grant EP/W011212/1 and an EU Horizon 2020 grant (agreement no.871042) ("SoBigData++: European Integrated Infrastructure for Social Mining and BigData Analytics" (http://www.sobigdata.eu)).

REFERENCES

[1] AFP. 2020. This video was taken during a protest at Ohio Statehouse | Fact Check. https://factcheck.afp.com/video-was-taken-during-protest-ohio-statehouse.

[2] Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, et al. 2023. The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *European Conference on Information Retrieval*. Springer, 506–517.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proc. of the 58th ACL*. Assoc. for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL'2019*. ACL, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research* 22, 107 (2021), 1–48.

[6] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852* (2020). https://arxiv.org/abs/2007.01852

[7] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine.* " O'Reilly Media, Inc.".

[8] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[9] Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. 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*. 266–285.

[10] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652* (2017). https://arxiv.org/abs/1705.00652

[11] K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management* 36, 6 (2000), 809–840.

[12] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proc. of EMNLP'2020*. ACL, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[13] Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. 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, 4504–4517. https://doi.org/10.18653/v1/2021.acl-long.347

[14] Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, Scott A Hale, and Rada Mihalcea. 2022. Matching Tweets With Applicable Fact-Checks Across Languages. *arXiv preprint arXiv:2202.07094* (2022).

[15] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. *arXiv preprint arXiv:2011.04088* (2020). https://arxiv.org/abs/2011.04088

[16] Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2022. Did I See It Before? Detecting Previously-Checked Claims over Twitter. In *European Conference on Information Retrieval*. Springer, 367–381.

[17] Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2023. This is not new! Spotting previously-verified claims over Twitter. *Information Processing & Management* 60, 4 (2023), 103414.

[18] Preslav Nakov, Giovanni Da San Martino, Firoj Alam, Shaden Shaar, Hamdy Mubarak, and Nikolay Babulkov. 2022. Overview of the CLEF-2022 CheckThat! lab task 2 on detecting previously fact-checked claims. (2022).

[19] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, et al. 2021. The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In *ECIR (2)*.

[20] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019). https://arxiv.org/abs/1901.04085

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[22] Julio Reis, Philipe de Freitas Melo, Kiran Garimella, and Fabrício Benevenuto. 2020. Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? *arXiv preprint arXiv:2006.02471* (2020). https://arxiv.org/abs/2006.02471

[23] Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2021. The Role of Context in Detecting Previously Fact-Checked Claims. *arXiv preprint arXiv:2104.07423* (2021). https://arxiv.org/abs/2104.07423

[24] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3607–3618. https://doi.org/10.18653/v1/2020.acl-main.332

[25] Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeno, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, et al. 2020. Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In *CLEF (Working Notes)*.

[26] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. *Online social networks and media* 22 (2021), 100104.

[27] Qiang Sheng, Juan Cao, H Russell Bernard, Kai Shu, Jintao Li, and Huan Liu. 2022. Characterizing multi-domain false news and underlying user effects on Chinese Weibo. *Information Processing & Management* 59, 4 (2022), 102959.

[28] Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article Reranking by Memory-Enhanced Key Sentence Matching for Detecting Previously Fact-Checked Claims. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5468–5481.

[29] Iknoor Singh, Kalina Bontcheva, and Carolina Scarton. 2021. The False COVID-19 Narratives That Keep Being Debunked: A Spatiotemporal Analysis. *arXiv preprint arXiv:2107.12303* (2021). https://arxiv.org/abs/2107.12303

[30] Iknoor Singh, Kalina Bontcheva, Xingyi Song, and Carolina Scarton. 2022. Comparative Analysis of Engagement, Themes, and Causality of Ukraine-Related Debunks and Disinformation. In *International Conference on Social Informatics*. Springer, 128–143.

[31] Iknoor Singh, Carolina Scarton, and Kalina Bontcheva. 2021. Multistage BiCross encoder for multilingual access to COVID-19 health information. *PloS one* 16, 9 (2021), e0256874.

[32] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. 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)*.

[33] Nguyen Vo and Kyumin Lee. 2020. Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7717–7731. https://doi.org/10.18653/v1/2020.emnlp-main.621

[34] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proc. of the 58th ACL: System Demonstrations*. ACL, Online, 87–94. https://doi.org/10.18653/v1/2020.acl-demos.12

[35] Samira Yousefinaghani, Rozita Dara, Samira Mubareka, Andrew Papadopoulos, and Shayan Sharif. 2021. An analysis of COVID-19 vaccine sentiments and

Table 9: English translations of the top six most frequent words between January 2020 and March 2021, grouped into quarters per language.

Language	Month				
	Jan-Mar 2020	April-Jun 2020	July-Sept 2020	Oct-Dec 2020	Jan-Mar 2021
Hindi	hindu, world, murder, police, delhi, people	death, corona, world, people, virus, religion	foreign, affairs, temple, hindu, lion, country	government, farmers, country, modi, supreme, hindu	city, india, farmers, people, father, movement
Portuguese	bolsonaro, water, mercury, venus, saturn, pyramids	lions, streets, supermarket, carnival, bahia, government	vaccine, world, covid, minister, france, palestine	vaccine, people, trump, years, votes, covid	people, vaccine, world, masks, health, woman
English	coronavirus, wuhan, china, patients, war, ronaldo	japan, coronavirus, epidemic, professor, italy, nobel	covid, mask, mother, kali, people, video, violence	passport, singapore, mask, pharaoh, crocodile, hyderabad	myanmar, pakistan, military, woman, khan, news, coup
Spanish	tarragona, explosion, day, petrochemical, girl, election	government, people, coronavirus, world, order, health	years, vaccine, age, spanish, influenza, deaths	october, vaccine, netherlands, nurse, cnn, video	vaccine, covid, people, spain, minister, trump

opinions on Twitter. *International Journal of Infectious Diseases* 108 (2021), 256–262.

A APPENDIX

A.1 Word Distribution

Table 9 shows the English translations of the top most frequent words per month. For a better understanding, these are grouped into three months each. Although our dataset spans from January 2020 to March 2021, we find that the tweets are topically diverse for each language. For instance, in Hindi tweets, misinformation related to the farmers' protest in Delhi[15] is highly concentrated after October 2020. In addition, misinformation related to the Ayodhya temple land dispute[16] is rampant between July and September 2020. We also find that vaccine misinformation is dominant after July 2020 in both Portuguese and Spanish misinformation tweets. The English misinformation tweets mostly contain COVID-19 related misinformation until the end of 2020, however, it's different from January to March 2021 (Table 9).

A.2 Hyperparameters

This section presents the training details. The first stage model is trained for four epochs with a batch size of 16, a learning rate of $4e-5$ and maximal input sequence length of 512. For the second stage, the cross-encoder model is trained with a batch size of 16 and $4e-5$ learning rate for two epochs. The subword tokens beyond 512 are truncated. For training the second stage model, we randomly sample ten negative debunks for each misinformation tweet. For all the models, we use linear warmup as the learning rate scheduler and AdamW as optimiser. The models are validated using the MMTweets training set and we manually tune the hyperparameters. The bounds for each hyperparameter are as follows: 1) 1 to 5 epoch 2) 1e-5 to 5e-5 learning rate 3) 8 to 64 batch size which is limited to model's GPU requirement. The training time for each epoch in first and second retrieval stage is 10 and 15 minutes respectively.

[15]https://en.wikipedia.org/wiki/2020%E2%80%932021_Indian_farmers%27_protest
[16]https://en.wikipedia.org/wiki/Ayodhya_dispute

Table 10: Results of BM25 and off-the-shelf MPT models.

Dataset	Metrics	BM25	mBERT	XLM-RoBERTa	LaBSE	USE
MMTweets-HI	**MAP@1**	**0.55**	0.01	0.00	0.45	0.27
	MAP@5	**0.61**	0.02	0.00	0.49	0.34
	MRR	**0.62**	0.04	0.01	0.50	0.37
MMTweets-PT	**MAP@1**	**0.65**	0.16	0.00	0.64	0.36
	MAP@5	**0.75**	0.26	0.00	0.72	0.48
	MRR	**0.76**	0.28	0.01	0.73	0.50
MMTweets-EN	**MAP@1**	**0.39**	0.01	0.00	0.21	0.31
	MAP@5	**0.52**	0.02	0.00	0.30	0.45
	MRR	**0.54**	0.02	0.00	0.33	0.47
MMTweets-ES	**MAP@1**	**0.66**	0.12	0.00	0.45	0.46
	MAP@5	**0.75**	0.18	0.00	0.57	0.53
	MRR	**0.76**	0.21	0.01	0.58	0.55
CLEF 21-AR	**MAP@1**	**0.76**	0.20	0.07	0.74	0.72
	MAP@5	**0.87**	0.23	0.08	0.86	0.84
	MRR	**0.90**	0.28	0.11	0.89	0.88
CLEF 22-AR	**MAP@1**	0.80	0.14	0.10	**0.88**	0.76
	MAP@5	0.86	0.17	0.11	**0.90**	0.80
	MRR	0.86	0.19	0.12	**0.90**	0.81
Average	**MAP@1**	**0.64**	0.11	0.03	0.56	0.48
	MAP@5	**0.73**	0.14	0.03	0.64	0.57
	MRR	**0.74**	0.17	0.04	0.66	0.59

All experiments are conducted on a machine with NVIDIA GeForce RTX 3090.

A.3 Results of MPT Models

Table 10 shows the results of the of the off-the-shelf models, namely mBERT, XLM-RoBERTa, LaBSE, and USE, while also incorporating the BM25 outcomes from Table 7 (6th column) for reference. The results indicate that BM25 performs consistently well across all datasets and metrics. Among the Transformer models, LaBSE and USE deliver the most favorable outcomes across most datasets and metrics, with LaBSE exhibiting superior performance on most metrics, especially on the MMTweets-PT and CLEF 21-AR datasets. However, other Transformer models, such as mBERT and XLM-RoBERTa, tend to perform poorly in most cases.

Table 11: Evaluation results of the multistage retrieval framework using various values of K, which represents the number of documents re-ranked in the second stage.

Dataset	Metrics	K=50	K=100	K=200	K=300	K=400
MMTweets-HI	**MAP@1**	0.75	0.75	0.76	0.76	0.76
	MAP@5	0.80	0.80	0.81	0.82	0.82
	MRR	0.80	0.80	0.82	0.82	0.82
MMTweets-PT	**MAP@1**	0.84	0.84	0.84	0.84	0.84
	MAP@5	0.89	0.89	0.89	0.89	0.89
	MRR	0.89	0.89	0.89	0.89	0.89
MMTweets-EN	**MAP@1**	0.45	0.44	0.44	0.44	0.43
	MAP@5	0.59	0.58	0.57	0.57	0.57
	MRR	0.60	0.59	0.58	0.58	0.58
MMTweets-ES	**MAP@1**	0.73	0.72	0.73	0.74	0.74
	MAP@5	0.81	0.80	0.80	0.82	0.82
	MRR	0.81	0.80	0.81	0.82	0.82
Average	**MAP@1**	0.69	0.69	0.69	**0.70**	0.69
	MAP@5	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**
	MRR	**0.78**	0.77	**0.78**	**0.78**	**0.78**

Table 12: Results of different MPT models used in the second stage of the multistage retrieval framework.

Dataset	Metrics	mBERT	XLM-RoBERTa	LaBSE
MMTweets-HI	**MAP@1**	0.49	0.17	**0.76**
	MAP@5	0.59	0.24	**0.81**
	MRR	0.60	0.27	**0.82**
MMTweets-PT	**MAP@1**	**0.87**	0.14	0.84
	MAP@5	**0.91**	0.30	0.89
	MRR	**0.91**	0.34	0.89
MMTweets-EN	**MAP@1**	0.38	0.03	**0.44**
	MAP@5	0.53	0.06	**0.57**
	MRR	0.54	0.11	**0.58**
MMTweets-ES	**MAP@1**	0.68	0.21	**0.73**
	MAP@5	0.78	0.37	**0.80**
	MRR	0.79	0.40	**0.81**
Average	**MAP@1**	0.61	0.14	**0.69**
	MAP@5	0.70	0.24	**0.77**
	MRR	0.71	0.28	**0.78**

In conclusion, LaBSE emerges as the optimal choice after BM25 as it outperforms other Transformer models in most cases. Thus, researchers and practitioners in the field of information retrieval may consider utilising LaBSE as a substitute for BM25 when they aim to achieve better performance without the need for machine translation that is required in BM25.

A.4 Influence of Number of Documents re-ranked

Table 11 shows the evaluation results of the multistage retrieval framework on MMTweets using various values of K, which represents the number of documents re-ranked in the second stage. The results show that increasing the value of K generally improves the model's performance, as indicated by higher MAP and MRR scores. For instance, the MAP@5 score for Hindi increases from 0.80 and 0.82, when K is increased from 50 to 400, indicating a consistent improvement in the model's performance. On the other hand, for Portuguese, all three metrics remain constant.

Overall, the results suggest that increasing the number of documents re-ranked in the second stage can improve the performance of the model, but the magnitude of the improvement may vary depending on the dataset and the evaluation metric used. Furthermore, it's worth noting that increasing K also results in a longer time taken to retrieve relevant documents, which can be a drawback in real-world applications. Therefore, in our experiments, we chose a value of K as 200 to balance the trade-off between performance and efficiency.

A.5 MPT Model in the Second Stage of Multistage Retrieval Framework

Table 12 shows the results of using three different MPT models (mBERT, XLM-RoBERTa, and LaBSE) in the second stage of the multistage retrieval framework. Please refer to Appendix A.2 for hyperparameter details. We find that the average performance across all datasets and metrics is highest for LaBSE, followed by mBERT and then XLM-RoBERTa. In particular, LaBSE outperforms the other two models significantly in the MMTweets-PT dataset, while XLM-RoBERTa performs the worst across all datasets and metrics. Therefore, we choose LaBSE as the model for the second stage of retrieval in our experiments.