Singh et al. *EPJ Data Science* (2023) 12:59
https://doi.org/10.1140/epjds/s13688-023-00437-y

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

UTDRM: unsupervised method for training debunked-narrative retrieval models

Iknoor Singh[1*], Carolina Scarton[1] and Kalina Bontcheva[1]

*Correspondence: i.singh@sheffield.ac.uk
[1] University of Sheffield, Sheffield, UK

Abstract

A key task in the fact-checking workflow is to establish whether the claim under investigation has already been debunked or fact-checked before. This is essentially a retrieval task where a misinformation claim is used as a query to retrieve from a corpus of debunks. Prior debunk retrieval methods have typically been trained on annotated pairs of misinformation claims and debunks. The novelty of this paper is an Unsupervised Method for Training Debunked-Narrative Retrieval Models (UTDRM) in a zero-shot setting, eliminating the need for human-annotated pairs. This approach leverages fact-checking articles for the generation of synthetic claims and employs a neural retrieval model for training. Our experiments show that UTDRM tends to match or exceed the performance of state-of-the-art methods on seven datasets, which demonstrates its effectiveness and broad applicability. The paper also analyses the impact of various factors on UTDRM's performance, such as the quantity of fact-checking articles utilised, the number of synthetically generated claims employed, the proposed *entity inoculation* method, and the usage of large language models for retrieval.

Keywords: Fact-checking; Misinformation detection; Information retrieval

1 Introduction

Automated fact-checking systems are pivotal not only for combatting false information on digital media but also for reducing the workload of fact-checkers [1, 2]. A key functionality of these systems is the retrieval of already debunked narratives for misinformation claims, which essentially means retrieving previously fact-checked similar claims [2–4]. This function is accomplished by training debunked-narrative retrieval models that utilise misinformation claims as queries to retrieve relevant debunked narratives.

Previous methods for training debunked-narrative retrieval models heavily rely on annotated pairs of misinformation claims and debunks [2, 4, 5]. However, the process of manually creating annotated pairs is time-consuming, labour-intensive, and often limited in scale, which can impede the performance of the retrieval models.

In this paper, we propose an *U*nsupervised method for *T*raining *D*ebunked-Narrative *R*etrieval *M*odels (*UTDRM*) that utilises synthetic claims to overcome the limitation of relying on manual annotations (see Fig. 1). Moreover, we hypothesise that UTDRM has the potential to detect topical misinformation by generating claims from incoming topical

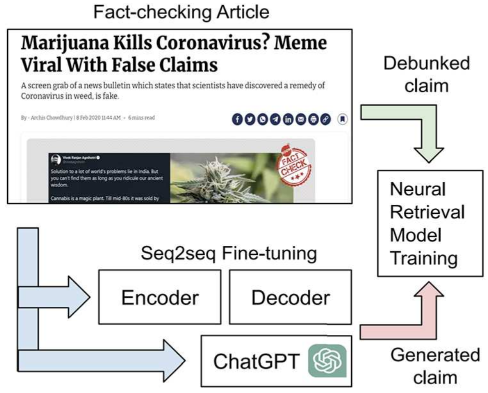
Fact-checking Article
Marijuana Kills Coronavirus? Meme Viral With False Claims
A screen grab of a news bulletin which states that scientists have discovered a remedy of Coronavirus in weed, is fake.
Debunked claim
Neural Retrieval Model Training
Seq2seq Fine-tuning
Encoder
Decoder
ChatGPT
Generated claim

Figure 1 End-to-end pipeline for `UTDRM`: a two-step method involving the generation of topical claims and the training of a neural retrieval model

fact-checks, thereby expanding its overall impact. Furthermore, our proposed *entity inoculation* method (Sect. 6.3) addresses the pressing challenge of similar false narratives evolving with different entities [6]. Our inspiration for this approach stems from an independent analysis, noting similar misinformation claims involving distinct entities. For example, misinformation about crocodile sightings during floods vary across locations – Hyderabad,[1] Patna,[2] Bengaluru,[3] and Florida[4] (see Appendix A.4 for more examples). By replacing named entities in generated claims, *entity inoculation* enhances the robustness of our `UTDRM` method, directly addressing the issue of narrative adaptability (see Sect. 6.3).

In particular, the research question addressed in this study is: how to train efficient debunked-narrative retrieval models without relying on human-annotated data?

The main contributions of this paper are:

- *UTDRM*, a two-step method for training debunked-narrative retrieval models that achieves comparable or superior retrieval scores to supervised models, all without relying on annotations. Figure 1 illustrates the UTDRM's end-to-end pipeline.
- A *large-scale dataset of synthetic topical claims* created using topical claim generation techniques based on text-to-text transformer-based models and large language models (LLMs).
- A *comprehensive performance evaluation* of `UTDRM` on seven publicly available datasets, demonstrating its effectiveness and generalisability in retrieving accurate debunks for misinformation in tweets, political debates, or speeches.
- *Extensive ablation experiments* that assess the impact of different factors on `UTDRM`'s performance. This includes: (1) the volume of fact-checking articles utilised, (2) the number of synthetically generated claims used for training, (3) the proposed *entity inoculation* method, and (4) the usage of LLMs, such as Large Language Model Meta AI (LLaMA 2) and Chat Generative Pre-trained Transformer (ChatGPT), for retrieval.

[1] https://factcheck.afp.com/no-footage-has-circulated-2019-reports-about-crocodile-west-india.

[2] https://www.boomlive.in/crocodile-spotted-during-bihar-floods-video-from-gujarat-shared-as-patna/.

[3] https://www.indiatoday.in/fact-check/story/fact-check-crocodile-spotted-waterlogged-bengaluru-viral-video-mp-1997133-2022-09-06.

[4] https://factcheck.afp.com/doc.afp.com.32KT6D7.

In the following sections, we discuss related work (Sect. 2) and our proposed UTDRM method (Sect. 3). Section 4 presents the various experimental methods and the datasets used for evaluation. The results and ablation experiments are presented in Sect. 5 and Sect. 6 respectively. Finally, we conclude the paper in Sect. 8.

2 Related work

Information retrieval involves the search and retrieval of relevant documents from a collection in response to a query. Initially, conventional lexical methods such as, Okapi Best Match 25 (BM25) [7], Term Frequency-Inverse Document Frequency (TF-IDF) weighting [8], Query Likelihood model (QL) [9], and Divergence From Randomness (DFR) [10], were the primary information retrieval techniques, which demonstrated the effectiveness of lexical and statistical approaches. However, these traditional approaches faced challenges in addressing lexical gaps and semantic issues in relevance matching [11]. In response to these challenges, recent recent Transformer-based methods [12] aim to harness the power of deep learning to enhance performance [13]. In the following sections, we review related work in two main areas: supervised and unsupervised methods for debunked-narrative retrieval.

2.1 Supervised training methods

Many existing methods for training debunked-narrative retrieval models rely on supervised learning techniques which typically leverage annotated pairs of misinformation claims and fact-checking articles as training data [3, 14–18]. For instance, Shaar et al. [16] train a pairwise learning-to-rank model for identifying debunked narratives. They also release Snopes and Politifact datasets [16], which we use for evaluation in this paper (Sect. 4.1). Similarly, Vo and Lee [19] train a ranking model that incorporates both textual and visual features to retrieve previously fact-checked content, while Shaar et al. [20] employ the Transformer-XH [21] to examine the role of context in political debates. On the other hand, Kazemi et al. [5, 22] address the task of debunked-narrative retrieval as a binary classification problem and train support vector machines model to classify misinformation tweets. However, formulating it as a classification problem is computationally not scalable due to its quadratic complexity.

The Conference and Labs of the Evaluation Forum (CLEF) CheckThat! Lab shared task 2020, 2021 and 2022 [2, 3, 14, 23] focus on debunked-narrative retrieval task and release different datasets for training and testing. In this paper, we utilise all of these CLEF test datasets for evaluation (Sect. 4.1). Teams in CLEF 22 use diverse methods, such as Sentence-T5 and GPT-Neo for re-ranking [24], Simple Contrastive Learning of Sentence Embeddings (SimCSE) [25], and data augmentation like back translation [26]. We utilise the state-of-the-art performance demonstrated by the shared task winners as a benchmark for comparing against our UTDRM method (Sect. 4.2).

While supervised training approaches require annotated training data, which can be costly and time-consuming to collect, this research proposes an alternative novel approach. By utilising fact-checking articles from professional fact-checking organisations, our method generates high-quality training data without the need for annotations. This methodology yields high scores in debunked-narrative retrieval (Sect. 5).

2.2 Unsupervised training methods

In recent years, unsupervised training methods for information retrieval have gained significant interest [13, 27–30]. Our proposed UTDRM method falls within this category. These unsupervised methods aim to overcome the challenges associated with acquiring annotated training data by utilising large corpora of unlabeled documents. For example, Lee et al. [27] introduce the Inverse Cloze Task (ICT) for training models using synthetic query-passage pairs by uniformly sampling sentences from random passages. Alternatively, Tranformer-based Denoising AutoEncoder (TSDAE) [28] encodes sentences with randomly deleted 60% of the tokens and the decoder to reconstruct the original sentences. Similarly, methods like SimCSE [25] and Contrastive Tension [31] focus on minimising the distance between embeddings from the same sentence. ICT, TSDAE, and SimCSE are among the unsupervised methods employed for comparison with our proposed UTDRM method (as discussed in Sect. 4.2).

Other line of unsupervised methods explore query generation as an alternative to improve retrieval performance. For eg. Nogueira et al. [32, 33] enhance traditional BM25 search by expanding passages with synthetic queries. On the other hand, Ma et al. [34] propose a zero-shot learning approach for passage retrieval using synthetic question generation, while Wang et al. [29] introduce Generative Pseudo Labeling (GPL), an unsupervised domain adaptation method that combines a T5-based query generator with pseudo labelling from a cross-encoder. However, these methods are not suitable for our specific use case since generating claims from fact-checking articles is a novel task in itself, and therefore, relying on pre-trained query generation models trained for different purposes is not appropriate. Additionally, the use of Margin Mean Squared Error (MarginMSE) [35] in GPL, which relies on a cross-encoder trained on Microsoft Machine Reading Comprehension (MSMARCO) data, may not be effective for our specific debunked-narrative retrieval task. This is because our task differs from general information retrieval tasks that typically require general queries as input, while the task in this paper specifically focuses on false claims on social media and political debates (Sect. 4.1).

While existing unsupervised methods show promising results, there is still room for improvement in retrieval performance and applicability. UTDRM aims to address these challenges by utilising unsupervised learning techniques tailored specifically for training debunked-narrative retrieval models. It focuses on generating high-quality topical misinformation claims from fact-checking articles (Sect. 3.1) which, to the best of our knowledge, has not been explored in previous work. These generated claims are employed to train the retrieval model in a zero-shot setting (Sect. 3.2).

Finally, this study is the first to assess the performance of LLMs (LLaMA 2 and ChatGPT) as listwise re-rankers on seven publicly available debunked-narrative retrieval datasets (Sect. 6.4). This assessment is conducted to examine how LLMs perform in comparison to other unsupervised methods, including our UTDRM.

3 UTDRM: unsupervised method for training debunked-narrative retrieval models

Debunked-narrative retrieval is a key task in a typical fact-checking workflow, where the verification professionals determine whether the claim or content that they need to verify has already been debunked in a publicly available debunking article posted by another fact-checking organisation. This is essentially a retrieval, where a misinformation claim serves as the query to extract relevant debunked claims (or fact-checked claims) from a database of already published publicly available debunking articles. It must be noted that if a claim has not already been debunked in a published article, there may not be suitable matches.

This section presents our proposed UTDRM method, which consists of two steps: (i) generation of topical claims (Sect. 3.1); and (ii) training of a debunked-narrative retrieval model (Sect. 3.2). Figure 1 illustrates the end-to-end pipeline for UTDRM.

3.1 Topical claim generation

We synthetically generate topical claims that resemble misinformation claims based on the debunked information provided by professional fact-checkers. To accomplish this, we propose two novel methods: the use of Text-to-Text Transfer Transformer (T5) and ChatGPT as claim generators. In this work, we specifically investigate the zero-shot scenario, where annotated pairs of social media posts and debunked claim pairs are unavailable, and only a large corpus for fact-checks is available.

3.1.1 T5 claim generator

The T5 claim generator is a sequence-to-sequence model based on the text-to-text transfer transformer (T5) [36]. We choose T5 model because of its proven effectiveness in various sequence-to-sequence tasks in prior research [29, 32, 36]. T5 is used to generate claims from fact-checking articles by framing the task as an encoder-decoder problem. The encoder is trained to understand and represent the fact-checking articles, while the decoder generates potential misinformation claims that can be effectively debunked using the corresponding fact-checking articles.

To train the T5 claim generator, first, we create a corpus of fact-checking articles published by different fact-checking organisations, namely Boomlive,[5] Agence France-Presse (AFP)[6] and Politifact.[7] We choose these fact-checking websites for their wide topic coverage, deferring the comparison of claim generators trained on different websites for future research. A total of 23,901 fact-checking articles were collected. For each fact-checking article, we collect the debunked claim statement, the title and the main body of the article. During fine-tuning, the input to the T5 model consists of the title and the main body of the fact-checking article, and the model is trained to generate the debunked claim statement. Since the generated claims are conditioned on the fact-checking article, they remain closely related to the actual claims being debunked in the fact-checking article. Please refer to Appendix A.1 for hyperparameter details.

3.1.2 ChatGPT claim generator

We use ChatGPT (*gpt-3.5-turbo*)[8] to generate tweets that are relevant to the debunked claims of fact-checking articles collected above. To achieve this, we provide an input prompt instructing the model to generate five different tweets about the text, ensuring that the generated tweets are not fact-checks or debunks. Additionally, we encourage the

[5] https://www.boomlive.in/.

[6] https://www.afp.com/.

[7] https://www.politifact.com/.

[8] https://platform.openai.com/docs/models.

diversity of hashtags in the generated tweets to enhance their variability. For this, we use the input prompt as:

````
Generate ten different tweets about the text delimited by triple
backticks.
Make sure that generated tweets should not be a fact-check or a
debunk.
Also, tweets should have different hashtags. ```{Debunked Claim}```
````

In summary, we use ChatGPT in conjunction with the T5 claim generator due to our observation that ChatGPT generates claims that are more diverse (Table 2) and closely resemble actual tweet claims (Sect. 3.1.3). Additionally, both T5 and ChatGPT claim generator can address emerging topics by generating claims from incoming topical fact-checks. These generated claims serve as valuable inputs for training our neural retrieval model (Sect. 3.2).

3.1.3 Generated claims

Table 1 showcases sample claims generated from T5 and ChatGPT. We present five random instances of debunked claims alongside three generated claims from each model. In the first example, T5 produces three claims pertaining to Senator Kamala Harris potentially violating laws during a visit to an Ohio voting site, while ChatGPT generates alternative claims with similar themes. Similarly, for the other examples, T5 and ChatGPT generate diverse variations of claims related to Dr Kafeel Khan's involvement in a farmers' rally in Delhi and a supposed COVID-19 cure by a Pondicherry University student.

In summary, both T5 and ChatGPT generate different types of claims with variations in wording, focus, and emphasis, while still conveying similar information related to the original debunked claims. Moreover, our analysis reveals that the claims generated by T5 exhibit simplicity and a higher level of similarity to the debunked claims. On the other hand, the claims generated by ChatGPT demonstrate greater diversity and closely resemble actual tweets, often incorporating hashtags (as shown in Table 2 – Sect. 3.1.4). Notably, some of the ChatGPT generated claims ask questions while stating the debunked claim (last example in Table 1). Finally, by using both T5 and ChatGPT, we can capture a broader range of claim styles and ensure comprehensive coverage for training debunked-narrative retrieval models.

3.1.4 Quality and diversity

Table 2 evaluates the generated claims using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [37] and self Bilingual Evaluation Understudy (selfBLEU) [38] metrics. Following previous work [29, 32], our evaluation does not involve human assessment. Instead, we rely on automatic metrics to assess the quality of generated claims. ROUGE measures the proximity of the generated claims to the reference debunked claims, while selfBLEU assesses the diversity among the generated claims. The choice of these metrics is justified by their close alignment with our research objectives, emphasising both quality and diversity as crucial evaluation criteria. We generate a total of six claims (three from each claim generator) from the collected fact-checking articles (Sect. 3.1), as it yields the best scores during experiments (see Sect. 6.2). The results in Table 2 indicate that T5 outperforms ChatGPT in ROUGE scores across all n-gram levels, indicating higher overlap with the reference debunked claims. This performance difference can be attributed to the fine-tuning of T5 in the T5 claim generator. Further evaluation of retrieval models trained

Table 1 Sample claims generated from T5 and ChatGPT claim generator (Sect. 3.1). The debunked claim statements from the fact-checking articles used to generate claims are also mentioned

Debunked Claims	T5 Generated Claims	ChatGPT Generated Claims
Sen. Kamala Harris "may have broken laws" when she visited a voting site in Ohio.	1. Kamala harris "said "this weekend, yells at people in line to vote in ohio, may have broken laws." 2. Sen. Kamala Harris broke ohio election laws. 3. Kamala harris broke ohio election laws by speaking to people in line to vote outside a polling site.	1. Did Sen. Kamala Harris cross the line during her Ohio voting site visit? Investigation underway. #votingrights #legalissues 2. Sen. Kamala Harris is facing accusations of breaking the law during a recent visit to an Ohio voting site. Stay tuned for updates! 3. Did Sen. Kamala Harris break the law during a recent visit to an Ohio voting site? The controversy continues to unfold. #VoterFraud
Dr Kafeel Khan driving a tractor at the farmers tractor rally in Delhi on Republic Day.	1. Dr Kafeel Khan was present at farmers tractor rally in Delhi. 2. Dr Kafeel Khan sitting on a red tractor in Delhi 3. Dr Kafeel Khan Was Present On Tractor Rally By Farmers On Republic Day	1. A picture worth a thousand words – Dr. Kafeel Khan driving a tractor at the farmers rally in Delhi on Republic Day. #Solidarity #India 2. Dr. Kafeel Khan showing his solidarity with farmers as he drives a tractor at the Republic Day rally in Delhi. #FarmersProtest #India 3. A powerful image of Dr. Kafeel Khan joining the farmers tractor rally in Delhi on Republic Day. #Solidarity #FarmersRights
Pondicherry University student found the cure for covid-19.	1. Pondicherry university student found a coronavirus cure by mixing ginger, black pepper powder and honey for five days. 2. "An Indian student from Pondicherry university found a home remedy cure for COVID-19 which is for the very first time accepted by WHO." 3. BJP member of parliament, Manaskar Singh, found "a home remedy cure for COVID-19, which is for the very first time accepted by WHO".	1. #Inspiration This is the kind of news that inspires us all. A student from Pondicherry University has made a groundbreaking discovery in the fight against COVID-19. Let's support and celebrate our young innovators! 2. The world is one step closer to defeating COVID-19 thanks to a brilliant mind from Pondicherry University. Let's celebrate this incredible achievement! #science #innovation 3. #Hope A glimmer of hope in these trying times! A student from Pondicherry University has uncovered a potential cure for COVID-19. Let's keep our fingers crossed for a breakthrough!
Photos show cash and gold recovered in income tax raids at MK Stalins daughters house.	1. Photos show money seised from the Income Tax raids conducted at the house of MK Stalin's daughter Senthamarai 2. Photos shows the money seized at the house of MK Stalins daughter Senthamarai Sabareesan 3. Photos show cash and gold taken in Income Tax raids at the house of MK Stalins daughter Senthamarai	1. The rich get richer? Income tax raids at MK Stalin's daughter's house expose a hoard of cash and gold. Will they be held accountable? #corruption #India 2. Shocking revelation! Income tax raids at MK Stalin's daughter's house reveal cash and gold – photos don't lie! #corruption #IndiaPolitics 3. The truth is out! Income tax raids at MK Stalin's daughter's house reveal a stash of cash and gold. Will justice prevail? #corruption #India
Countries were exporting COVID-19 "diagnostic test instruments" in 2018.	1. The world bank created a list of COVID-19 tests before the coronavirus pandemic. 2. "Coronavirus tests exist in 2018, predates the pandemic." 3. A list of exports of COVID-19 diagnostic tests "predates" the coronavirus.	1. Did you know that countries were already exporting #COVID19 diagnostic test instruments as early as 2018? Interesting discovery. 2. This is mind-boggling: countries were exporting COVID-19 "diagnostic test instruments" in 2018! How did we not see this coming? 3. It's surprising to learn that #COVID19 diagnostic test instruments were already being exported by countries in 2018. What does this tell us?

Table 2 ROUGE and selfBLEU scores for claims generated from T5 and ChatGPT claim generator. Lower selfBLEU scores indicate higher diversity, while higher ROUGE scores indicate greater overlap with the reference debunked claims

Metrics	ROUGE-1	ROUGE-2	ROUGE-L	selfBLEU1	selfBLEU2	selfBLEU3
T5	0.563	0.423	0.541	0.553	0.493	0.444
ChatGPT	0.272	0.119	0.237	0.250	0.142	0.085

on generated claims will provide insights into the claim quality and their alignment with task requirements (Sect. 5).

Table 2 also presents the selfBLEU scores, which computes the similarity between the generated claims, with lower scores indicating higher diversity. T5 exhibits higher selfBLEU scores across all N-gram levels, indicating more similarity among its generated claims. In contrast, ChatGPT achieves lower selfBLEU scores, suggesting greater diversity and distinctiveness in its generated claims.

3.2 Neural retrieval model

The neural retrieval model is a transformer model fine-tuned on the generated claim and the original debunked claim statement pairs using multiple negatives ranking loss (MNRL) [39, 40]. In this, consider a dataset of synthetically generated claims $g = (g_1, \ldots, g_N)$ along with their corresponding debunked claim statements $d = (d_1, \ldots, d_N)$. During fine-tuning, each batch of size K contains one generated claim g_i and one corresponding relevant debunked claim statement d_i, which is the same debunked claim used for generating g_i. The remaining $K - 1$ elements in the batch are irrelevant debunked claim statements which are the hard negatives mined using a pretrained retrieval model. Every debunked claim statement d_j is a negative candidate for generated claim g_i if $i \neq j$. The loss for a single batch of size K is defined as,

$$-\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp(\mathrm{Sim}(f_\theta(g_i), f_\theta(d_i)))}{\sum_{j=1}^{K}\exp(\mathrm{Sim}(f_\theta(g_i), f_\theta(d_j)))}, \tag{1}$$

where f_θ is the sentence encoder using the transformer model and Sim is the similarity between the encoded embeddings. We employ cosine similarity function with the mean-pooling technique due to its proven effectiveness in prior research [41]. MNRL aims to maximise the similarity between the generated claim and its relevant debunked claim statement while minimising the similarity with irrelevant statements. Hyperparameter details are in Appendix A.1.

4 Experimental setup

4.1 Evaluation datasets

We evaluate the models on the test set of seven publicly available datasets. The datasets are divided into two types based on whether the claims are sourced from Twitter or from political debates or speeches:

- *Twitter-based* datasets: *Snopes* [16] and CLEF CheckThat! Lab task datasets which include *CLEF 22 2A* [3], *CLEF 21 2A* [14] and *CLEF 20 2A* [2].
- *Political-based* datasets: *Politifact* [16] and CLEF CheckThat! Lab task datasets which include *CLEF 22 2B* [3] and *CLEF 21 2B* [14].

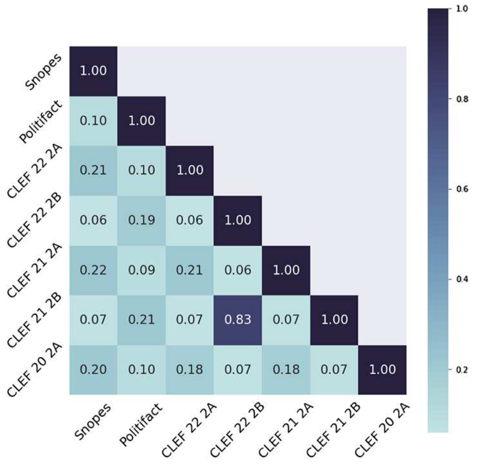

Figure 2 Heatmap for dataset domain overlap

To assess the diversity of domains, we calculate the pairwise domain overlap between all the claims in the datasets using a weighted Jaccard similarity measure [42]. Figure 2 shows a heatmap illustrating the pairwise weighted Jaccard similarity scores. Besides CLEF 22 2B and CLEF 21 2B, the results indicate a relatively low overlap among most datasets, suggesting that the evaluation of UTDRM is conducted on diverse data.

In order to avoid any data leakage with the fact-checking articles utilised for claim generation (Sect. 3.1), we exclude all fact-checking articles that exhibit a Jaccard similarity of 0.5 or higher between the debunked claim statements. Please note that fact-checking articles used for claim generation are removed and are not from the evaluation datasets.

4.2 Baselines

Okapi BM25 We use the ElasticSearch[9] [43] implementation of BM25 [44], with default parameters in ElasticSearch ($k = 1.2$ and $b = 0.75$).

Out-of-the-box models We use two strong out-of-the-box pre-trained models for information retrieval. We test these models in their default configuration without any supervision from the generated claims to assess their zero-shot performance. The models are: (1) *Sentence-Transformer's* model based on Masked and Permuted Pre-training for Language Understanding (*MPNet*) [45] *all-mpnet-base-v2*[10] which has been trained on a large and diverse dataset of over a billion training examples. (2) Approximate Nearest Neighbor Negative Contrastive Estimation (*ANCE*), which is a RoBERTa [46] model fine-tuned on MSMARCO dataset [47] with hard negatives selected using approximate nearest neighbor [48].

Unsupervised methods We use five different unsupervised methods which utilise the same set of fact-checking articles for training, as used in the claim generation process (Sect. 3.1): (1) *ICT* [27] is employed to generate pseudo-claims by uniformly sampling

[9] https://www.elastic.co/elasticsearch/.

[10] https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

sentences from the fact-checking articles. MNRL loss (Sect. 3.2) is then applied to train the model using the pairs of pseudo and debunked claim statements. (2) Back-Translation (*BT*) [49] involves translating all debunked claim statements to Hindi and then back to English. The resulting pairs of back-translated claim and the original debunked claim statement are further used for training the model using MNRL loss. (3) *SimCSE* [25] encodes the same debunked claim statement twice with different dropout masks and utilises MNRL loss for training. (4) *TSDAE* [28] pre-trains a retrieval model using a denoising autoencoder. It encodes debunked claim statements with randomly deleted 60% of the tokens and the decoder reconstructs the original debunked claim statements [28]. All unsupervised methods employ a distilled version of the RoBERTa-base [46][11] as the underlying model. Hyperparameter details are in Appendix A.1.

Supervised methods We also report previous State-Of-The-Art (SOTA) performance achieved by the winners of the shared tasks on the test set, as published in their respective papers [2, 3, 14, 16]. Please note that these supervised methods benefit from annotated training data, which enables them to utilise specific information pertaining to real-world instances of misinformation claims and their corresponding debunks.

For example, the winning team of CLEF 22 2A [24] use Sentence-T5 [50] for candidate selection and GPT-Neo [51] for re-ranking. The winning team in CLEF 22 2B [52] employ a combination of semantic and lexical similarity features between claims and debunks for retrieval. In CLEF 21 2A [14], the top-performing team utilise a combination of TF-IDF, Sentence-BERT, and Lambda Multiple Additive Regression Trees (LambdaMART) for ranking [53], while the winning team in CLEF 21 2B [54] combines the Sentence-BERT model with a custom neural network to get the final list of sorted debunks based on relevance. The top-performing team in CLEF 20 2A [55] use a fine-tuned RoBERTa model for retrieval.

Lastly, for Snopes and Politifact, we directly report scores from Shaar et al. [16], who utilise a pairwise learning-to-rank model for debunk retrieval.

4.3 Experimental details

`UTDRM` is tested on two models: a distilled version of the RoBERTa-base model (`UTDRM-RoBERTa`) and the MPNet model (`UTDRM-MPNet`) (Sect. 5). We generate six topical claims (three from each claim generator) for all the collected 23,901 fact-checking articles (Sect. 3.1), as this approach yields the best scores during experiments (Sect. 6.2). Following previous work [29], we employ nucleus sampling during generation, using a *Top-k* value of 25 and a *Top-p* value of 0.95. For the ChatGPT claim generator, we keep all API parameters at their default values, except for the temperature, which is set to 0.7 to ensure diversity. The total cost of using ChatGPT to generate the claims was 14 GBP. Finally, a total of 1,43,406 (23,901x6) generated claims are used for training the neural retrieval model.

4.4 Evaluation metrics

For evaluation, we employ two widely used ranking metrics [3, 14]: Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP). MRR computes the score based on

[11] https://huggingface.co/distilroberta-base.

the highest-ranked relevant debunk for each misinformation tweet and is defined as MRR = $\frac{1}{|C|}\sum_{i=1}^{|C|}\frac{1}{\text{rank}_i}$, where $|C|$ is the number of input claims used as query and rank_i is the rank of the relevant debunk for the *ith* claim. The higher the MRR score the better. MAP, on the other hand, measures the precision of the system in returning relevant results for a given query. We use two variations of MAP: MAP@1 and MAP@5, which evaluate the top one and top five retrieved documents, respectively. A higher MAP@k score indicates better performance.

5 Results and discussion

Table 3 reports the results of UTDRM evaluation divided into two parts: the top part presents the individual and average results for *Twitter-based* datasets (Snopes, CLEF 22 2A-EN, CLEF 21 2A-EN & CLEF 20 2A-EN), while the bottom part showcases the individual and average results for *political-based* datasets (Politifact, CLEF 22 2B-EN & CLEF 21 2B-EN).

BM25 and out-of-the-box models These models consistently achieve high retrieval scores across all metrics, with MPNet outperforming the others (Table 3 column 3–5). This indicates that leveraging models trained on other information retrieval datasets can improve retrieval effectiveness (Sect. 4.2). However, it is important to note that there are variations in performance among the datasets, suggesting that the models' effectiveness might depend on the specific characteristics of the dataset.

Among the *Twitter-based* datasets, MPNet stands out as the best-performing model with the highest average scores. It achieves an average MAP@1 score of 0.841, MAP@5 score of 0.886, and MRR score of 0.888. In contrast, when considering the *political-based* datasets (Politifact, CLEF 22 2B-EN, and CLEF 21 2B-EN), BM25 emerges as the top-

Table 3 Performance of BM25, out-of-the-box, unsupervised and SOTA supervised models. The first part of the table shows the individual and average results for *Twitter-based* datasets, while the second part shows the individual and average results for *political-based* datasets. UTDRM results are highlighted in blue. The highest scores for each dataset and metric are in *bold*

Datasets	Metrics	Elastic	Out-of-the-box		Unsupervised Methods						Supervised
		BM25	MPNet	ANCE	BT	ICT	SimCSE	TSDAE	UTDRM-RoBERTa	UTDRM-MPNet	Prev SOTA
Snopes	MAP@1	0.557	0.776	0.662	0.333	0.627	0.545	0.458	0.716	**0.831**	0.691
	MAP@5	0.690	0.840	0.752	0.406	0.737	0.643	0.532	0.811	**0.889**	0.782
	MRR	0.786	0.843	0.759	0.418	0.745	0.652	0.548	0.815	**0.890**	0.788
CLEF 22 2A	MAP@1	0.823	0.866	0.761	0.368	0.756	0.589	0.469	0.823	0.933	**0.943**
	MAP@5	0.856	0.898	0.800	0.425	0.797	0.661	0.520	0.857	0.946	**0.956**
	MRR	0.862	0.899	0.807	0.444	0.804	0.674	0.539	0.861	0.948	**0.957**
CLEF 21 2A	MAP@1	0.797	0.837	0.767	0.332	0.762	0.644	0.510	0.817	**0.906**	0.861
	MAP@5	0.844	0.881	0.815	0.386	0.819	0.694	0.564	0.863	**0.933**	0.883
	MRR	0.849	0.885	0.823	0.406	0.825	0.704	0.579	0.869	**0.936**	0.884
CLEF 20 2A	MAP@1	0.834	0.884	0.869	0.372	0.769	0.673	0.578	0.874	**0.945**	0.897
	MAP@5	0.869	0.924	0.893	0.416	0.836	0.711	0.642	0.913	**0.961**	0.929
	MRR	0.878	0.925	0.896	0.436	0.840	0.722	0.653	0.915	**0.961**	0.927
Average Twitter-based	MAP@1	0.753	0.841	0.765	0.351	0.729	0.613	0.504	0.808	**0.904**	0.848
	MAP@5	0.815	0.886	0.815	0.408	0.797	0.677	0.565	0.861	**0.932**	0.888
	MRR	0.844	0.888	0.821	0.426	0.804	0.688	0.580	0.865	**0.934**	0.889
Politifact	MAP@1	0.467	0.413	0.428	0.387	0.445	0.355	0.424	0.426	0.516	**0.531**
	MAP@5	0.503	0.494	0.499	0.446	0.512	0.404	0.477	0.507	**0.600**	0.588
	MRR	0.541	0.524	0.532	0.464	0.543	0.431	0.504	0.539	**0.627**	0.608
CLEF 22 2B	MAP@1	0.308	0.285	0.331	0.277	0.254	0.254	0.269	0.369	0.392	**0.408**
	MAP@5	0.371	0.344	0.368	0.295	0.336	0.276	0.309	0.408	0.431	**0.459**
	MRR	0.419	0.374	0.413	0.337	0.377	0.319	0.349	0.459	0.467	**0.475**
CLEF 21 2B	MAP@1	0.285	0.247	0.272	0.222	0.215	0.209	0.241	0.310	**0.348**	0.304
	MAP@5	0.343	0.308	0.310	0.239	0.282	0.226	0.276	0.340	**0.392**	0.346
	MRR	0.377	0.333	0.344	0.268	0.317	0.262	0.307	0.386	**0.422**	0.350
Average Political-based	MAP@1	0.353	0.315	0.344	0.295	0.305	0.273	0.311	0.368	**0.419**	0.414
	MAP@5	0.406	0.382	0.392	0.327	0.377	0.302	0.354	0.418	**0.474**	0.464
	MRR	0.446	0.410	0.430	0.357	0.412	0.337	0.387	0.461	**0.505**	0.478

performing model with the average MAP@1 score of 0.353, MAP@5 score of 0.406, and MRR score of 0.446, indicating its effectiveness in retrieving relevant information from political speech datasets. Overall, the average scores suggest that the models perform better on the *Twitter-based* datasets compared to the *political-based* datasets. This difference in performance can be attributed to the fact that *political-based* claims pose greater challenges for the models.

Unsupervised methods Table 3 reports the results of the unsupervised methods, including the baselines BT, ICT, SimCSE, TSDAE (columns 6–9), as well as the proposed `UTDRM-RoBERTa` (Table 3 columns 10). All these methods utilise a distilled RoBERTa model, as described in Sect. 4.2. Among the baselines, ICT achieves the highest scores across all metrics, followed by SimCSE and TSDAE. However our proposed `UTDRM-RoBERTa` achieves the highest average scores for both *Twitter-based* and *political-based* datasets, followed by ICT and SimCSE. Additionally, the table reveals that each method has its own strengths and weaknesses on different datasets. For instance, `UTDRM-RoBERTa` performs well on all datasets except Politifact, where it is surpassed by ICT.

Furthermore, given the impressive performance of the out-of-the-box MPNet model, we also test `UTDRM` on the MPNet model (Table 3 column 11). `UTDRM-MPNet` outperforms all other methods, achieving the highest scores across all evaluation metrics. It obtains an average MAP@1, MAP@5, and MRR of 0.904, 0.932, and 0.934, respectively, for *Twitter-based* datasets. For *political-based* datasets, it achieves an average MAP@1, MAP@5, and MRR of 0.419, 0.474, and 0.505, respectively. Overall, `UTDRM-MPNet` consistently achieves the highest scores across all datasets, demonstrating its effectiveness. `UTDRM-RoBERTa` also performs well, albeit slightly lower than `UTDRM-MPNet`.

Supervised methods Table 3 (last column) reports the results for the previous SOTA methods (Sect. 4.2). These methods benefit from annotated training data, allowing them to leverage specific information about real-life misinformation claims and debunked claim statements (Sect. 4.2). In contrast, the `UTDRM` does not have access to any annotated training data. Surprisingly, the `UTDRM-MPNet` model, despite being an unsupervised method, achieves comparable or even superior retrieval scores compared to the SOTA supervised models. This demonstrates the effectiveness of `UTDRM` without the need for any annotations.

Summary We find that the choice of method depends on specific requirements, data availability, and the desired performance-resource trade-off. `UTDRM-RoBERTa` and `UTDRM-MPNet` consistently yield the highest retrieval scores, while the out-of-the-box models offer viable alternatives without the need for any training data whatsoever for debunked-narrative retrieval. Additionally, our proposed method, `UTDRM`, has the potential to detect topical misinformation claims by generating claims from incoming topical fact-checks; thus allowing it to address emerging topics and contribute to the timely detection of misinformation.

Table 4 Influence of fact-checking articles on UTDRM. The highest scores for each dataset and metric are in **bold**

Datasets	Metrics	Fact-checking Articles			
		1K	5K	10K	All
Snopes	MAP@1	0.750	0.794	0.803	**0.831**
	MRR	0.810	0.842	0.865	**0.890**
CLEF 22 2A	MAP@1	0.871	0.914	0.919	**0.933**
	MRR	0.904	0.932	0.936	**0.948**
CLEF 21 2A	MAP@1	0.851	0.891	**0.906**	**0.906**
	MRR	0.895	0.925	**0.937**	0.936
CLEF 20 2A	MAP@1	0.894	0.940	0.935	**0.945**
	MRR	0.931	0.957	0.957	**0.961**
Average Twitter-based	MAP@1 MRR	0.842 0.885	0.885 0.914	0.891 0.924	**0.904** **0.934**
Politifact	MAP@1	0.428	0.484	0.508	**0.516**
	MRR	0.541	0.602	0.618	**0.627**
CLEF 22 2B	MAP@1	0.285	0.362	0.377	**0.392**
	MRR	0.381	0.436	0.450	**0.467**
CLEF 21 2B	MAP@1	0.259	0.323	0.335	**0.348**
	MRR	0.346	0.390	0.402	**0.422**
Average Political-based	MAP@1 MRR	0.324 0.423	0.390 0.476	0.407 0.490	**0.419** **0.505**

6 Analysis

6.1 Influence of fact-checking articles

Table 4 shows the results UTDRM-MPNet when trained using different numbers of fact-checking articles (1K, 5K, 10K, and *All*). Due to space limitations, the table reports only the MAP@1 and MRR metrics.

The results suggest that the size of the corpus does have a postive effect on the performance of UTDRM, but the extent of the improvement may vary depending on the specific dataset and corpus size being used (Table 4). For instance, the CLEF 21 2A (*Twitter-based* dataset) shows an increasing trend until the number of fact-checking articles reaches 10K, after which it becomes relatively constant. On the other hand, for *political-based* datasets, the average performance continues to increase as the number of fact-checking articles increases, suggesting that a larger corpus of fact-checking articles has a more pronounced impact on improving retrieval performance.

6.2 Influence of the generated claims

Table 5 shows results of UTDRM-MPNet using different numbers of generated claims for training N: 2, 6, 10, 20. It should be noted that the proportion of claims generated using T5 and ChatGPT is kept the same for all cases. The individual performance of models trained on T5 and ChatGPT generated claims separately is generally lower (Appendix A.3 and A.2).

Table 5 demonstrates an overall improvement in performance as the number of generated claims increases from $N = 2$ to $N = 6$ and $N = 10$ across most datasets. However, performance either declines or stabilises beyond $N = 10$. For instance, in the Snopes dataset, MAP@1 and MRR scores show a slight decline from $N = 6$ to $N = 20$. Similar trends are observed in the CLEF 22 2A, CLEF 21 2A, and CLEF 20 2A datasets, where MAP@1 performance peaks at $N = 6$ and then plateaus or slightly decreases. In contrast, the CLEF 22 2B and CLEF 21 2B datasets reach their peak performance at $N = 10$. In general, the results suggest that $N = 6$ is the optimal value for the number of generated claims, as it yields

Table 5 Influence of the generated claims on UTDRM. The highest scores for each dataset and metric are in **bold**

Datasets	Metrics	Generated Claims			
		$N = 2$	$N = 6$	$N = 10$	$N = 20$
Snopes	MAP@1	0.821	**0.831**	0.830	0.829
	MRR	0.881	**0.890**	**0.890**	0.889
CLEF 22 2A	MAP@1	0.914	**0.933**	**0.933**	**0.933**
	MRR	0.934	0.948	0.948	**0.949**
CLEF 21 2A	MAP@1	0.906	**0.906**	0.901	0.896
	MRR	0.936	**0.936**	0.932	0.931
CLEF 20 2A	MAP@1	0.935	**0.945**	**0.945**	**0.945**
	MRR	0.957	0.961	0.963	**0.964**
Average	MAP@1	0.894	**0.904**	0.902	0.901
Twitter-based	MRR	0.927	**0.934**	0.933	0.933
Politifact	MAP@1	0.508	**0.516**	0.500	0.496
	MRR	0.616	**0.627**	0.619	0.615
CLEF 22 2B	MAP@1	0.362	0.392	**0.400**	0.392
	MRR	0.441	0.467	**0.473**	0.468
CLEF 21 2B	MAP@1	0.323	0.348	**0.354**	0.335
	MRR	0.394	0.422	**0.424**	0.416
Average	MAP@1	0.397	**0.419**	0.418	0.408
Political-based	MRR	0.484	**0.505**	**0.505**	0.499

the highest average retrieval performance, while going beyond this range may introduce noise and decrease performance.

6.3 Influence of *entity inoculation*

We propose an *entity inoculation* method, which involves replacing a random named entity in the generated claims with another random named entity to simulate real-world scenarios where similar misinformation narratives spread with different entities (see Appendix A.4 for examples). By training the model with these modified claims, it is expected to become more robust in retrieving debunked narratives regardless of the specific entities involved. Table 6 presents the results of *entity inoculation* using different entity types: geopolitical entities (GPE), person (PERSON), and organisation name (ORG), as well as a combined approach that uses all types. The *Default* column represents the performance of UTDRM-MPNet without *entity inoculation* (from Table 3).

Entity inoculation shows positive results on *political-based* datasets with an average increase of two MRR points with the combined approach as compared to the *Default* performance without *entity inoculation*. This indicates the effectiveness of *entity inoculation* in handling misinformation narratives in political contexts. On the other hand, for *Twitter-based* datasets, the impact of *entity inoculation* is less pronounced. While *entity inoculation* shows benefits in making models' entities agnostic, we hypothesise that its effectiveness may be limited to datasets that contain cases where similar narratives are spread with different entities. Examples of such false narratives can be found in Appendix A.4.

6.4 Influence of large language models (LLMs)

Large Language Models (LLMs) have consistently demonstrated impressive performance across a wide range of natural language processing (NLP) tasks [56, 57]. However, their application in information retrieval tasks remains an ongoing area of research, with the aim of optimising their ability to retrieve relevant information from large corpora in response to a given input query [58, 59]. Therefore, to assess the performance of LLMs in

Table 6 Influence of *entity inoculation* on UTDRM. UTDRM is the deafult UTDRM-MPNet performance from Table 3. The highest scores for each dataset and metric are in **bold**

Datasets	Metrics	*Entity Inoculation*				UTDRM
		GPE	PERSON	ORG	Combine	Default
Snopes	MAP@1	0.831	0.831	**0.841**	0.821	0.831
	MRR	0.889	0.891	**0.893**	0.881	0.890
CLEF 22 2A	MAP@1	0.928	0.919	0.923	0.919	**0.933**
	MRR	0.942	0.936	0.942	0.935	**0.948**
CLEF 21 2A	MAP@1	0.916	0.901	0.901	0.906	0.906
	MRR	0.940	0.929	0.932	0.932	0.936
CLEF 20 2A	MAP@1	0.940	0.930	0.940	0.935	**0.945**
	MRR	0.957	0.955	0.958	0.955	**0.961**
Average	MAP@1	0.904	0.895	0.901	0.895	**0.904**
Twitter-based	MRR	0.932	0.928	0.931	0.926	**0.934**
Politifact	MAP@1	0.492	**0.527**	0.512	0.512	0.516
	MRR	0.613	**0.637**	0.631	0.633	0.627
CLEF 22 2B	MAP@1	0.415	0.400	0.415	**0.423**	0.392
	MRR	0.482	0.471	0.482	**0.495**	0.467
CLEF 21 2B	MAP@1	0.367	0.354	0.367	**0.373**	0.348
	MRR	0.433	0.423	0.433	**0.442**	0.422
Average	MAP@1	0.425	0.427	0.431	**0.436**	0.419
Political-based	MRR	0.509	0.510	0.515	**0.524**	0.505

comparison to our UTDRM method, we employ a Listwise Re-ranker with a Large Language Model (LRL) [59] to re-rank the *Top-k* documents retrieved by the initial stage ranker. In this context, the LLM is provided with the following instruction template:

```
Passage1 = {Debunk_1}
...
PassageM = {Debunk_M}
Query = {Claim}
Passages = [Passage1, ..., PassageM]
Sort the Passages by their relevance to the Query.
Sorted Passages = [
```

Please note that due to LLM memory constraints, input sequences may exceed the maximum input sequence length. In such cases, we implement progressive re-ranking ($M = 20$) following the approach of Ma et al. [59]. This technique re-ranks M debunks at a time and incrementally shifts the window by $M/2$ towards the beginning of the retrieved debunks, leading to an enhancement in the top-ranked results. In this work, we test two types of LLMs: (1) the open-sourced LLaMA 2 13B [56, 57];[12] and (2) the private LLM ChatGPT (*gpt-3.5-turbo*). LLaMA 2 was hosted on our local server (2x24GB NVIDIA GeForce RTX 3090) and for ChatGPT, we use OpenAI API.[13] The total cost of testing using ChatGPT was 20 GBP.

Table 7 shows the results of LRL using BM25 and UTDRM-MPNet as first-stage rankers. For eg. "BM25+ChatGPT" (column 5 – Table 7) signifies that BM25 performs the first-stage ranking, and ChatGPT conducts the second-stage ranking. Following the methodology from prior work [59], the LLM is used to re-rank 100 documents on top of BM25 and 20 documents on top of UTDRM. The results indicate that ChatGPT outperforms LLaMA 2 across all datasets and metrics. Moreover, we find that re-ranking on top of UTDRM yields

[12] We use OpenAssistant's LLaMA 2 13B model for our experiments, accessible at https://huggingface.co/OpenAssistant/llama2-13b-orca-8k-3319.

[13] https://platform.openai.com/docs/models.

Table 7 Influence of large language models (LLaMA 2 and ChatGPT) as a second stage retriever to re-rank the top candidate claims retrieved by BM25 and `UTDRM`. `UTDRM` is the deafult `UTDRM-MPNet` performance from Table 3. UTDRM+ChatGPT signifies that `UTDRM-MPNet` performs the initial ranking, and ChatGPT conducts the second-stage ranking. The highest scores for each dataset and metric are in **bold**

Datasets	Metrics	BM25+LLaMA2	UTDRM+LLaMA2	BM25+ChatGPT	UTDRM+ChatGPT	UTDRM
Snopes	MAP@1	0.460	0.657	0.667	**0.841**	0.831
	MRR	0.659	0.728	0.862	**0.890**	**0.890**
CLEF 22 2A-EN	MAP@1	0.794	0.890	0.895	0.919	**0.933**
	MRR	0.835	0.913	0.916	0.936	**0.948**
CLEF 21 2A-EN	MAP@1	0.782	0.911	0.906	**0.926**	0.906
	MRR	0.836	0.939	0.927	**0.949**	0.936
CLEF 20 2A-EN	MAP@1	0.673	0.729	0.849	0.925	**0.945**
	MRR	0.724	0.762	0.894	0.948	**0.961**
Average	MAP@1	0.679	0.819	0.822	0.895	**0.904**
Twitter-based	MRR	0.777	0.860	0.902	0.925	**0.934**
Politifact	MAP@1	0.260	0.293	0.512	**0.561**	0.516
	MRR	0.333	0.417	0.607	**0.680**	0.627
CLEF 22 2B-EN	MAP@1	0.285	0.346	**0.400**	**0.400**	0.392
	MRR	0.383	0.426	0.486	**0.493**	0.467
CLEF 21 2B-EN	MAP@1	0.266	0.310	**0.361**	**0.361**	0.348
	MRR	0.347	0.388	**0.445**	0.425	0.422
Average	MAP@1	0.270	0.316	0.424	**0.441**	0.419
Political-based	MRR	0.354	0.411	0.513	**0.533**	0.505

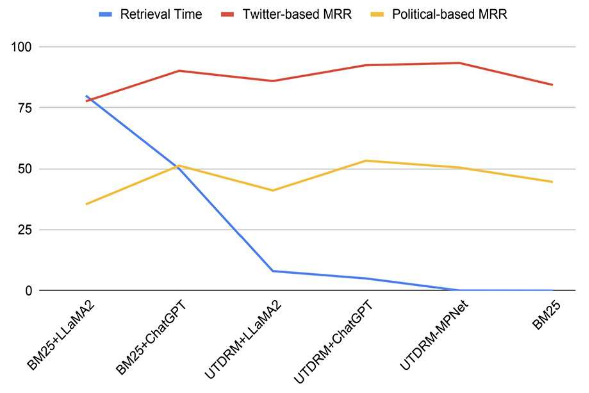

Figure 3 Retrieval times (in seconds per query) and average MRR performance (scaled by a factor of 100) comparison of different retrieval methods

superior scores compared to re-ranking on top of BM25 (Table 7). Figure 3 visually depicts the average MRR performance of different retrieval methods.

For the *Twitter-based* datasets, although `UTDRM` achieves the highest average scores, UTDRM+ChatGPT outperforms `UTDRM` in Snopes (MAP@1) and in CLEF 21 2A-EN (MAP@1 and MRR). For the *political-based* datasets, notably, UTDRM+ChatGPT beats `UTDRM` and attains the highest performance in MAP@1, MAP@5, and MRR across all datasets.

While LLMs exhibit impressive performance, it is important to consider the trade-offs, one of which is retrieval cost and latency. We conduct experiments to measure the time taken per claim to retrieve debunks for each method and we observe notable differences in retrieval speed. Figure 3 shows retrieval times and average MRR performance comparison of different retrieval methods. We find that BM25+LLaMA2 and BM25+ChatGPT exhibit longer retrieval times, averaging around 80 seconds and 50 seconds per claim, respectively. In contrast, UTDRM+LLaMA2 and UTDRM+ChatGPT significantly reduce retrieval time, taking only 8 seconds and 5 seconds per claim, respectively, possibly due to the fewer number of debunks to be re-ranked. Remarkably, `UTDRM-MPNet` on its own achieves an exceptionally low retrieval time of just 0.04 seconds per claim. These findings underscore that, despite LLMs' impressive performance in relevance ranking, they often come at the cost of extended retrieval times, whereas our proposed `UTDRM-MPNet` approach offers both high relevance and exceptional retrieval speed.

7 Error analysis

The evaluation of UTDRM would be incomplete without a thorough examination of the types of errors it may produce. To address this, we manually review cases where the retrieval model fails to rank the most relevant debunked claim at the top. We conduct this analysis by inspecting the retrieved debunked claims for 50 randomly selected cases from the Snopes and Politifact datasets. We find that the primary cause of such errors is when a misinformation claim is associated with multiple debunked claims (19 out of 50). For instance, the false claim "African Union warning African citizens against the safety of travelling to the United States" in Snopes has multiple relevant debunked claims. In such instances, the model assigns highly similar high scores to all relevant debunked claims, even though each misinformation claim is linked to a single debunked claim in the dataset. This highlights inconsistencies in the existing datasets and the need for further improvement.

The second type of error occurs when the retrieved debunked claim is not entirely relevant, but there is some degree of relevance to the input misinformation claim (16 out of 50). For instance, for the claim "Governor Christie has endorsed many of the ideas that Barack Obama supports, whether it is gun control or the appointment of Sonia Sotomayor", the top retrieved debunked claim discusses Governor Chris Christie and Barack Obama sharing similar views on gay marriage. This highlights the challenge of distinguishing closely related debunked claims, emphasising the need for continued refinement in retrieval models for enhanced precision. Moreover, we hypothesise that this may also be attributed to limitations in the claim generation model, where it generates claims that, while not entirely irrelevant, are only tangentially related to the intended debunked claim. Such errors suggest the propagation of errors in the retrieval process and suggests the need for improvement in the claim generation model.

The third category, accounting for 15 out of 50 cases, involves errors that occur when a misinformation claim lacks sufficient context to find the relevant debunked claim. For example, one of the misinformation claims in the Politifact dataset states "very few children" which is ambiguous and makes finding a relevant debunk challenging. Moreover, the task becomes even more challenging when misinformation claims span multiple modalities, such as combining text and images. For instance, one of the misinformation claims is a X (formerly Twitter) post stating "Botswana condemns remarks made by President Trump", along with an image containing details of the remarks. In such cases, retrieval models also require information contained in the image, as the text of the tweet alone is not sufficient. This motivates future work on multimodal debunked-narrative retrieval, where models can exploit joint information from different modalities.

8 Conclusion

This paper presents UTDRM, an unsupervised method for training debunked-narrative retrieval models that effectively overcomes the reliance on manually annotated training data. UTDRM introduces a novel approach to synthetically generate large-scale topical claims from fact-checking articles. A comprehensive comparison with other out-of-the-box, unsupervised, and supervised models confirm the efficacy of UTDRM in retrieving accurate debunked claims. In general, UTDRM-MPNet and UTDRM-RoBERTa consistently achieve the highest scores across all datasets, with UTDRM-MPNet exhibiting slightly better performance.

Furthermore, this study emphasises the importance of corpus size, demonstrating that larger corpora contribute to improved retrieval performance. The paper also examines how different factors, such as the quantity of synthetically generated claims used and the *entity inoculation* method, influence the performance of UTDRM. While *entity inoculation* shows benefits in making models entity agnostic, its effectiveness may be limited to cases involving narratives that adapt and propagate with different entities.

Additionally, this paper experiments with state-of-the-art LLMs as listwise re-rankers and compares them to our UTDRM method. While LLMs exhibit slight performance improvements over UTDRM on some datasets, their use comes at the cost of lower computational efficiency, making UTDRM a more practical choice for real-time applications.

Finally, UTDRM allows models to adapt and learn from synthetically generated topical claims in real-time; thus providing significant benefits in combating ever-evolving topical misinformation.

9 Limitations and future work

The present work acknowledges certain limitations and identifies several avenues for future improvement. Firstly, this study focused solely on English-language datasets and did not explore cross-lingual retrieval. However, the UTDRM approach can be replicated and adapted to other languages using pre-trained multilingual language models. Conducting cross-lingual experiments would provide a more comprehensive understanding of UTDRM's performance and applicability in diverse linguistic contexts, thereby extending its potential impact in combating misinformation on a global scale. Additionally, future work can include testing on a broader range of fact-checking articles and exploring novel approaches to further improve the information retrieval models used in UTDRM.

Appendix

A.1 Hyperparameters

For the T5 claim generator, we fine-tune the base variant of the T5 model[14] using a constant learning rate of $1e$–4 for 2 epochs, with a batch size of 12. The maximum input tokens allowed is 512, and the maximum output tokens is set to 64.

The training details for the neural retrieval model are as follows. UTDRM-RoBERTa is fine-tuned for two epochs with a batch size of 64 and a learning rate of $4e$–5. For UTDRM-MPNet, we fine-tune it for one epoch with a batch size of 64 and a learning rate of $8e$–7. The maximum input sequence length is set to 350, the optimiser used is AdamW and we

[14] https://huggingface.co/t5-base.

use linear warmup as the learning rate scheduler. Hard negatives for training the neural retrieval model are mined using the *all-mpnet-base-v2*[15] and *all-MiniLM-L12-v2*[16] models because of their demonstrated efficacy .[17] Both `UTDRM-RoBERTa` and `UTDRM-MPNet` are validated using the respective dataset's validation set, and we manually tune the hyperparameters based on the evaluation metrics (Sect. 4.3). The hyperparameter bounds are as follows: 1) Epochs range from 1 to 5, 2) Learning rate ranges from $1e–7$ to $1e–5$, and 3) Batch size ranges from 8 to 64, limited by the GPU requirements of the model. The training time for each epoch ranges from 10 to 15 minutes.

For the baselines, BT and ICT use the same hyperparameters as `UTDRM-RoBERTa` to ensure a fair comparison. For SimCSE and TSDAE, we use the same hyperparameters as stated by the authors in their respective papers [25, 28]. Finally, all experiments are conducted on a machine with a 24GB NVIDIA GeForce RTX 3090.

A.2 Influence of ChatGPT claims

Table 8 shows the performance of the `UTDRM-MPNet` model trained using different numbers of generated claims using ChatGPT ($N = 1$, $N = 2$, $N = 6$, $N = 10$). The datasets are divided into two categories: *Twitter-based* datasets (Snopes, CLEF 22 2A, CLEF 21 2A, CLEF 20 2A) and *political-based* datasets (Politifact, CLEF 22 2B, CLEF 21 2B).

From Table 8, we can observe that the model generally performs better on *Twitter-based* datasets, with the highest MAP@1 and MRR values of 0.945 and 0.962 respectively, recorded on the CLEF 20 2A dataset with $N = 6$ and $N = 10$ generated claims. In contrast, performance on *political-based* datasets is comparatively lower, with the highest MAP@1 and MRR values of 0.512 and 0.612 respectively, both recorded on the Politifact

Table 8 Influence of ChatGPT generated claims. The highest scores for each dataset and metric are in **bold**

Datasets	Metrics	ChatGPT Generated Claims			
		$N = 1$	$N = 2$	$N = 6$	$N = 10$
Snopes	MAP@1	0.811	**0.826**	0.813	0.811
	MRR	0.869	**0.882**	0.880	0.879
CLEF 22 2A	MAP@1	0.904	0.909	0.919	**0.923**
	MRR	0.926	0.929	0.937	**0.940**
CLEF 21 2A	MAP@1	0.876	**0.901**	**0.901**	0.896
	MRR	0.915	**0.931**	**0.931**	0.929
CLEF 20 2A	MAP@1	0.940	0.935	**0.945**	**0.945**
	MRR	0.956	0.955	**0.962**	**0.962**
Average Twitter-based	MAP@1	0.883	0.893	**0.894**	**0.894**
	MRR	0.917	0.924	**0.928**	**0.928**
Politifact	MAP@1	0.461	0.477	**0.512**	0.496
	MRR	0.572	0.593	**0.612**	0.606
CLEF 22 2B	MAP@1	0.346	0.346	0.377	**0.385**
	MRR	0.423	0.424	0.448	**0.458**
CLEF 21 2B	MAP@1	0.310	0.310	0.335	**0.342**
	MRR	0.379	0.381	0.403	**0.411**
Average Political-based	MAP@1	0.372	0.378	**0.408**	0.407
	MRR	0.458	0.466	0.488	**0.492**

[15] https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[16] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2.

[17] https://www.sbert.net/docs/pretrained_models.html.

Table 9 Influence of T5 generated claims. The highest scores for each dataset and metric are in **bold**

Datasets	Metrics	T5 Generated Claims			
		N = 1	*N* = 2	*N* = 6	*N* = 10
Snopes	MAP@1	0.811	0.821	0.846	**0.851**
	MRR	0.870	0.879	0.898	**0.900**
CLEF 22 2A	MAP@1	0.900	0.909	**0.928**	**0.928**
	MRR	0.923	0.930	**0.946**	0.945
CLEF 21 2A	MAP@1	0.886	0.901	**0.916**	0.906
	MRR	0.923	0.933	**0.938**	0.937
CLEF 20 2A	MAP@1	0.935	**0.935**	**0.935**	**0.935**
	MRR	0.954	0.955	**0.957**	0.955
Average	MAP@1	0.883	0.891	**0.906**	0.905
Twitter-based	MRR	0.917	0.924	**0.935**	0.934
Politifact	MAP@1	0.484	**0.516**	**0.516**	0.500
	MRR	0.598	0.628	**0.637**	0.627
CLEF 22 2B	MAP@1	0.392	0.408	**0.415**	**0.415**
	MRR	0.451	0.473	0.487	**0.490**
CLEF 21 2B	MAP@1	0.348	0.361	0.354	**0.367**
	MRR	0.403	0.420	0.431	**0.439**
Average	MAP@1	0.408	**0.428**	**0.428**	0.427
Political-based	MRR	0.484	0.507	0.518	**0.519**

dataset with six generated claims ($N = 6$). Furthermore, the performance generally tends to improve with more generated claims, however, there are exceptions. On the Snopes and CLEF 21 2A datasets, performance dips slightly when increasing generated claims from $N = 2$ to $N = 10$. Overall, these observations suggest that the optimal number of claims to generate for best performance can vary depending on the specific dataset and whether it is *Twitter-based* or *political-based*.

A.3 Influence of T5 claims

Table 9 shows the performance of the `UTDRM-MPNet` model trained using different numbers of generated claims using T5 ($N = 1$, $N = 2$, $N = 6$, $N = 10$). On the *Twitter-based* datasets, the model reaches peak performance on the CLEF 20 2A dataset with $N = 6$ generated claims (MAP@1 = 0.935 and MRR = 0.957). On *political-based* datasets, the model achieves maximum performance on the Politifact dataset with $N = 6$ generated claims (MAP@1 = 0.516 and MRR = 0.637). In general, finding an optimal number of generated claims for the best performance varies depending on the dataset, and the pattern is different from that of the ChatGPT generated claims (Sect. A.2).

A.4 *Entity inoculation* motivation

Table 10 illustrates an intriguing aspect of misinformation – it tends to replicate across diverse contexts and entities, applying similar narratives or themes to varied situations. The first example shows similar claims about "Crocodiles". These falsehoods involve the sighting of crocodiles in flooded city streets but vary by location — Hyderabad, Patna, Bengaluru, Aligarh and Florida. This shows how a single false narrative can be adapted to fit multiple geographical contexts, fueling misinformation in different locations. Similarly, The second example shows claims around "Sushant Singh Rajput's Death" (Table 10). These false narratives revolve around the demand for a CBI inquiry into the actor's death. The narrative remains consistent but the entities change – one claim implicates Rajput's father, KK Singh, and the other brings in PM Modi and Amit Shah. These falsehoods illustrate how misinformation can persist by switching the characters.

Table 10 Examples showcasing the variation of similar debunked claims across multiple entities and contexts, with corresponding fact-check links. The text in **bold** shows difference in named entities between the claims

Debunked Claim	Fact-check Link
Claim 1: Crocodile swimming on a flooded street in the south Indian city of **Hyderabad**.	https://factcheck.afp.com/no-footage-has-circulated-2019-reports-about-crocodile-west-india
Claim 2: Crocodile seen during flood in **Patna**, **Bihar**.	https://www.boomlive.in/crocodile-spotted-during-bihar-floods-video-from-gujarat-shared-as-patna/
Claim 3: Crocodile spotted in the waterlogged streets of **Bengaluru**.	https://www.indiatoday.in/fact-check/story/fact-check-crocodile-spotted-waterlogged-bengaluru-viral-video-mp-1997133-2022-09-06
Claim 4: Crocodile seen during flood in **Aligarh**, **Uttar Pradesh**.	https://factly.in/this-video-of-a-crocodile-swimming-in-a-flooded-street-was-captured-in-madhya-pradesh-not-aligarh/
Claim 5: Crocodile seen during flood in **Florida**.	https://factcheck.afp.com/doc.afp.com.32KT6D7
Claim 1: Sushant Singh Rajputs father **KK Singh** has demanded a CBI inquiry into his death.	https://www.boomlive.in/fake-news/sushant-singh-rajput-ians-jagran-fall-for-fake-account-demanding-cbi-probe-8750
Claim 2: **PM Modi** has ordered for a CBI inquiry into Sushant Singh Rajputs death.	https://www.boomlive.in/fake-news/no-pm-modi-did-not-order-cbi-inquiry-into-sushant-singh-rajputs-death-8736
Claim 3: **Amit Shah** ordered CBI probe for investigating Sushant Singh Rajput's death.	https://www.boomlive.in/fake-news/no-amit-shah-did-not-order-cbi-probe-into-sushant-singh-rajputs-death-8939
Claim 1: A video in which a **woman** suffers a seizure on the floor in an **Argentine hospital** after the woman was vaccinated against covid-19 .	https://checamos.afp.com/mulher-que-sofreu-uma-convulsao-em-um-hospital-argentino-nao-foi-vacinada-contra-covid-19
Claim 2: A video shows a **man** fainting after receiving the Covid-19 vaccine in **Indonesia's West Nusa Tenggara** province.	https://factcheck.afp.com/video-actually-shows-simulation-exercise-indonesia-not-real-covid-19-vaccination
Claim 1: A video shows meeting of the **Pacific Ocean** and **Atlantic ocean**, but without that they mix.	https://checamos.afp.com/este-video-mostra-o-rio-fraser-se-encontrando-com-o-oceano-pacifico-no-canada
Claim 2: A video which shows a place where the **Indian Ocean** meets the **Atlantic ocean**, and the waters of the two oceans do not mix.	https://napravoumiru.afp.com/toto-video-nezachycuje-misto-kde-se-setkavaji-dva-oceany
Claim 3: A video shows meeting of the **Gulf of Mexico** and **Mississippi River**, but without mixing.	https://www.snopes.com/fact-check/mississippi-meets-gulf-mexico/

In summary, Table 10 highlights the importance of our adopted approach of *entity inoculation*, as detailed in Sect. 6.3. This method involves replacing one randomly chosen named entity in the generated claims with another random named entity, with the intent to mimic real-world scenarios where similar misinformation narratives disseminate involving different entities. This emphasises both the adaptability and resilience of misinformation, underlining the need for effective methods like *entity inoculation* to detect debunked narratives.

Acknowledgements
We thank the anonymous reviewers for their comments and suggestions.

Funding
This research has been partially supported by a UKRI grant EP/W011212/1 and the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities" – Grant Agreement n. 871042 ("SoBigData++: European Integrated Infrastructure for Social Mining and BigData Analytics" (http://www.sobigdata.eu)).

Abbreviations
LLMs, Large Language Models; BERT, Bidirectional Encoder Representations from Transformers; RoBERTa, Robustly Optimized BERT Approach; UTDRM, Unsupervised Method for Training Debunked-Narrative Retrieval Models; LLaMA, Large Language Model Meta AI; ChatGPT, Chat Generative Pre-trained Transformer; BM25, Best Match 25; TF-IDF, Term Frequency – Inverse Document Frequency; QL, Query Likelihood model; DFR, Divergence From Randomness; CLEF, Conference and Labs of the Evaluation Forum; T5, Text-to-Text Transfer Transformer; SimCSE, Simple Contrastive Learning of Sentence Embeddings; ICT, Inverse Cloze Task; TSDAE, Tranformer-based Denoising AutoEncoder; GPL, Generative Pseudo Labeling; MSMARCO, Microsoft Machine Reading Comprehension; MSE, Mean Squared Error; AFP, Agence France-Presse; GPT-3.5, Generative Pretrained Transformer versions 3.5; ROUGE, Recall-Oriented Understudy for Gisting Evaluation; BLEU, Bilingual Evaluation Understudy; MNRL, Multiple Negatives Ranking Loss; MPNet, Masked and Permuted Pre-training for Language Understanding; ANCE, Approximate Nearest Neighbor Negative Contrastive Estimation; BT, Back-Translation; SOTA, State-Of-The-Art; LambdaMART, Lambda Multiple Additive Regression Trees; MRR, Mean Reciprocal Rank; MAP, Mean Average Precision; GPE, Geopolitical Entities; PERSON, Person; ORG, Organisation; NLP, Natural Language Processing; LRL, Listwise Re-ranker with a Large Language Model.

Data availability
The datasets supporting the conclusions of this article are publicly available: (1) Snopes and Politifact [16] https://github.com/sshaar/That-is-a-Known-Lie (2) CLEF 22 2A-EN and 2B-EN [3] https://sites.google.com/view/clef2022-checkthat (3) CLEF 21 2A-EN and 2B-EN [14] https://sites.google.com/view/clef2021-checkthat (4) CLEF 20 2A-EN [2] https://sites.google.com/view/clef2020-checkthat. To facilitate repeatability, UTDRM models and code are made publicly available at https://github.com/iknoorjobs/UTDRM.

Declarations

Ethics approval and consent to participate
This research has received ethics approval by the Sheffield University Ethics Board. While this paper involves generating false claims for research purposes, its overarching goal is to develop effective techniques for identifying already debunked narratives. The synthetically generated false claims dataset is solely for evaluation and will only be made available to academic researchers following careful vetting and a signed contract, in order to prevent public harm or spreading of misinformation. Furthermore, the research demonstrates that `UTDRM` is an effective method for training debunked-narrative retrieval models without the need for annotations, which are often time-consuming, expensive, and limited in scale. The overall aim is to promote ethical technology use and advance misinformation debunking efforts for the benefit of fact-checkers and users in general.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author contributions
IS developed the UDTM method, performed the experiments, and wrote the original paper draft. CS and KB provided direction with conceptualisation and methodology. All authors edited and submitted the final manuscript. All authors read and approved the final manuscript.

Received: 11 September 2023 Accepted: 1 December 2023 Published online: 13 December 2023

References

1. Procter R, Catania MA, He Y, Liakata M, Zubiaga A, Kochkina E, Zhao R (2023) Some observations on fact-checking work with implications for computational support. arXiv preprint. arXiv:2305.02224

2. Shaar S, Nikolov A, Babulkov N, Alam F, Barrón-Cedeno A, Elsayed T, Hasanain M, Suwaileh R, Haouari F, Da San Martino G et al (2020) Overview of CheckThat! 2020 English: automatic identification and verification of claims in social media. In: CLEF (working notes)
3. Nakov P, Da San Martino G, Alam F, Shaar S, Mubarak H, Babulkov N (2022) Overview of the CLEF-2022 CheckThat! lab task 2 on detecting previously fact-checked claims
4. Nakov P, Corney D, Hasanain M, Alam F, Elsayed T, Barrón-Cedeño A, Papotti P, Shaar S, Martino GDS (2021) Automated fact-checking for assisting human fact-checkers. ArXiv preprint. arXiv:2103.07769
5. Kazemi A, Garimella K, Gaffney D, Hale S (2021) Claim matching beyond English to scale global fact-checking. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp 4504–4517. https://doi.org/10.18653/v1/2021.acl-long.347. https://aclanthology.org/2021.acl-long.347
6. Singh I, Bontcheva K, Scarton C (2021) The false covid-19 narratives that keep being debunked: a spatiotemporal analysis. ArXiv preprint. arXiv:2107.12303
7. Robertson S, Zaragoza H et al (2009) The probabilistic relevance framework: Bm25 and beyond. Found Trends Inf Retr 3(4):333–389
8. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523
9. Ponte JM, Croft WB (2017) A language modeling approach to information retrieval. In: ACM SIGIR forum, vol 51. ACM, New York, pp 202–208
10. Amati G, Van Rijsbergen CJ (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans Inf Syst 20(4):357–389. https://doi.org/10.1145/582415.582416
11. Berger A, Caruana R, Cohn D, Freitag D, Mittal V (2000) Bridging the lexical chasm: statistical approaches to answer-finding. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, pp 192–199
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
13. Thakur N, Reimers N, Rücklé A, Srivastava A, Gurevych I (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)
14. Nakov P, Da San Martino G, Elsayed T, Barrón-Cedeno A, Míguez R, Shaar S, Alam F, Haouari F, Hasanain M, Babulkov N et al (2021) The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: ECIR, vol 2
15. Hardalov M, Chernyavskiy A, Koychev I, Ilvovsky D, Nakov P (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, pp 266–285
16. Shaar S, Babulkov N, Da San Martino G, Nakov P (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, pp 3607–3618. https://doi.org/10.18653/v1/2020.acl-main.332. https://aclanthology.org/2020.acl-main.332
17. Sheng Q, Cao J, Zhang X, Li X, Zhong L (2021) Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp 5468–5481. https://doi.org/10.18653/v1/2021.acl-long.425. https://aclanthology.org/2021.acl-long.425
18. Bhatnagar V, Kanojia D, Chebrolu K (2022) Harnessing abstractive summarization for fact-checked claim detection. In: Proceedings of the 29th international conference on computational linguistics, pp 2934–2945
19. Vo N, Lee K (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, pp 7717–7731. https://doi.org/10.18653/v1/2020.emnlp-main.621. https://aclanthology.org/2020.emnlp-main.621
20. Shaar S, Alam F, Da San Martino G, Nakov P (2022) The role of context in detecting previously fact-checked claims. In: Findings of the association for computational linguistics: NAACL 2022. Association for Computational Linguistics, Seattle, pp 1619–1631. https://doi.org/10.18653/v1/2022.findings-naacl.122. https://aclanthology.org/2022.findings-naacl.122
21. Zhou J, Han X, Yang C, Liu Z, Wang L, Li C, Sun M (2019) GEAR: graph-based evidence aggregating and reasoning for fact verification. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, pp 892–901. https://doi.org/10.18653/v1/P19-1085. https://aclanthology.org/P19-1085
22. Kazemi A, Li Z, Pérez-Rosas V, Hale SA, Mihalcea R (2022) Matching tweets with applicable fact-checks across languages. ArXiv preprint. arXiv:2202.07094
23. Barrón-Cedeño A, Alam F, Caselli T, Da San Martino G, Elsayed T, Galassi A, Haouari F, Ruggeri F, Struß JM, Nandi RN et al (2023) The CLEF-2023 CheckThat! Lab: checkworthiness, subjectivity, political bias, factuality, and authority. In: European conference on information retrieval. Springer, Berlin, pp 506–517
24. Shliselberg S-HM, Dori-Hacohen S (2022) RIET Lab at CheckThat! 2022: improving decoder based re-ranking for claim matching. Working Notes of CLEF
25. Gao T, Yao X, Chen D (2021) SimCSE: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. https://aclanthology.org/2021.emnlp-main. 552

26. Frick RA, Vogel I (2022) Fraunhofer SIT at CheckThat! 2022: ensemble similarity estimation for finding previously fact-checked claims. Working Notes of CLEF
27. Lee K, Chang M-W, Toutanova K (2019) Latent retrieval for weakly supervised open domain question answering. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, pp 6086–6096. https://doi.org/10.18653/v1/P19-1612. https://aclanthology.org/P19-1612
28. Wang K, Reimers N, Gurevych I (2021) TSDAE: using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In: Findings of the association for computational linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, pp 671–688. https://doi.org/10.18653/v1/2021.findings-emnlp.59. https://aclanthology.org/2021.findings-emnlp.59
29. Wang K, Thakur N, Reimers N, Gurevych I (2022) GPL: generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics, Seattle, pp 2345–2360. https://doi.org/10.18653/v1/2022.naacl-main.168. https://aclanthology.org/2022.naacl-main.168
30. Chang W, Yu FX, Chang Y, Yang Y, Kumar S (2020) Pre-training Tasks for Embedding-based Large-scale Retrieval. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020). https://openreview.net/forum?id=rkg-mA4FDr
31. Carlsson F, Gyllensten AC, Gogoulou E, Hellqvist EY, Sahlgren M (2021) Semantic Re-tuning with Contrastive Tension. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 OpenReview.net. https://openreview.net/forum?id=Ov_sMNau-PF
32. Nogueira R, Lin J, Epistemic A (2019) From doc2query to docTTTTTquery. Online preprint
33. Nogueira R, Yang W, Lin J, Cho K (2019) Document expansion by query prediction. ArXiv preprint. arXiv:1904.08375
34. Ma J, Korotkov I, Yang Y, Hall K, McDonald R (2021) Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, pp 1075–1088. https://doi.org/10.18653/v1/2021.eacl-main.92. https://aclanthology.org/2021.eacl-main.92
35. Hofstätter S, Althammer S, Schröder M, Sertkan M, Hanbury A (2020) Improving efficient neural ranking models with cross-architecture knowledge distillation. ArXiv preprint. arXiv:2010.02666
36. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21:140–114067
37. Lin C-Y (2004) ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out. Association for Computational Linguistics, Barcelona, pp 74–81. https://aclanthology.org/W04-1013
38. Shu R, Nakayama H, Cho K (2019) Generating diverse translations with sentence codes. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1823–1827
39. Oord AVD, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint. arXiv:1807.03748
40. Henderson M, Al-Rfou R, Strope B, Sung Y-H, Lukács L, Guo R, Kumar S, Miklos B, Kurzweil R (2017) Efficient natural language response suggestion for smart reply. ArXiv preprint. arXiv:1705.00652
41. Reimers N, Gurevych I (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, pp 3982–3992. https://doi.org/10.18653/v1/D19-1410. https://aclanthology.org/D19-1410
42. Ioffe S (2010) Improved consistent sampling, weighted minhash and l1 sketching. In: 2010 IEEE international conference on data mining. IEEE, Los Alamitos, pp 246–255
43. Gormley C, Tong Z (2015) Elasticsearch: the definitive guide: a distributed real-time search and analytics engine. O'Reilly Media
44. Jones KS, Walker S, Robertson SE (2000) A probabilistic model of information retrieval: development and comparative experiments: part 2. Inf Process Manag 36(6):809–840
45. Song K, Tan X, Qin T, Lu J, Liu T (2020) MPNet: masked and permuted pre-training for language understanding. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html
46. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized bert pretraining approach. ArXiv preprint. arXiv:1907.11692
47. Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, Deng L (2016) MS MARCO: a human generated machine reading comprehension dataset. In: Besold TR, Bordes A, d'Avila Garcez AS, Wayne G (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, CEUR workshop proceedings, vol 1773. CEUR-WS.org
48. Xiong L, Xiong C, Li Y, Tang K, Liu J, Bennett PN, Ahmed J, Overwijk A (2021) Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 OpenReview.net. https://openreview.net/forum?id=zeFrfgyZln
49. Sennrich R, Haddow B, Birch A (2016) Improving neural machine translation models with monolingual data. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 86–96. https://doi.org/10.18653/v1/P16-1009. https://aclanthology.org/P16-1009
50. Ni J, Abrego GH, Constant N, Ma J, Hall K, Cer D, Yang Y (2022) Sentence-t5: scalable sentence encoders from pre-trained text-to-text models. In: Findings of the association for computational linguistics: ACL 2022, pp 1864–1874
51. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N et al (2020) The pile: an 800gb dataset of diverse text for language modeling. arXiv preprint. arXiv:2101.00027
52. Hövelmeyer A, Boland K, Dietze S (2022) SimBa at CheckThat! 2022: lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way. Working Notes of CLEF

53. Chernyavskiy A, Ilvovsky D, Nakov P (2021) Aschern at CheckThat! 2021: lambda-calculus of fact-checked claims. Faggioli et al. [12]
54. Mihaylova S, Borisova I, Chemishanov D, Hadzhitsanev P, Hardalov M, Nakov P (2021) Dips at checkthat! 2021: verified claim retrieval. In: CLEF (working notes), pp 558–571
55. Bouziane M, Perrin H, Cluzeau A, Mardas J, Sadeq A (2020) Team buster. Ai at checkthat! 2020 insights and recommendations to improve fact-checking. In: CLEF (working notes)
56. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al (2023) LLaMA: open and efficient foundation language models. ArXiv preprint. arXiv:2302.13971
57. Köpf A, Kilcher Y, von Rütte D, Anagnostidis S, Tam Z-R, Stevens K, Barhoum A, Duc NM, Stanley O, Nagyfi R et al (2023) Openassistant conversations–democratizing large language model alignment. arXiv preprint. arXiv:2304.07327
58. Ai Q, Bai T, Cao Z, Chang Y, Chen J, Chen Z, Cheng Z, Dong S, Dou Z, Feng F et al (2023) Information retrieval meets large language models: a strategic report from chinese ir community. AI Open
59. Ma X, Zhang X, Pradeep R, Lin J (2023) Zero-shot listwise document reranking with a large language model. arXiv preprint. arXiv:2305.02156

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.