RESEARCH-ARTICLE

Multimodal and Multilingual Fact-Checked Article Retrieval

STEFANOS IORDANIS PAPADOPOULOS, Centre for Research and Technology-Hellas, Thessaloniki, Macedonia, Greece

IVANA BEŇOVÁ, Kempelen Institute of Intelligent Technologies, Bratislava, Bratislava Region, Slovakia

SEBASTIAN KULA, Kempelen Institute of Intelligent Technologies, Bratislava, Bratislava Region, Slovakia

MICHAL GREGOR, Kempelen Institute of Intelligent Technologies, Bratislava, Bratislava Region, Slovakia

GEORGE KARANTAIDIS, Centre for Research and Technology-Hellas, Thessaloniki, Macedonia, Greece

TOMÁŠ JAVŮREK, Kempelen Institute of Intelligent Technologies, Bratislava, Bratislava Region, Slovakia

View all

Published: 30 June 2025

Citation in BibTeX format

ICMR '25: International Conference on Multimedia Retrieval
June 30 - July 3, 2025
IL, Chicago, USA

Conference Sponsors:
SIGMM

ICMR '25: Proceedings of the 2025 International Conference on Multimedia Retrieval (June 2025)
https://doi.org/10.1145/3731715.3733402
ISBN: 9798400718779

Multimodal and Multilingual Fact-Checked Article Retrieval

Stefanos-Iordanis Papadopoulos
Centre for Research and Technology Hellas, Information Technologies Institute
Thessaloniki, Greece
stefpapad@iti.gr

Ivana Beňová
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
ivana.benova@kinit.sk

Sebastian Kula
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
sebastian.kula@kinit.sk

Michal Gregor
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
michal.gregor@kinit.sk

George Karantaidis
Centre for Research and Technology Hellas, Information Technologies Institute
Thessaloniki, Greece
karantai@iti.gr

Tomáš Javůrek
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
tomas.javurek@kinit.sk

Marián Šimko
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
marian.simko@kinit.sk

Symeon Papadopoulos
Centre for Research and Technology Hellas, Information Technologies Institute
Thessaloniki, Greece
papadop@iti.gr

Abstract

Fact-Check Retrieval (FCR) plays a crucial role in automated fact-checking by retrieving relevant fact-checked articles for disputed claims. While recent work has explored text-based, multilingual, and multimodal FCR, most efforts remain unimodal or limited to English. To bridge this gap, we introduce *M3-Check*, the first FCR dataset combining multilingual texts and images from social media posts with fact-check articles from diverse, credible sources. Furthermore, we introduce *FACTOR* a two-tower Transformer-based architecture that employs cross-tower parameter sharing and modality-wise aligned weight initialization; that outperforms zero-shot baselines, two-tower linear models, and vanilla Transformers, achieving a 17% improvement over the latter. Moreover we conduct modality ablations and compare state-of-the-art encoders, showing that multilingual encoders like multi-E5 can provide an additional 13% in performance without requiring English translations.

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**; **Multilingual and cross-lingual retrieval**; • **Computing methodologies** → *Image representations*; *Information extraction.*

ICMR '25, Chicago, IL, USA

ACM ISBN 979-8-4007-1877-9/2025/06
https://doi.org/10.1145/3731715.3733402

Keywords

Deep Learning, Multimodal Learning, Multilingual Learning, Information Retrieval, Automated Fact-checking

ACM Reference Format:
Stefanos-Iordanis Papadopoulos, Ivana Beňová, Sebastian Kula, Michal Gregor, George Karantaidis, Tomáš Javůrek, Marián Šimko, and Symeon Papadopoulos. 2025. Multimodal and Multilingual Fact-Checked Article Retrieval. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25), June 30-July 3, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3731715.3733402

1 Introduction

The online spread of misinformation has become a serious societal issue, with consequences such as undermining public health [10, 27], marginalizing vulnerable groups [12, 28], and threatening democratic processes [3, 5]. Organized fact-checking initiatives have emerged to combat misinformation, but the process remains time-consuming and labor-intensive, often involving tedious retrieval tasks. In an effort to assist human fact-checkers, researchers are increasingly trying to automate aspects of the fact-checking process, including claim detection, evidence collection, verdict prediction, and explanation generation [14]. In this context, Fact-Check Retrieval (FCR) has also become a vital tool in combating misinformation by identifying verified articles from reputable fact-checking organizations to address specific claims, debunk falsehoods, and promote media literacy through credible, evidence-based sources.

While recent research has made significant strides in developing datasets and retrieval systems for FCR [2, 6], existing efforts have largely focused on one of three areas: (1) text-based FCR with English-only claims and articles [15], (2) multilingual claims and

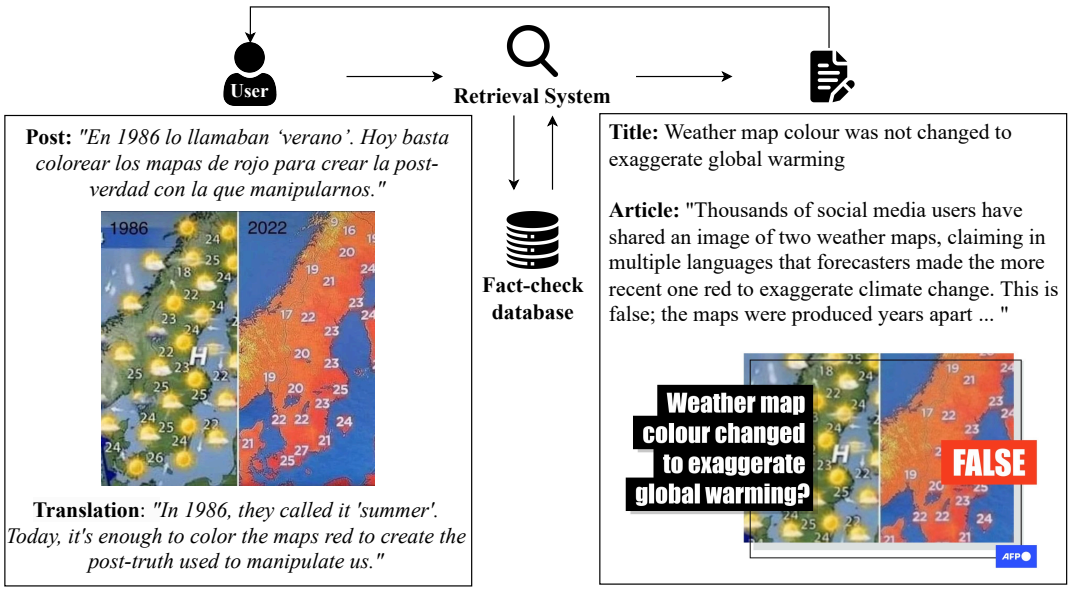

Figure 1: A high-level overview of the Fact-Check Retrieval (FCR). A user's post is processed by the retrieval system, which searches a fact-check database for relevant fact-checked articles. In this example, both multilinguality and multimodality are essential for accurately linking the user post to the relevant fact-checked article.

articles [24], or (3) multimodal setups combining English text and images [34]. Yet, there is a critical lack of a large-scale dataset for FCR that is both multilingual and multimodal. This gap is especially urgent given the growth of misinformation, which increasingly transcends linguistic and geographic boundaries [7] and spans multiple modalities such as text and images [38]. For example, Fig. 1 shows a Spanish social media post claiming that weather maps are intentionally colored red to manipulate perceptions of global warming. This post is linked to an English article[1] debunking the claim, with the shared image acting as a key multimodal connection.

To address these challenges, we present the *M3-Check* dataset, which includes images and diverse multilingual texts, encompassing both social media posts from three major platforms (Twitter/X, Instagram, and Facebook) as well as articles from a wide range of credible sources. By addressing multilinguality and multimodality simultaneously, *M3-Check* enables experiments that more accurately reflect real-world scenarios and serves as a valuable resource to drive future research in the field.

Furthermore, we propose *FACTOR* (FAct-Check TransfORmer), a two-tower Transformer-based architecture that leverages cross-tower parameter sharing and applies modality-wise aligned weight initialization to the linear projection layers. We compare *FACTOR* against strong baselines, including zero-shot retrieval with CLIP features [25], a two-tower linear model, and a 'vanilla' Transformer. Notably, *FACTOR* consistently outperforms other methods, surpassing its 'vanilla' counterpart by 17% in terms of Hits@1. Additionally, we perform an ablation study to assess the contribution of each modality in FCR systems by comparing text-only, image-only, and multimodal approaches, while also incorporating OCR-extracted texts. Finally, we conduct a comparison of state-of-the-art backbone encoders, including CLIP, E5, DINOv2, and multi-E5, and examine the trade-offs between using multilingual texts or English translations. Our analysis emphasizes that every modality (images, texts, OCR-extracted texts) is crucial for optimal retrieval performance, and that leveraging robust multilingual models (i.e., multi-E5) can yield strong performance without the need for English translations.

We make our code and data publicly available at: https://github.com/kinit-sk/M3-Check.

2 Background

2.1 Preliminaries

We define the task of FCR as follows: Let there be a set of social media user posts:

$$\mathcal{P} = \{(T_i^p, I_i^p, O_i^p)\}_{i=1}^{N} \tag{1}$$

where each post consists of some text T_i^p (the full text of the post), image I_i^p, and OCR-extracted text O_i^p from the image. Let there be a set of articles from fact-checking organizations:

$$\mathcal{A} = \{(T_j^a, I_j^a, O_j^a)\}_{j=1}^{M} \tag{2}$$

where each article consists of a text $\mathcal{T}_j^a$ (the fact-checked claim, not the full text of the fact-check), an image $\mathcal{I}_j^a$, and OCR-extracted text O_j^a. Finally, let there be an *annotation mapping*:

$$\mathcal{L} : \mathcal{P} \times \mathcal{A} \rightarrow \{0, 1\}, \tag{3}$$

which provides the ground-truth links between posts $\mathcal{P}$ and their corresponding relevant articles in $\mathcal{A}$. The FCR task is to learn a

[1] https://factcheck.afp.com/doc.afp.com.32C27XA

scoring function:

$$S : \mathcal{P} \times \mathcal{A} \rightarrow \mathbb{R}, \tag{4}$$

such that for each post $p \in \mathcal{P}$, relevant articles $A_{p+} = \{a \in \mathcal{A} : \mathcal{L}(p, a) = 1\}$ are ranked above irrelevant articles $A_{p-} = \mathcal{A} - A_{p+}$:

$$\begin{aligned} S(p_i, a^+) > S(p_i, a^-) \\ \forall a^+, a^- \in A_{p+} \times A_{p-} \end{aligned} \tag{5}$$

2.2 Related Work

In recent years, there has been growing interest in automated fact-checking [14], encompassing tasks such as claim detection [1], evidence filtering and retrieval [8, 39], detection of fake news [31, 37], deepfakes [26], and multimodal misinformation [22], alongside the retrieval of fact-checked articles. As a result, numerous datasets have been developed to support these tasks, including text datasets like LIAR [37] and FEVER [31], multimodal datasets such as Twitter MediaEval [4], Weibo [16], Fakeddit [19], CHASMA and VERITE [23], and multi-task datasets like MOCHEG [39] and Factify [18], which address a range of tasks, including evidence retrieval, verdict classification and explanation generation.

Fact-Check Retrieval (FCR), sometimes referred to as verified claim retrieval [2], or claim matching [17], is a related but distinct task that focuses on linking new claims or social media posts to corresponding fact-check articles, with the goal of examining the veracity of a claim by comparing it to existing verified information. Recent studies on FCR have focused on either text-based retrieval, whether monolingual or multilingual, or multimodal retrieval incorporating additional data types such as images.

2.3 Textual FCR datasets

CrowdChecked is a large-scale FCR dataset comprising 10,340 articles and 316,564 social media posts, created by searching for fact-check URLs on Twitter and collecting English tweets from the corresponding threads. To address the inherent noise in the process, the authors propose various noise filtering techniques [15]. The CheckThat! Dataset, used in Task 2 of the CheckThat! Challenge on retrieving previously fact-checked claims [2], focuses on English and Arabic tweets referenced in fact-checks. It includes 2,259 tweets, 44,164 articles, and 2,440 links. Kazemi et al. [17] collected over 2 million messages from public chat groups in English, Bengali, Hindi, Malayalam, and Tamil, along with around 150,000 fact-checks and selected 2,343 messages which were annotated for their claim similarity, of which 258 were linked to fact-check articles. More recently, the multilingual MultiClaim dataset was proposed for FCR [24], which contains roughly 28,092 social media posts in 27 languages from three social media platforms, 205,751 fact-checks in 39 languages written by professional fact-checkers, and 31,305 links between the two, making it the largest and most linguistically diverse dataset of its kind.

2.4 Multimodal FCR datasets

Vo and Lee [34] used existing datasets from Snopes and PolitiFact, manually annotating them to link user posts with corresponding fact-checks, with the Snopes dataset containing 11,202 positive pairs from 11,167 tweets and 1,703 fact-check articles, and the PolitiFact dataset containing 2,037 positive pairs from 2,026 tweets and 467 fact-check articles, both of which also include images, and comprise English texts. The MuMiN dataset is a large-scale resource containing millions of user posts from Twitter/X, which are linked to 10,920 fact-check articles [20]. However, there are only 6,573 images linked to user posts, while the fact-check articles themselves do not contain any visual content. Although the dataset could theoretically be used for FCR, it is structured as a heterogeneous graph and centers on two graph classification tasks: Claim Classification, predicting whether a claim and its surrounding subgraph indicate misinformation or factuality, and Tweet Classification, determining whether a source tweet and its associated subgraph pertain to a claim whose verdict is misinformation or factual.

Table 1: Overview of publicly available datasets for Fact-Check Retrieval.

Dataset	Posts	Articles	Multimodal	Multilingual
CrowdChecked [15]	316,564	10,340	–	–
MultiClaim [24]	28,092	18,107	–	✓
Politifact [34]	2,026	467	✓	–
Snopes [34]	11,167	1,703	✓	–
M3-Check (Ours)	28,092	18,107	✓	✓

2.5 Limitations in existing FCR datasets

Based on the above, and as shown in Table 1, there is a significant gap in large-scale multilingual and multimodal datasets for FCR. Existing datasets fail to integrate both images and text from social media posts and fact-check articles and lack coverage of multiple major platforms such as Instagram, Facebook, and Twitter/X.

3 Dataset Description

M3-Check was created to support the task of FCR, comprising social media posts from Twitter/X, Facebook, and Instagram linked to articles from credible fact-checking organizations. Both the posts and articles are multimodal, containing text and images, and the texts are in multiple languages. *M3-Check* expands upon the MultiClaim dataset [24] by introducing multimodality in the form of images for both posts and articles. Additionally, it also incorporates full fact-checked articles, English translations and OCR-processed images, enhancing its linguistic and visual diversity. Translations from the original languages to English were performed using the Googletrans Python library[2].

3.1 Data Collection

The development of *M3-Check* was done through the aggregation of data from multiple diverse sources, divided into two distinct categories: a) social media platforms (Facebook, Instagram, and Twitter/X) and b) articles (textual data) and images sourced from the websites of fact-checking organizations. In total 74 different sources (web domains) linked with 41 different fact-check organizations were sampled, with a substantial proportion of data originating from the French Press Agency (AFP) fact check website [3].

In the data collection process, we tried to maintain linguistic neutrality, refraining from favoring any particular language. Our

[2]https://pypi.org/project/googletrans/
[3]https://factcheck.afp.com/

objective was to maximize the linguistic diversity of the dataset, thereby we employed multilingual sources spanning a wide range of languages. However, due to the fact that English is a high-resource language in both social media and fact check websites, its somewhat over-represented (28 fact check sources). In addition to English, sources in low-resource languages, including Polish, Slovak, and Greek, were also considered. The dataset comprises sources in 28 distinct languages. Beyond English (28 sources), other well-represented languages include Spanish (7), Portuguese (5), French (4), German (4), Hindi (3), Chinese (2), Polish (2), Indonesian (2), Arabic (2) and Bengali (2), while the rest of languages are represented by a single source.

3.2 OCR Text Extraction

During the image data collection process, it has been observed that images from social media posts and images collected from fact check articles often feature overlay text in a variety of forms: comments placed directly on the image referring to its content, screenshots showing written reactions to a given image on social media, fact-checking explanations, banner text at public gatherings, and others. It was determined by images inspection that the text contained in the images would be useful in the FCR task; therefore all images containing text were processed by OCR.

Text can appear in different areas of an image and there is no clear common pattern that would apply to all images. Text can be of different sizes and with different font types at the same time, which meant that it was necessary to employ a robust multilingual OCR approach. We assessed EasyOCR[4], Gemini (1.5)[5], and GPT-4o[6], finding that the GPT-4o model delivers the highest quality and accuracy. For OCR text extraction with GPT-4, we use the following prompt:

Prompt applied for OCR with GPT-4o

"You are a professional and friendly OCR, AI based assistant. Extract all the text from the image. Display only extracted text, do not add any explanations or comments. Do not display named entities like usernames, names and surnames of private persons. Do not display information about the number of comments, shares, likes, replies, also do not display URLs, e-mail addresses. Do not display information about technical specification of the image, like image size, image resolution, number of pixels and similar information."

In addition to performing OCR, the GPT-4o model was tasked with text filtering, namely URL addresses, names and usernames of users who are not public figures and technical information about the photos. These elements were excluded as they do not contribute meaningfully to the FCR task. In total, text was extracted using the GPT-4o model for 16,792 images from social media posts and 29,726 images from fact-check articles. Fig.2 provides an example of text extracted from an image using GPT-4 OCR. The extracted text is as follows: *"Dr. Roberto Petrella on COVID-19 You will not be able to travel without a vaccine, you will not be able to go to the cinema, and in the future you won't even be able to leave your own house."*

[4]https://www.jaided.ai/easyocr
[5]https://gemini.google.com
[6]https://openai.com/index/hello-gpt-4o

3.3 Dataset Statistics

M3-Check comprises records, which describe posts, fact-check articles and links between them. The records contain post texts, paths to the images from posts, URLs of fact-check articles, fact-check text, OCRs of images from post, OCRs of images from fact-check articles, OCR and text translations into English if the original language is different than English, fact-check claim and fact-check title. In total 28,909 unique images from posts, and 31,453 from fact-check articles were collected and are available by accessing the URLs of fact-check articles. The dataset includes 28,092 unique posts and 18,107 unique fact-check articles, with 31,305 pairs of posts and fact-check articles. Each pair represents a connection between a single post and a single fact-check article (i.e. the fact-check article addresses the claim made in the post); the relationship between posts and fact-checks is many-to-many in its nature.

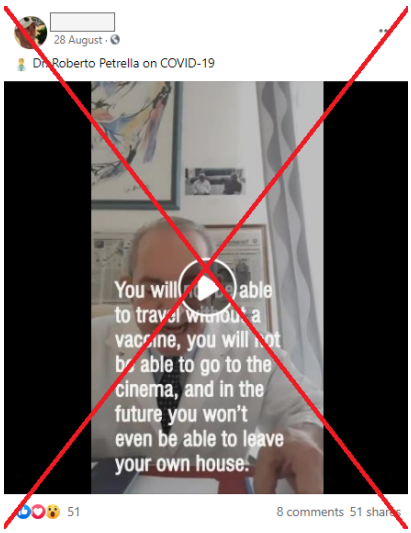

Figure 2: Example of an image from which text is extracted using OCR with GPT-4.

The dataset was divided into three subsets for the purpose of experimentation: a training subset (70% of data), a validation subset (15%) and a test subset (15%). This partitioning scheme enables the evaluation of model performance on unseen data.

4 Methods

4.1 Pre-Trained Encoders

To improve robustness and generalization capabilities of our models, we start by extracting features from each item using a pre-trained encoder; we experimented with the following alternatives:

- CLIP[7] [25]: A vision-language model that (independently) embeds texts and images into a joint embedding space, and which therefore seems particularly well-suited for this task;

[7]HuggingFace ID: `openai/clip-vit-large-patch14`.

- E5[8] [35] and multilingual-E5 [36]: A well-performing English and multi-lingual text embedding model;
- DINOv2[9] [21]: a popular embedding model for vision, trained through self-supervision.

4.2 Model Architecture

To produce the final embeddings, features extracted using pre-trained encoders are passed through a dual-encoder architecture – i.e. through a *post encoder*, which aggregates the features from a user post's image, text, and OCR:

$$z_p = f_p(\phi_v(I^p), \phi_l(T^p), \phi_l(O^p)), \tag{6}$$

and an *article encoder*, which does the same for the fact-check article:

$$z_a = f_a(\phi_v(I^a), \phi_l(T^a), \phi_l(O^a)). \tag{7}$$

For the post encoder f_p and the article encoder f_a, we explore a simple linear projection baseline and a transformer model.

4.2.1 Linear Projection Baseline. With the linear projection baseline, we concatenate the features provided by the pre-trained encoders to get:

$$\begin{aligned} c_p &= \phi_v(I^p) \oplus \phi_l(T^p) \oplus \phi_l(O^p) \\ c_a &= \phi_v(I^a) \oplus \phi_l(T^a) \oplus \phi_l(O^a), \end{aligned} \tag{8}$$

where $\oplus$ denotes the concatenation operator. We then apply a single linear layer to each:

$$\begin{aligned} z_p &= c_p W_p^\top + b_p \\ z_a &= c_a W_a^\top + b_a \end{aligned} \tag{9}$$

Embeddings of all our modalities have the same dimensionality $e_{\text{dim}} = 768$, which we also use for our joint embedding, therefore $W_p, W_a \in \mathbb{R}^{e_{\text{dim}} \times 3e_{\text{dim}}}$.

4.2.2 Fact-Check Transformer. We leverage the Transformer instance used by the Llama family of models ([11, 32, 33]), leveraging: (i) rotary positional embeddings [30]; (ii) the SwiGLU activation function [29]; (iii) RMSnorm [40] to normalize the input of each transformer sub-layer. Since our model is not generative, unlike the Llama family, we do not use causal masking.

We first apply a linear projection to each modality, to produce the embeddings z_p, z_a:

$$c_{d,M} = \phi_v(M^d) W_{d,M}^\top + b_{d,M}, \tag{10}$$

where $d \in \{p, a\}$ represents the type of data (posts/articles) and $M \in \{T, I, O\}$ represents the modality (text, image, OCR).

Modality-wise Aligned Initialization: To ensure consistency across modality projection layers, we apply modality-wise aligned weight initialization, where the weights of projection layers per modality are initialized with the same weights and biases at the start of training, as follows:

$$\begin{aligned} W_{d,O}^{init}, W_{d,I}^{init} &\leftarrow W_{d,T}^{init} \\ b_{d,O}^{init}, b_{d,I}^{init} &\leftarrow b_{d,T}^{init} \end{aligned} \tag{11}$$

This approach helps prevent misalignment of embeddings at the start of training.

[8] `intfloat/e5-base`, `intfloat/multilingual-e5-base`.
[9] `facebook/dinov2-base`.

The projected modalities are then concatenated into sequences and fed into their respective transformers f_p^{tr}, f_a^{tr} for posts and articles respectively, as follows:

$$\begin{aligned} z_p &= f_p^{tr}([c_{p,T} \oplus c_{p,I} \oplus c_{p,O}])[0] \\ z_a &= f_a^{tr}([c_{a,T} \oplus c_{a,I} \oplus c_{a,O}])[0], \end{aligned} \tag{12}$$

where $[0]$ denotes indexing the representation of the first token in the sequence. When a post or article lacks text, image, or OCR data, we replace its embedding with an all-zero vector.

Cross-Tower Parameter Sharing: We reason that the transformer's expressiveness allows for effective cross-tower parameter sharing. Therefore, we also explore the possibility of setting $f_a^{tr} = f_p^{tr}$, encouraging the model to learn a unified representation space that captures common semantic structures and relationships between user posts and fact-check articles. Nevertheless, in the ablation study, we also explore the more conventional approach with two separate transformer encoders ($f_p^{tr} \neq f_a^{tr}$).

Orthogonal Initialization. Finally, we opted for orthogonal weight initialization for all linear layers. This choice is motivated by the intuition that, particularly in modality-specific projection layers, an orthogonal transformation—including rotation—can help preserve the structure of embeddings, potentially facilitating their alignment across modalities. A related idea appears in cross-lingual word embedding alignment (e.g., [9]), where learned orthogonal mappings, such as rotations, help align embeddings from different languages while preserving their geometric relationships. In contrast, general linear projections can limit the expressive capacity of the embeddings, potentially obscuring some of the information they contain.

4.3 Similarity Function

Let $z_p^{(i)}$ and $z_a^{(j)}$ denote the embeddings of the i-th post and the j-th article, respectively, produced by encoders f_p and f_a. The similarity between the post and the article is computed by the inner product:

$$\text{sim}(z_p^{(i)}, z_a^{(j)}) = z_p^{(i)} \cdot z_a^{(j)}. \tag{13}$$

4.4 Loss Function

During training, we employ cross-entropy loss and batches of positive pairs are drawn from the dataset, where each pair consists of a post and one of its linked articles [25]. In our proposed approach, we utilize the annotation mapping $\mathcal{L}$ to compute the actual label matrix rather than assuming a unit matrix. When a post is linked to multiple articles in the batch (as per $\mathcal{L}$), all linked articles are treated as positive examples.

5 Experimental Setup

5.1 Evaluation Protocol

We evaluate the ranking induced by S using Hits@k ($k \in \{1, 5, 10\}$), which measures the proportion of cases where at least one $a \in A_{p+}$ appears among the top-k retrieved articles given post p. During evaluation, we consider all M articles as candidates, pooling them from the entire train-validation-test split, as excluding articles at this stage would artificially simplify the retrieval task, leading to an overestimation of overall performance.

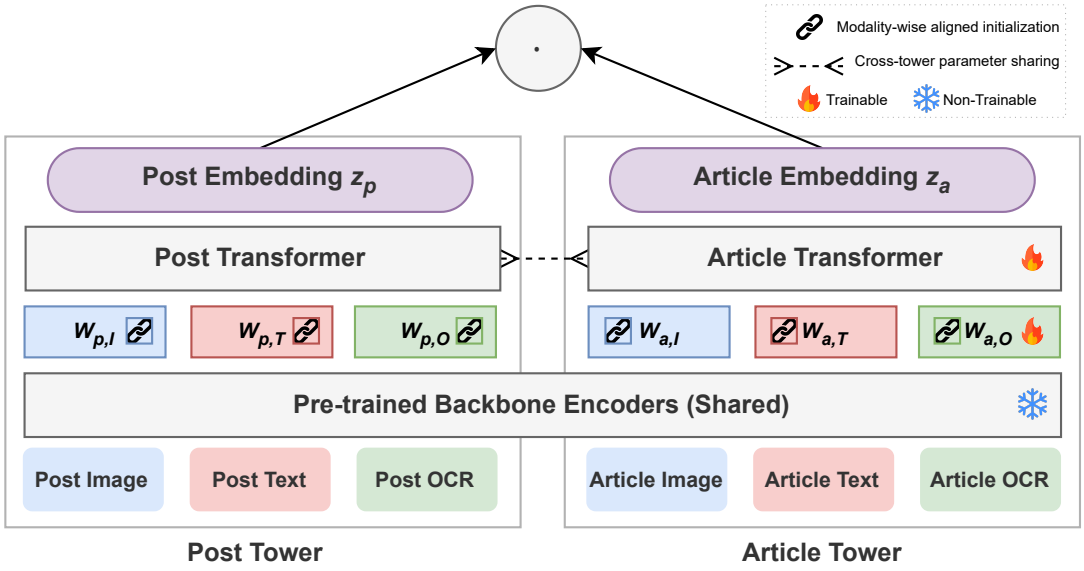

Figure 3: High-level overview of *FACTOR*, a two-tower architecture that leverages pre-trained backbone encoders (e.g., CLIP, E5, or multi-E5), linear projection layers with modality-wise aligned weight initialization, and cross-tower parameter sharing via a LLaMa-based Transformer encoder.

Note that each post p_i may have multiple instances due to the sharing and copying of social media posts. While all instances share the same text T_i^p, they are often associated with different sets of images. To simplify our experiments, we consider only the first instance of each post and select its first image as I_i^p. For OCR-extracted text, we concatenate the OCR output from all images of the first instance.

5.2 Implementation Details

Each experiment runs for up to 500 epochs, with early stopping based on the Hits@1 metric. Patience is set to 150 to avoid premature termination. At the end of training, the model is restored to the weights achieving the highest Hits@1. Experiments are conducted with 5 different seeds, i.e., $0, \ldots, 4$. We also evaluate optimization-free baselines, using raw embeddings from modality encoders or their concatenations, which are deterministic and computed only once. Statistical significance is assessed using the Kruskal-Wallis H test ($p < 0.01$). Uncertainty, including error bars and table values, is reported as ± standard deviation unless stated otherwise.

For training, we employ the AdamW optimizer with a learning rate of $5 \cdot 10^{-5}$, keeping all other parameters at default values. The batch size is set to 512. Regarding *FACTOR*, we use a compact configuration of the Transformer architecture with $d_{\text{model}} = e_{\text{dim}}$, an intermediate size of 64, 4 attention heads, 3 hidden layers, a maximum of 32 positional embeddings, and RoPE $\theta = 10^4$. Various configurations were explored during the design phase, and this setup was chosen based on the best validation performance.

6 Results

In the first set of experiments, we evaluate *FACTOR* against the linear projection baseline and several non-trainable baselines that rely on raw embeddings from pre-trained encoders for each modality. For all experiments, we use CLIP embeddings, applying cosine similarity for optimization-free baselines (as done in [21, 25, 35]) instead of the inner product for embedding comparisons. As shown in Fig. 4, the linear projection baseline significantly outperforms the optimization-free approaches and FACTOR further enhances performance, exceeding the linear baseline by around 12% in terms of Hits@1. Importantly, *FACTOR* provides significant performance gains at only minimal computational overhead compared to the linear projection baseline. As shown in Table 2, inference time, measured by iterating over the training set 10 times and computing the per-epoch duration with `torch.compile`, shows no statistically significant difference. During training, the per-epoch overhead is approximately 32%, which reduces to around 25% when factoring in early stopping.

6.1 Modality Ablations

Given the use of multiple modalities, it is crucial to assess the impact of each modality on the model's overall performance. In Table 3, we present several modality ablations, where specific modalities are masked (replacing their embeddings with all-zeroes during training and inference). We opted not to exclude the corresponding "modality tokens" entirely from the transformer, as doing so could reduce its representational capacity and potentially affect performance.

Ablating individual modalities (e.g., `[-img]`, `[-text]`, `[-ocr]`) reveals that each modality significantly contributes to performance.

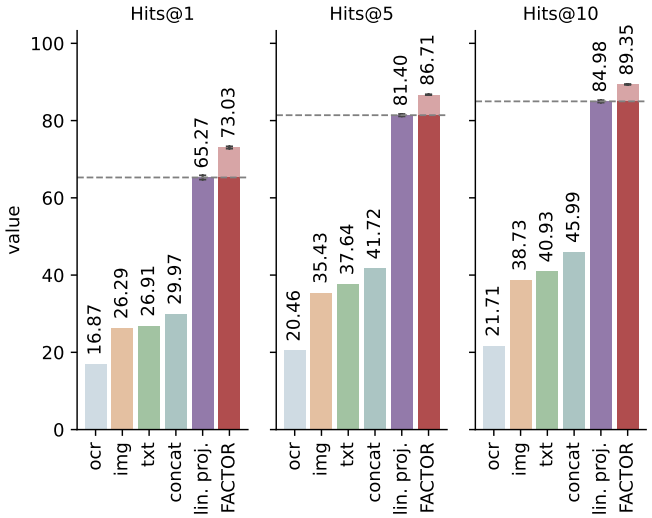

Figure 4: Comparison of 1) optimization-free baselines using raw embeddings for image, text, and OCR, as well as their concatenation, 2) the linear projection model, and 3) the proposed *FACTOR*. CLIP embeddings were used for all experiments.

Table 2: Training and inference times for linear projection and *FACTOR*, with overhead showing *FACTOR*'s increase over linear projection and statistical significance at $p < 0.01$. Inference time is averaged over 5 runs, each iterating 10 times, with per-epoch time computed. `torch.compile()` is used for inference, with one batch run before measurements.

	training		inference
	total time	per epoch	per epoch
lin. proj.	1364.64 ±8.57	2.72 ±0.02	2.28 ±0.02
FACTOR	1703.94 ±162.82	3.59 ±0.02	2.27 ±0.02
overhead	+24.86%	+31.95%	+0.23%
significant	yes	yes	no

Specifically, text is the most impactful, with a 43% drop in performance when removed, followed by images and OCR, with images resulting in a performance drop of over 13% and OCR of almost 7% in terms of Hits@1. In regards to unimodal performance, the image-only model (39.44%) is followed by the text-only model (35.18%), and then by the OCR-only model (21.28%).

6.2 Methodological Ablations

In the methodology section, we introduced several design choices that distinguish *FACTOR* from the vanilla Transformer. In the following experiments, presented in Table 4, we perform ablations to assess their impact on performance. We find that *FACTOR* outperforms the vanilla Transformer by 17% in Hits@1. Moreover, we observe that removing cross-tower parameter sharing (using two separate transformers F_a^{tr} and F_p^{tr}) causes the most significant performance drop (-8.15%), followed closely by removing modality-wise aligned weight initialization (-8.07%) and by the removal of orthogonal initialization (-3.23%). Overall, the ablation study highlights the critical role of each component in driving *FACTOR*'s performance.

Table 3: Results of *FACTOR* leveraging CLIP embeddings with different modality ablations, compared against the base version (*FACTOR*`[all]`); `[-mod]` indicates that modality `mod` is ablated; `[mod]` indicates that only modality `mod` is used.

	Hits@1	Hits@5	Hits@10
FACTOR[all]	**73.03 ± 0.29**	**86.71 ± 0.09**	**89.35 ± 0.07**
FACTOR[-img]	63.17 ± 0.57	76.83 ± 0.41	80.34 ± 0.28
FACTOR[-txt]	41.26 ± 0.53	54.59 ± 0.42	57.53 ± 0.43
FACTOR[-ocr]	68.19 ± 0.59	83.46 ± 0.25	86.45 ± 0.31
FACTOR[img]	39.44 ± 0.25	52.30 ± 0.15	55.18 ± 0.08
FACTOR[txt]	35.18 ± 0.38	47.74 ± 0.43	51.71 ± 0.38
FACTOR[ocr]	21.28 ± 0.22	26.31 ± 0.28	27.53 ± 0.25

Table 4: Comparison of *FACTOR* with its ablated versions: without modality-wise aligned initialization, without orthogonal initialization, without the cross-tower parameter sharing ($F_a^{tr} \neq F_p^{tr}$) and without any additional components (denoted as 'vanilla transformer').

	Hits@1	Hits@5	Hits@10
FACTOR (ours)	**73.03 ± 0.29**	**86.71 ± 0.09**	**89.35 ± 0.07**
w/o modality-wise init	67.14 ± 0.16	82.63 ± 0.35	86.06 ± 0.25
w/o orthogonal init.	70.75 ± 0.32	85.15 ± 0.23	88.15 ± 0.23
w/o cross-tower	67.08 ± 0.31	82.71 ± 0.33	86.13 ± 0.18
vanilla transformer	62.39 ± 0.38	79.72 ± 0.13	83.85 ± 0.33

Table 5: Performance of different encoders: CLIP, E5, multilingual-E5 (with CLIP for image encoding), and DINOv2 (with CLIP for text encoding), comparing text encoders (CLIP, E5, multilingual-E5) with and without machine translation of non-English texts.

	Hits@1	Hits@5	Hits@10
CLIP	73.03 ± 0.29	86.71 ± 0.09	89.35 ± 0.07
E5	69.87 ± 0.28	83.70 ± 0.28	86.82 ± 0.35
DINOv2	69.13 ± 0.35	83.74 ± 0.21	86.82 ± 0.12
multi-E5	**82.55 ± 0.40**	**93.74 ± 0.14**	**95.40 ± 0.19**
CLIP+translation	70.61 ± 0.51	89.14 ± 0.19	92.34 ± 0.27
E5+translation	67.76 ± 0.38	85.97 ± 0.34	89.70 ± 0.16
multi-E5+translation	72.33 ± 0.51	89.72 ± 0.31	92.86 ± 0.24

6.3 Comparing Backbone Encoders

Furthermore, we compare the performance of *FACTOR* with various backbone encoders: CLIP, E5, and DINOv2. As shown in Table 5, neither the E5 model nor DINOv2 provides an improvement over CLIP. However, given the multilingual nature of *M3-Check*, we then evaluated the multilingual version of the E5 model, which resulted in a significant relative improvement of around 13%. This highlights the critical role of multilinguality in enhancing performance, with *multilingual FACTOR* emerging as our best-performing model.

6.4 Using Machine Translation

As an additional analysis, we examined whether machine-translating non-English texts into English could match or surpass the performance of multilingual FACTOR. The results of this comparison are shown in Table 5. Regarding Hits@5 and Hits@10, both CLIP and E5 demonstrate a clear improvement when texts are translated to English, as expected. However, neither model outperforms multilingual FACTOR, and for multilingual-E5, translation actually leads to a decline in performance.

Interestingly, when evaluating Hits@1, translation consistently hinders performance. This is likely due to same-language bias: given how *M3-Check* was constructed, not all connections between matching posts and articles are actually annotated by post-article links and the links are much more likely to exist between posts and articles in the same language. As a result, when using original languages, the model is more likely to retrieve a relevant same-language article in the top-1. However, with English translations, the top-1 result is often a relevant article in a different language, which is less likely to be annotated. As expected, this same-language bias has less impact on Hits@5 and Hits@10, which are less prone to this issue.

6.5 Inference Examples

Fig.5 illustrates two inference examples, taken from the test set of *M3-Check*, by *multilingual FACTOR*, where given a social media post, the model successfully retrieves the correct fact-checked article. The top example, taken from AFP[10], presents a relatively straightforward case where the user post's textual content closely matches the article title, making it sufficient for retrieval while OCR and image information play a minimal role. In contrast, the bottom example[11] presents a more challenging scenario, as the user post contains no textual information apart from a "surprise" emoji. Here, the OCR-extracted text becomes crucial, and a multilingual backbone encoder is necessary to handle both English and non-English text. Additionally, while the post image is not identical to the article image, it shares some visual similarities, which can further aid retrieval. These examples highlight how text, images, and OCR-extracted text each play a vital role in different situations in order to retrieve the correct fact-checked article.

Figure 5: Inference examples: Given a social media post, *FACTOR* retrieves the most relevant fact-checked article; correctly matching the ground truth.

7 Conclusion

In this paper, we presented *M3-Check*, a dataset designed to advance the FCR task by integrating multilingual textual data, multimodal media, and OCR-extracted texts in order to facilitate more robust research in FCR and automated fact-checking. Moreover, we introduced *FACTOR*, a two-tower Transformer-based architecture with post-article and modality-wise weight initializations. To assess its effectiveness, we conducted extensive evaluations, comparing it against strong baselines, including linear models, vanilla Transformers, and state-of-the-art encoders such as CLIP, E5, DINOv2, and multi-E5. Our ablation studies and comparative experiments highlighted the crucial role of multilinguality and multimodality in enhancing FCR systems. These findings pave the way for more effective and scalable automated fact-checking solutions, capable of operating across diverse linguistic and media-rich environments.

Moreover, *FACTOR* or other methods trained on *M3-Check*, could be seamlessly integrated into fact-checking pipelines by retrieving relevant fact-checked articles, helping determine whether a claim has already been verified before consulting more complex misinformation detection models [14]. An additional use of *FACTOR* is as a tool for identifying "leaked evidence", meaning the use of "external information" or "evidence" gathered directly from fact-checking websites and used during the training of misinformation detection systems, which can create an unrealistic setting for early misinformation detection systems [8, 13].

Finally, in regards to future research, while *M3-Check* establishes a strong multilingual and multimodal foundation for FCR research, it currently focuses on images and text, omitting other modalities like audio and video, which are increasingly important vectors of misinformation on social media. Future works should also explore the incorporation of audio and video data to further enhance the capabilities of FCR systems [14].

Acknowledgments

This work is funded by the European Research Executive Agency under Grant No.:101079164 as part of the "DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies" project, and supported by Slovak Research and Development Agency (Slovak APVV) under contract No. APVV-22-0414 Modermed (Multimodal Detection of Toxic Behaviour in Social Media).

[10] https://factcheck.afp.com/covid-19-vaccines-do-not-contain-tracking-devices
[11] https://checamos.afp.com/essa-imagem-nao-e-de-um-episodio-de-os-simpsons-sobre-um-ataque-ao-congresso-dos-eua

References

[1] Alberto Barrón-Cedeño, Firoj Alam, Julia Maria Struß, Preslav Nakov, Tanmoy Chakraborty, Tamer Elsayed, Piotr Przybyła, Tommaso Caselli, Giovanni Da San Martino, Fatima Haouari, et al. 2024. Overview of the CLEF-2024 CheckThat! lab: check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 28–52.

[2] Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. 215–236.

[3] W Lance Bennett and Steven Livingston. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication* 33, 2 (2018), 122–139.

[4] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc Tien Dang Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. 2015. Verifying multimedia use at mediaeval 2015. In *MediaEval 2015*. Vol. 1436. CEUR-WS.

[5] Michele Cantarella, Nicolò Fraccaroli, and Roberto Volpe. 2023. Does fake news affect voting behaviour? *Research Policy* 52, 1 (2023), 104628.

[6] Tanmoy Chakraborty, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. 2023. Information retrieval algorithms and neural ranking models to detect previously fact-checked information. *Neurocomputing* 557 (2023), 126680.

[7] Razieh Chalehchaleh, Reza Farahbakhsh, and Noel Crespi. 2024. Multilingual fake news detection: A study on various models and training scenarios. In *Intelligent Systems Conference*. Springer, 73–89.

[8] Zacharias Chrysidis, Stefanos-Iordanis Papadopoulos, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2024. Credible, Unreliable or Leaked?: Evidence verification for enhanced automated fact-checking. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*. 73–81.

[9] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* (2017).

[10] Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization* 100, 9 (2022), 544.

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[12] José Gamir-Ríos, Raquel Tarullo, Miguel Ibáñez-Cuquerella, et al. 2021. Multimodal disinformation about otherness on the internet. The spread of racist, xenophobic and Islamophobic fake news in 2020. *Anàlisi* (2021), 49–64.

[13] Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 5916–5936.

[14] 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.

[15] 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 (Volume 1: Long Papers)*. 266–285.

[16] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.

[17] Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. Claim Matching Beyond English to Scale Global Fact-Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4504–4517.

[18] Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya N Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P Sheth, Asif Ekbal, et al. 2022. FACTIFY: A Multi-Modal Fact Verification Dataset.. In *DE-FACTIFY@ AAAI*.

[19] Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 6149–6157.

[20] Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 3141–3153.

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[22] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2023. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*. 36–44.

[23] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval* 13, 1 (2024), 4.

[24] Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Mária Bieliková. 2023. Multilingual Previously Fact-Checked Claim Retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 16477–16500.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[26] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. 2022. Deepfake detection: A systematic literature review. *IEEE access* 10 (2022), 25494–25513.

[27] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* 7, 10 (2020), 201199.

[28] Hilda Ruokolainen and Gunilla Widén. 2020. Conceptualising misinformation in the context of asylum seekers. *Information Processing & Management* 57, 3 (2020), 102127.

[29] Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202* (2020).

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.

[31] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 809–819.

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[33] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[34] 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)*. 7717–7731.

[35] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* (2022).

[36] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672* (2024).

[37] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 422–426.

[38] Anna Wilson, Seb Wilkes, Yayoi Teramoto, and Scott Hale. 2023. Multimodal analysis of disinformation and misinformation. *Royal Society Open Science* 10, 12 (2023), 230964.

[39] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2733–2743.

[40] Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems* 32 (2019).