Awakened at CheckThat! 2026: X-IBIS: an Iterative Bi-Encoder Fine-Tuning with Hard-Negative Mining for Cross-Lingual Scientific Source Retrieval

CheckThat! 2026 at CLEF 2026

Maria-Diana Cotelin[1,†], Ciprian-Octavian Truică[1,2,†] and Elena-Simona Apostol[1,*,†]

[1]National University of Science and Technology Politehnica Bucharest, Splaiul Independenței 313, București 060042, Romania
[2]Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania

Abstract

In this paper, we present X-IBIS, an iterative bi-encoder fine-tuning multi-stage cross-lingual retrieval system for CheckThat! 2026 Task 1, which requires retrieving the scientific paper cited by a social media post from a collection of 10,000 English papers. Queries arrive in three languages (English, German, French), creating a vocabulary and register mismatch with formal scientific abstracts. Our pipeline addresses this through: (1) GPT-4o-mini translation of non-English queries, (2) document-side augmentation with synthetic tweet-like summaries to bridge informal and formal register, (3) two-round iterative bi-encoder fine-tuning with hard-negative mining over *multilingual-e5-large*, (4) dense first-stage retrieval, and (5) *BGE-M3* cross-encoder reranking with score interpolation. An extensive ablation study shows that bi-encoder fine-tuning alone yields gains of +51.6% (DE), +50.3% (FR), and +23.9% (EN) in MRR@5 over the zero-shot baseline, with the full pipeline reaching +62.8%, +58.0%, and +34.7%, respectively. Our system achieved an average test MRR@5 of 0.640, placing **10th out of 37 teams** on the official evaluation leaderboard.

Keywords

scientific source retrieval, cross-lingual retrieval, bi-encoder fine-tuning, hard-negative mining, cross-encoder reranking

1. Introduction

Fact-checking becomes more and more important due to the rapid growth of misinformation [1, 2, 3]. A key task involved in the fact-checking process is called *source retrieval*: given a claim, find the scientific papers that can be used to confirm or deny it.

Aligned with this real-world need, CheckThat! 2026 Task 1 [4, 5, 6, 7] requires solving the cross-lingual source retrieval problem, where queries are represented by tweets (social media posts) written informally in English, German, or French, and the scientific papers are a collection of 10,000 documents. The difference between queries and abstracts of scientific papers in terms of their language and format poses a non-trivial task for any retrieval model.

To address Task 1 of the CheckThat! competition, this paper presents *X-IBIS*, an iterative bi-encoder fine-tuning with hard-negative mining for cross-lingual scientific source retrieval. To tackle the cross-lingual problem, our solution includes the following components:

- GPT-4o-mini-based translations of tweets from German and French to English;
- Document side augmentation with synthetic tweet-like summaries;
- Bi-encoder fine-tuning through iterative hard negative sampling from *BGE-M3* with score interpolation.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
maria_diana.cotelin@stud.acs.upb.ro (M. Cotelin); ciprian.truica@upb.ro (C. Truică); elena.apostol@upb.ro (E. Apostol)
https://sites.google.com/view/ciprian-octavian-truica (C. Truică); https://sites.google.com/view/elena-simona-apostol (E. Apostol)
0009-0006-6414-9834 (M. Cotelin); 0000-0001-7292-4462 (C. Truică); 0000-0001-6397-4951 (E. Apostol)

We perform an extensive ablation study across four configurations, revealing how each component helps to increase MRR@5. Our final model placed **10th** among 37 teams participating in the official leaderboard, yielding an average MRR@5 score of 0.640 across all three language splits.

The rest of this paper is organized as follows. Section 2 reviews the related work on this task.Section 3 describes our methodology. Section 4 presents and analyzes the experimental results. Lastly, Section 5 concludes our paper.

2. Related Work

With the spread of misinformation and disinformation on online platforms, the detection and mitigation of such content has attracted the interest of practitioners and researchers. For misinformation and disinformation detection, methods range from word embeddings [8, 9, 10], to transformer-based embeddings [11, 12, 1, 13, 14, 15], a mixture of transforms [12, 16, 17], and Large Language Models [18], while the evaluation of the models use Accuracy, Precision and Recall [19]. In the literature, immunization strategies [20, 21, 22] as well as full detection and mitigation systems have also been proposed [23, 24, 25].

Since the beginning of the CheckThat! competition, multiple methods have been employed for various misinformation detection and fact-checking subtasks. In 2018, the highest scores were achieved using a Multilayer Perceptron with a feature-rich representation [26] and Support Vector Machines [27]. In 2019, the approaches shifted toward supervised text classification models such as Gradient Boosting and Random Forest [28], as well as token-level BERT embeddings for ranking pages [29]. Variants of BERT dominated in 2020, the highest score being achieved by a pre-trained and fine-tuned version of RoBERTa[30], followed by a BERT model with cascade fine-tuning[31]. A fine-tuned pre-trained sentence-BERT combined with TF-IDF is used as well in 2021 [32] and shown to be effective. More recently, the competition has expanded to cover increasingly complex tasks along the fact-checking pipeline, including source retrieval, numerical claim verification, and full fact-checking article generation [4].

2.1. Scientific Source Retrieval

Task 1 of this year's edition is a follow-up to Task 4b from the 2025 edition, which posed an analogous scientific source retrieval challenge. Top-performing systems at that shared task established several key findings that inspired our approach. The second place [33] demonstrates that hard negative mining via BM25[34], where retrieved but incorrect papers serve as negatives, yielded the greatest improvements. They further showed that cross-encoder reranking over the top-10 candidates produced a +10.6% relative gain, and that extending the reranking pool beyond the top-10 degraded performance due to noise. Another approach [35] uses BM25 with dense hybrid retrieval via Reciprocal Rank Fusion [36], finding that neither sparse nor dense retrieval alone was sufficient. Query reformulation strategies [37] find that appending a reformulated version of the query (rather than replacing it) improves retrieval performance. They also compare seven cross-encoder rerankers, consistently finding MS MARCO-trained models outperforming GooAQ-trained variants[38] [39].

2.2. Cross-Lingual Retrieval

Cross-lingual information retrieval (CLIR) has been substantially advanced by multilingual dense encoders such as mBERT[40] and multilingual E5[41], which project texts from different languages into a shared embedding space. Nonetheless, lexical retrieval methods like BM25[34] remain effective for exact-match signals (author names, journal names, specific terminology) that dense models may miss.

2.3. Bi-Encoder Fine-Tuning and Reranking

The retrieve-and-rerank paradigm[42] has become standard for dense retrieval: a bi-encoder performs efficient first-stage retrieval over large collections, followed by a cross-encoder that jointly encodes query-document pairs for precise reranking of the top candidates.

Doc2query [43], which augments document representations with generated queries, has been applied to bridge vocabulary gaps between documents and informal query formulations.

3. Methodology

In this section, we present the methodology used to address the Source Retrieval for Scientific Web Claims. X-IBIS's pipeline (Figure 1) consists of five stages:

1. Query translation
2. Document-side augmentation
3. Iterative Bi-encoder Fine-Tuning with hard-negative mining
4. First-Stage Dense Retrieval
5. Cross-encoder Reranking with score interpolation

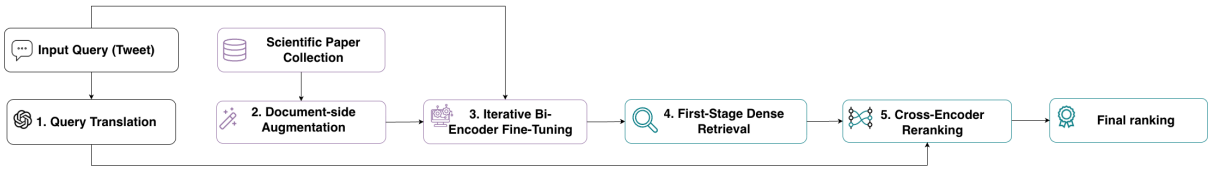

Figure 1: X-IBIS pipeline

3.1. Query Translation

Since the paper collection is in English, German and French queries have little to no BM25 lexical overlap with the documents when used in their original language. Initial experiments confirmed that applying this strategy directly to DE/FR queries degraded the MRR@5. Because of that, we decided to translate all DE and FR queries to English using GPT-4o-mini. We provided the following prompt to the LLM along with the query:

> *"Translate the following [Language] social-media post to English. Preserve scientific terminology, drug/compound names, journal names, study venue names, and author names exactly as they appear. Output only the English translation, nothing else.*
>
> *Text: [Query Text]"*

3.2. Document Side-Augmentation

To resolve the vocabulary gap between social media posts and formal scientific abstracts, each paper from the 10,000 collection is augmented with synthetic tweet-like summaries generated by GPT-4o-mini. For each paper, the model is given the prompt to produce two short claims in the style of a tweet that a user might post when citing the paper:

> *"You write short social-media posts that cite scientific papers. Given a paper, output a JSON object {`"tweets"`: [...]_containing exactly 2 short tweet-like claims (1-2 sentences each) that a real user might post when implicitly referencing this paper. Vary the framing: one may mention the journal/venue, another a key finding, an author, a specific drug/compound/method. Use natural language, not jargon. Output ONLY the JSON object, no markdown."*

An example of the synthetic generations produced by this pipeline is provided in Table 1.

Table 1
Example of synthetic tweet-style expansions generated by `gpt-4o-mini` for document augmentation.

Source	Generated Synthetic Claims
Tweet 1	"A recent study in Pharmaceuticals highlights the antiviral potential of Hypericum perforatum against SARS-CoV-2. Could natural remedies be part of our fight against COVID-19?"
Tweet 2	"Researchers Fakry F. Mohamed and Darisuren Anhlan found that hypericin and pseudohypericin from St. John's Wort show promise in inhibiting SARS-CoV-2. Nature might hold the key!"

The augmented document text used for retrieval has the form:

$$\{\texttt{title}\}\ \{\texttt{venue}\}\ \{\texttt{abstract}\}\ \{\texttt{authors}\}\ ||\ \{\texttt{synthetic\_tweet}_1\}\ \{\texttt{synthetic\_tweet}_2\} \quad (1)$$

This augmentation aims to teach the retrieval model vocabulary that can be found in both tweets and documents datasets. A non-augmented version of each document is reserved for the cross-encoder reranker, where the augmented text would introduce noise into the relevance scoring.

3.3. Iterative Bi-Encoder Fine-Tuning with Hard-Negative Mining

We fine-tuned *intfloat/multilingual-e5-large*[41] (that has 559M parameters and 1024 dimensional embeddings) using *MultipleNegativesRankingLoss*. The model is trained for two rounds of iterative hard-negative mining.

Round 1. The zero-shot mE5-large model encodes the training queries, then for each query, the top 50 retrieved papers that are not the correct answer become candidates for hard negatives. In this step, 5 are sampled per query. The training examples become *query, positive paper, hard-negative paper* triplets. The DE and FR tweets are paired directly with EN papers during training to teach the multilingual encoder to solve the language gap without translations for the dense component.

Round 2. The process repeats, producing harder negatives as the model improves, and providing better gradient signal in the second round. The final fine-tuned bi-encoder re-encodes the full paper collection for first-stage retrieval.

3.4. First-Stage Dense Retrieval

This step uses the fine-tuned bi-encoder with cosine similarity against the augmented collection embeddings. Queries are prefixed with *query* and documents with *passage* following the mE5 asymmetric retrieval format.

3.5. Cross-Encoder Reranking and Score Interpolation

The top 10 candidates from the previous step are reranked using *BGE-M3*[44][45], a multilingual cross-encoder based on XLM-RoBERTa that encodes together query and document pairs. The reranker receives English-translated queries paired with English paper text (without augmentation). Reranking depth is fixed at top 10, following the AIRwaves [33] finding that deeper pools introduce noise and degrade performance beyond top 10.

The final ranking score (Equation (2)) combines the cross-encoder score with the first-stage dense score as a weighted sum after min-max normalisation of both score distributions. For computing the score, we set α to 0.15.

$$final\_score = CE\_score\_norm + \alpha \cdot first\_stage\_score\_norm \quad (2)$$

The weighted α was manually selected based on development-set ablation behavior, rather than obtained from a grid search. Regarding implementation, the score obtained by the cross-encoder model is taken

as the main source of information regarding the relevancy, whereas the dense score obtained from the first stage is only used to make secondary changes among the top-10 candidates reranked using the cross-encoder. Using $\alpha = 0.15$, it was possible to give enough power to the dense retriever to make a decision between two equally relevant cross-encoder matches, while keeping the impact of the dense retriever substantially lower than the normalized cross-encoder score.

An earlier version used a *cross-encoder/mmarco-mMiniLMv2-L12-H384-v1*[46][47] (a smaller mMiniLM-based multilingual model). The cross-encoder reranking is over the top 20 candidates from the first-stage retrieval.

3.6. Additional Methods

3.6.1. BM25 Hybrid Retrieval (Reciprocal Rank Fusion)

BM25 is built over the collection using language-aware tokenization: spaCy lemmatizers for German and French (*de_core_news_sm* and *fr_core_news_sm*) and simple whitespace tokenization for English. We used the default *BM25Okapi* parameters from *rank-bm25*: $k_1 = 1.5$, $b = 0.75$, and $\epsilon = 0.25$. Scores are fused with dense cosine similarity via Reciprocal Rank Fusion (Equation (3)).

$$rrf\_score[i] = \frac{1}{k + dense\_rank[i]} + \frac{1}{k + bm25\_rank[i]} \tag{3}$$

3.6.2. Hypothetical Document Embedding

We also experimented with Hypothetical Document Embedding (HyDE) [48]. For each query tweet, GPT-4o-mini is prompted to generate a hypothetical 256-token paper abstract that the tweet might be citing. The hypothetical abstract is then embedded with *intfloat/multilingual-e5-large*[41] and used as the retrieval query in place of the raw tweet. The intuition is that a full-sentence academic abstract sits closer in embedding space to real paper abstracts than a short informal tweet, but during experiments, HyDE did not outperform simple dense retrieval.

3.6.3. Single-Round Bi-Encoder Fine-Tuning with BM25 Hard Negatives

Another method we employ is to perform a single round of fine-tuning using hard negatives mined from the BM25 + dense hybrid scores (top 5 non-relevant papers per query)

The final system shows improvements in: (1) hard negatives are mined from the dense model alone (after removing BM25), and (2) a second round of mining with the Round 1 checkpoint provided harder, model-specific negatives that gave better gradient signal. The single-round version serves as a useful starting point but underperforms the two-round iterative system.

4. Experimental Results

In this section, we present the experimental results of the proposed architecture for CheckThat! Task 1 Source Retrieval for Scientific Web Claims.

4.1. Dataset

The CheckThat! 2026 Task 1 is a Source Retrieval for Scientific Web Claims task. The dataset for this consists of 10,000 English scientific papers (title, venue, abstract, authors) and social media queries (tweets) from three language splits Table 2: German, French, and English.

4.2. Experimental Setup

All experiments are executed on Google Colab using an NVIDIA A100 GPU (40 GB HBM2). Bi-encoder fine-tuning used mixed-precision (fp16), AdamW optimiser, batch size 32, and a linear warmup over

Table 2
Dataset Split Statistics by Language

Language	Train	Dev	Test
German (DE)	1,460	386	876
French (FR)	2,807	702	1,220
English (EN)	14,977	3,905	6,076

10% of training steps, for 3 epochs per round. The implementation is available online on GitHub at the following url: https://github.com/DS4AI-UPB/Awakened-CLEF2026-CheckThat.

4.3. Result

We evaluate our proposed pipeline using Mean Reciprocal Rank at 5 (MRR@5). Table 6 details the incremental performance gains during our ablation study on the development set.

Experiment 1. Table 3 presents the BM25 hybrid retrieval, HyDE query expansion, and a small mMiniLM cross-encoder, all using the original (not translated) queries. Feeding raw DE/FR tweets into a BM25 index built on English text does not have a good outcome: BM25 RRF reduces DE MRR@5 by 38.5% (0.374 $\rightarrow$ 0.230) and FR MRR@5 by 43.5% (0.455 $\rightarrow$ 0.257) relative to the zero-shot baseline. HyDE expansion partially recovers the cross-lingual signal (DE: +2.6%, FR: +23.5% over BM25 RRF), but remains well below the zero-shot baseline for non-English languages. The mMiniLM cross-encoder further degrades EN performance by 30.0% relative to the baseline, suggesting that a shallow cross-encoder without translation is insufficient for this task.

Table 3
Ablation Summary for Experiment 1

Stage	DE MRR@5	FR MRR@5	EN MRR@5
Zero-shot mE5-large (baseline)	0.374	0.455	0.498
+ BM25 Hybrid RRF	0.230	0.257	0.532
+ HyDE Expansion	0.236	0.317	0.537
+ Cross-Encoder (mMiniLM)	0.231	0.346	0.378

Experiment 2. Table 4 shows the results when using query translation for DE/FR tweets before BM25, upgrading the cross-encoder to *BGE-M3*, and adding single-round bi-encoder fine-tuning with BM25-mined hard negatives. Translation alone recovers the BM25 signal for non-English languages: DE BM25 RRF improves from 0.230 to 0.366 (+59.1% over Experiment 1's BM25), and FR from 0.257 to 0.506 (+96.9%). Adding cross-encoder reranking over the top 10 yields further gains (+16.6% DE, +24.2% FR, +11.4% EN over the zero-shot baseline). The full Experiment 2 pipeline (i.e., fine-tuning plus BM25 hybrid plus reranker) reaches DE: 0.609, FR: 0.719, EN: 0.671, representing improvements of +62.8%, +58.0%, and +34.7% respectively over the zero-shot baseline.

Table 4
Ablation Summary for Experiment 2

Stage	DE MRR@5	FR MRR@5	EN MRR@5
Zero-shot mE5-large (baseline)	0.374	0.455	0.498
+ BM25 RRF	0.366	0.506	0.532
+ Cross-encoder reranking (top 10)	0.436	0.565	0.555
Fine-tuned dense + BM25 RRF	0.413	0.578	0.560
Full pipeline (fine-tune + BM25 + reranker)	0.609	0.719	0.671

Experiment 3. Table 5 introduces *BGE-M3* as an alternative first-stage retriever. BGE-M3 natively

combines dense, sparse (SPLADE), and ColBERT multi-vector scores in a single forward pass. Its hybrid score improves over the zero-shot mE5 baseline by +8.8% DE, +16.5% FR, and +9.4% EN. Ensembling BGE-M3 with mE5-large pushes these gains to +17.6% DE, +23.7% FR, and +15.7% EN, comparable to adding cross-encoder reranking in Experiment 2, suggesting that retriever diversity can partially substitute for a reranker.

Table 5
Ablation Summary for Experiment 3

Stage	DE MRR@5	FR MRR@5	EN MRR@5
Zero-shot mE5-large (baseline)	0.374	0.455	0.498
BGE-M3 Hybrid (dense + sparse + ColBERT)	0.407	0.530	0.545
BGE-M3 + mE5 Ensemble	0.440	0.563	0.576

X-IBIS final pipeline (Table 6) combines doc2query augmentation, two-round iterative hard-negative mining, dense-only first-stage retrieval (BM25 dropped), *BGE-M3* cross-encoder reranking (top 10), and score interpolation. The single largest gain comes from iterative bi-encoder fine-tuning: moving from zero-shot to the fine-tuned first-stage retriever improves MRR@5 by +51.6% (DE: 0.374 → 0.567), +50.3% (FR: 0.455 → 0.684), and +23.9% (EN: 0.498 → 0.617). Cross-encoder reranking then adds a further +5.6% DE, +4.5% FR, and +8.6% EN over the fine-tuned retriever alone. Score interpolation provides a small but consistent final boost (+1.7% DE, +0.6% FR, +0.1% EN), yielding peak development set scores of DE: 0.609, FR: 0.719, EN: 0.671.

Table 6
Ablation Summary Awakened pipeline

Stage	DE MRR@5	FR MRR@5	EN MRR@5
Zero-shot mE5-large (baseline)	0.374	0.455	0.498
Fine-tuned mE5-large (first-stage)	0.567	0.684	0.617
+ Cross-encoder reranking (top 10)	0.599	0.715	0.670
+ Score interpolation	**0.609**	**0.719**	**0.671**

Table 7 compares the best-performing configuration from each experiment. The progression from Experiment 1 (best: HyDE, EN: 0.537) to Experiment 2 (best: full pipeline, EN: 0.671) highlights the outsized impact of query translation and bi-encoder fine-tuning. Experiment 3's BGE-M3 ensemble (EN: 0.576) surpasses Experiment 1 without any reranker, but could not match the fine-tuned pipeline of Experiment 2. The final X-IBIS system outperforms the best alternative approach (Experiment 2 full pipeline on DE: 0.455 → 0.609, +33.8%; FR: 0.601 → 0.719, +19.6%; EN: 0.561 → 0.671, +19.6%) through two-round iterative hard-negative mining and the stronger *BGE-M3* cross-encoder.

Table 7
Experiments Comparison

Stage	Method	DE MRR@5	FR MRR@5	EN MRR@5
Experiment 1	HyDE Expansion	0.236	0.317	0.537
Experiment 2	Fine-tune + BM25 + reranker	0.455	0.601	0.561
Experiment 3	BGE-M3 + mE5 Ensemble	0.440	0.563	0.576
X-IBIS	**CE + score interpolation**	**0.609**	**0.719**	**0.671**

To assess the robustness of our approach, we compare our final development set performance against the official leaderboard test results in Table 8.

The difference between the dev and the test set is stable: average MRR@5 drops by only 1.2% (0.648 → 0.640). German performance is essentially unchanged (+0.7%; 0.609 → 0.613), English declines by

Table 8
Comparison between Dev and Test results for Awakened

Set	Avg. MRR@5	DE MRR@5	FR MRR@5	EN MRR@5
Dev	0.648	0.609	0.719	0.671
Test	0.640	0.613	0.664	0.644

4.0% (0.671 $\rightarrow$ 0.644), and French shows the largest drop of 7.7% (0.719 $\rightarrow$ 0.664), likely reflecting a distributional shift in the French test queries. Overall, our solution secures the **10th place out of 37** participants on the evaluation leaderboard and 10th out of 42 on the development leaderboard, confirming the efficacy of our multi-stage retrieval architecture.

5. Conclusion

We presented *X-IBIS*, a multi-stage cross-lingual scientific source retrieval system for CheckThat! 2026 Task 1. Our pipeline combines GPT-4o-mini query translation, document-side tweet augmentation, two-round iterative bi-encoder fine-tuning with hard-negative mining, dense first-stage retrieval, and *BGE-M3* cross-encoder reranking with score interpolation.

The ablation study demonstrates that query translation and bi-encoder fine-tuning are the most impactful components: fine-tuning alone improves MRR@5 by +51.6% for German, +50.3% for French, and +23.9% for English over the zero-shot baseline. Adding cross-encoder reranking and score interpolation further brings the total gain to +62.8%, +58.0%, and +34.7%, respectively, achieving development set scores of DE: 0.609, FR: 0.719, EN: 0.671. Our system generalizes well to the test set (avg. MRR@5: 0.640), placing 10th out of 37 teams.

Key findings from the ablation study are: (1) lexical retrieval (BM25) is harmful for non-English queries without prior translation; (2) HyDE expansion can partially compensate for language mismatch, but is outperformed by fine-tuning; (3) iterative hard-negative two-round mining provides substantially better gradient signal than single-round fine-tuning; and (4) a small $\alpha = 0.15$ interpolation of the dense first-stage score with the cross-encoder score consistently improves the final ranking.

Future work could explore larger document augmentation budgets, adaptive reranking depth, and fine-tuning the cross-encoder jointly with the bi-encoder to reduce the retrieval-reranking mismatch.

Acknowledgments

The research presented in this paper is supported in part by The Academy of Romanian Scientists, through the funding of the project "NetGuardAI: Intelligent system for harmful content detection and immunization on social networks" (AOȘR-TEAMS-IV).

Declaration on Generative AI

During the preparation of this work, the authors used Claude 4.6 Opus for rephrasing in order to improve clarity and style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C.-O. Truică, E.-S. Apostol, MisRoBÆRTa: Transformers versus Misinformation, Mathematics 10 (2022) 1–25(569). doi:10.3390/math10040569.

[2] C.-O. Truică, E.-S. Apostol, M. Marogel, A. Paschke, GETAE: Graph Information Enhanced Deep Neural NeTwork Ensemble ArchitecturE for fake news detection, Expert Systems with Applications 275 (2025) 126984. doi:10.1016/j.eswa.2025.126984.

[3] C.-O. Truică, E.-S. Apostol, P. Karras, DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection, Knowledge-Based Systems 294 (2024) 111715. doi:10.1016/j.knosys.2024.111715.

[4] J. M. Struß, S. Schellhammer, S. Dietze, V. V., V. Setty, T. Chakraborty, P. Nakov, A. Anand, P. Chungkham, S. Hafid, D. Sahnan, K. Todorov, The clef-2026 checkthat! lab: Advancing multilingual fact-checking, in: R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, Z. Ren, S. Verberne, N. Bai, M. Mansoury (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2026, pp. 325–335.

[5] J. M. Struß, S. Schellhammer, S. Dietze, V. V., V. Setty, T. Chakraborty, P. Nakov, A. Anand, P. Chungkham, S. Hafid, D. Sahnan, K. Todorov, Overview of the CLEF-2026 CheckThat! Lab: Advancing multilingual fact-checking, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, Alba García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[6] E. S. Salido, A. Barrón-Cedeño, Alba García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CLEF 2026, Jena, Germany, 2026.

[7] S. Schellhammer, S. Hafid, Y. S. Kartal, K. Boland, D. Dimitrov, K. Todorov, S. Dietze, Overview of the CLEF-2026 CheckThat! lab task 1 on source retrieval for scientific web claims, in: [6], 2026.

[8] V.-I. Ilie, C.-O. Truică, E.-S. Apostol, A. Paschke, Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings, IEEE Access 9 (2021) 162122–162146. doi:10.1109/ACCESS.2021.3132502.

[9] C.-O. Truică, E.-S. Apostol, It's all in the Embedding! Fake News Detection using Document Embeddings, Mathematics 11 (2023) 1–29(508). doi:10.3390/math11030508.

[10] C.-O. Truică, E.-S. Apostol, T. Ștefu, P. Karras, A Deep Learning Architecture for Audience Interest Prediction of News Topic on Social Media, in: International Conference on Extending Database Technology (EDBT2021), 2021, pp. 588–599. doi:10.5441/002/EDBT.2021.69.

[11] M.-D. Cotelin, E.-S. Apostol, C.-O. Truică, NetGuardAI at EXIST2025: Sexism Detection using mDeBERTa, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2025, pp. 1889–1897.

[12] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for EXIST2024, in: Working Notes of the Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1157–1164.

[13] D.-E. Burghelea, C.-O. Truică, E.-S. Apostol, VERIT-ALBERT: A Finetuned LLM Approach for Verifying Information Credibility, in: The 2025 24th RoEduNet Conference: Networking in Education and Research (RoEduNet), 2025, pp. 1–6. doi:10.1109/RoEduNet68395.2025.11208267.

[14] N. C. J. Le Pollotec, E. S. Apostol, C.-O. Truică, Netguardai at multipride: Multilingual detection of reclaimed language, in: EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26– 27, Bari, IT, 2026, pp. 1–8.

[15] C.-O. Truică, E.-S. Apostol, A. Paschke, Awakened at CheckThat! 2022: Fake News Detection using BiLSTM and sentence transformer, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2022, pp. 749–757.

[16] A. Petrescu, E.-S. Apostol, C.-O. Truică, Awakened at EXIST2025: Adaptive Mixture of Transformers, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2025, pp. 2104–2110.

[17] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for Sexism Identification in Social Networks, in: Conference and Labs of the Evaluation Forum, 2025, pp. 142–155. doi:10.1007/978-3-032-04354-2_10.

[18] Ciprian-Octavian, E.-S. Apostol, A.-G. Ilie, A. Paschke, HarmLLaMA: Harmful Language Detection with Large Language Models, in: International Conference on Intelligent Computer Communica-

tion and Processing (ICCP 2025), 2025, pp. 1–8. doi:10.1109/ICCP68926.2025.11427180.
[19] C.-O. Truică, C. A. Leordeanu, Classification of an imbalanced data set using decision tree algorithms, University Politehnica of Bucharest Scientific Bulletin - Series C Electrical Engineering and Computer Science 79 (2017) 69–84.
[20] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, P. Karras, MCWDST: A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media, IEEE Access 11 (2023) 125861–125873. doi:10.1109/ACCESS.2023.3331220.
[21] A. Petrescu, C.-O. Truică, E.-S. Apostol, P. Karras, Sparse Shield: Social Network Immunization vs. Harmful Speech, in: ACM International Conference on Information and Knowledge Management (CIKM2021), ACM, 2021, pp. 1426–1436. doi:10.1145/3459637.3482481.
[22] E.-S. Apostol, Özgur Coban, C.-O. Truică, CONTAIN: A community-based algorithm for network immunization, Engineering Science and Technology, an International Journal 55 (2024) 1–10(101728). doi:10.1016/j.jestch.2024.101728.
[23] M.-D. Cotelin, C.-O. Truică, E.-S. Apostol, Cleannews: a network-aware fake news mitigation architecture for social media, arXiv preprint arXiv:2509.04489 (2025).
[24] C.-O. Truică, A.-T. Constantinescu, E.-S. Apostol, StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms, in: IEEE International Conference on Intelligent Computer Communication and Processing, 2024, pp. 1–5. doi:10.1109/ICCP63557.2024.10793051.
[25] E.-S. Apostol, C.-O. Truică, A. Paschke, ContCommRTD: A Distributed Content-Based Misinformation-Aware Community Detection System for Real-Time Disaster Reporting, IEEE Transactions on Knowledge and Data Engineering (2024) 1–12. doi:10.1109/tkde.2024.3417232.
[26] C. Zuo, A. Karakas, R. Banerjee, A hybrid recognition system for check-worthy claims using heuristics and supervised learning, in: CEUR workshop proceedings, volume 2125, 2018.
[27] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 267–276.
[28] L. Favano, M. J. Carman, P. Lanzi, et al., Theearthisflat's submission to clef'19checkthat! challenge, in: Working Notes of {CLEF} 2019-Conference and Labs of the EvaluationForum, Lugano, Switzerland, September 9-12, 2019, volume 2380, CEUR-WS. org, 2019, pp. 1–11.
[29] F. Haouari, Z. S. Ali, T. Elsayed, bigir at clef 2019: Automatic verification of arabic claims over the web., in: CLEF (working notes), 2019.
[30] M. Bouziane, H. Perrin, A. Cluzeau, J. Mardas, A. Sadeq, Team buster. ai at checkthat! 2020 insights and recommendations to improve fact-checking., in: CLEF (working notes), 2020, p. 12.
[31] L. Passaro, A. Bondielli, A. Lenci, F. Marcelloni, et al., Unipi-nle at checkthat! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers, in: CEUR Workshop Proceedings, volume 2696, CEUR, 2020.
[32] A. Chernyavskiy, D. Ilvovsky, P. Nakov, Aschern at checkthat! 2021: lambda-calculus of fact-checked claims, Faggioli et al.[12] (2021).
[33] C. Ashbaugh, L. Baumgärtner, T. Gress, N. Sidorov, D. Werner, Airwaves at checkthat! 2025: Retrieving scientific sources for implicit claims on social media with dual encoders and neural re-ranking, 2025. URL: https://arxiv.org/abs/2509.19509. arXiv:2509.19509.
[34] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, volume 4, Now Publishers Inc, 2009.
[35] P. J. Sager, A. Kamaraj, B. F. Grewe, T. Stadelmann, Deep retrieval at checkthat! 2025: identifying scientific papers from implicit social media mentions via hybrid retrieval and re-ranking, arXiv preprint arXiv:2505.23250 (2025).
[36] G. V. Cormack, C. L. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 758–759.
[37] J. Schofield, S. Tian, H. T. T. Truong, M. Heil, Ds@ gt at checkthat! 2025: exploring retrieval and

reranking pipelines for scientific claim source retrieval on social media discourse, arXiv preprint arXiv:2507.06563 (2025).

[38] D. Khashabi, A. Ng, T. Khot, A. Sabharwal, H. Hajishirzi, C. Callison-Burch, Gooaq: Open question answering with diverse answer types, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 421–433.

[39] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019, pp. 3982–3992. URL: https://arxiv.org/abs/1908.10084.

[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[41] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).

[42] R. Nogueira, K. Cho, Passage re-ranking with bert, arXiv preprint arXiv:1901.04085 (2019).

[43] M. Gospodinov, S. MacAvaney, C. Macdonald, Doc2query–: when less is more, in: European Conference on Information Retrieval, Springer, 2023, pp. 414–422.

[44] C. Li, Z. Liu, S. Xiao, Y. Shao, Making large language models a better foundation for dense retrieval, 2023. arXiv:2312.15503.

[45] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[46] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 5776–5788.

[47] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2140–2151. URL: https://aclanthology.org/2021.findings-acl.188. doi:10.18653/v1/2021.findings-acl.188.

[48] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023, pp. 1762–1777. URL: https://aclanthology.org/2023.acl-long.99/. doi:10.18653/v1/2023.acl-long.99.