DisAI.eu

D1.4
Shared task report

Project Title	Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies
Contract Nº.	101079164
Type of Action	HORIZON-CSA
Topic	Disinformation Combating
Project start date	1st Dec 2022
Duration	36 months

Funded by
the European Union

Deliverable title	Shared task report
Deliverable number	D1.4
Deliverable version	1.0
Contractual date of delivery	30 November 2025
Actual date of delivery	28 November 2025
Nature of deliverable	Report
Dissemination level	Public
Work Package	WP1
Task(s)	T1.3
Partner responsible	UCPH
Author(s)	Qiwei Peng (UCPH), Robert Moro (KInIT), Michal Gregor (KInIT), Ivan Srba (KInIT), Simon Ostermann (DFKI)

Abstract	This report presents our SemEval-2025 Task 7 "Multilingual and Crosslingual Fact-Checked Claim Retrieval". Using the MultiClaim dataset and newly collected testing data, the task evaluates systems on previously fact-checked claim retrieval. With 179 participants and 23 submitted system description papers, the results show that contrastively fine-tuned embedding models – often enhanced with hard negatives and multi-stage pipelines – achieve the best performance.
Keywords	Shared Task, SemEval, Claim Matching, Claim Retrieval, Fact-checked claim detection, MultiClaim dataset, Multilingual NLP, Crosslingual NLP

Table of Contents

1 Introduction

The sheer amount of disinformation on the Internet is proving to be a societal problem (Vosoughi et al., 2018). Fact-checking organizations have made progress in manually and professionally fact-checking various contents (Vlachos and Riedel, 2014; Micallef et al., 2022). However, manual fact-checking at scale is too costly. To reduce some of the fact-checkers' manual efforts and make their work more effective, recent studies have examined their needs and identified tasks that could be automated (Nakov et al., 2021; Zeng et al., 2021; Hrckova et al., 2024), such as evidence retrieval (Schuster et al., 2020; Liao et al., 2023) and verdict prediction (Nakashole and Mitchell, 2014; Thorne et al., 2018a).

In this work, we put focus on the task of claim retrieval, also called claim matching (Kazemi et al., 2021), searching for fact-checked information (Vo and Lee, 2020), or fact-checked claim detection (Shaar et al., 2020). Claim retrieval is an NLP task that addresses one significant problem that fact-checkers face: How to find out whether a similar claim has already been fact-checked before, even in a different language (Hrckova et al., 2024). Solving this problem would reduce duplicate work and improve the efficiency of fact-checking efforts. Previously, this task was mostly done in English. Other languages that have been considered include Arabic, Bengali, Hindi, and Tamil (Kazemi et al., 2021; Nakov et al., 2022). However, many other languages or even entire major language families have not been considered at all.

To close this gap, we organized the SemEval 2025[1] Task 7 "Multilingual and Crosslingual FactChecked Claim Retrieval". We formulate the claim retrieval task as an information retrieval task: Based on an input text (we use social media posts), the goal is to find an appropriate claim from a database of claims that have been already fact-checked by professional fact-checkers. Consider the example in Table 1. A user has written a post making a claim worth fact-checking. The idea of claim retrieval is for a model to find a semantically similar claim from a list of previously fact-checked claims. We base the task on our already published multilingual dataset MultiClaim (Pikuliak et al., 2023) that consists of two parts: (1) A list of 206k fact-checked claims, and (2) 28k social media posts (SMPs) with references to relevant fact-checked claims from the list. With this data, we can simulate a situation where a fact-checker is asked to fact-check an SMP, and they want to search through the list of already available fact-checks to see whether the same claim was fact-checked before. Our task features subtracks on both a monolingual and a crosslingual setup. To construct the training and development sets, we use a subset of the MultiClaim data set covering 8 different languages in the monolingual track and 14 different languages (52 combinations) in the crosslingual track. We further collect new resources to construct a test set which covers two additional previously unannounced languages and more than 4,000 new SMP-Claim pairs.

Our shared task was organized along with the prestigious SemEval workshop. The task attracted 179 participants. There were in total 52 test submissions by 31 teams, and 23 teams submitted system description papers. These systems tackling the multilingual and

[1] https://semeval.github.io/SemEval2025/tasks

Table 1. An example from our dataset. The social media post was written by a user. The main text of the post is in italics, the OCR transcript from the attached images is in regular font. The claim comes from the fact-check assigned by a fact-checker.

	Original text (German)	English translation
Social Media Post	Wer an Wahlen glaubt, ist auch der Meinung, dass das Ordnungsamt die Küche putzt.} Wussten Sie eigentlich das Das Besatzerstatut besagt: dass die Fremdverwaltung allein Durch die Wahlen vom Wahler akzeptiert wird ** Bei unter 50% Wahlbeteiligung wäre dies hinfällig Denn es gabe dann ein rechtliches Problem Das Besatzerstatut wäre ungültig...	*Anyone who believes in elections also believes that the regulatory office cleans the kitchen.* Did you actually know that? The occupier statute states: that the foreign administration is only accepted by the voters through the elections ** If the turnout is less than 50%, this would be obsolete Because then there was a legal problem The occupier statute would be invalid...
Claim	Deutschland ist ein besetztes Land.	Germany is an occupied country.

crosslingual claim retrieval task provide valuable insights on applying state-of-the-art techniques to the real-world task and highlight future directions on better solving the claim retrieval problem. Both our shared task paper and the accompanying system papers are published and available online. This deliverable reports on the organization and execution of the shared task. In this report, we describe in detail the setup and implementation of our shared task along with datasets and submitted machine learning systems.

Note that the organization of a shared task was one of the planned activities under the project proposal. Beyond the original plan, where the shared task was envisioned as a standalone activity, with results disseminated mainly through the project website, we managed (as already mentioned) to get the task accepted at the well-known SemEval workshop, which significantly increased its visibility and impact. We may thus conclude that the activity was successfully completed without major deviations from the original plan.

Regarding the scope of the task, we deliberately decided to focus on a multilingual text-only retrieval task rather than a multimodal one. Early consultations with partner institutions indicated that a well-defined multilingual shared task would likely be better positioned to achieve broad participation and would also reduce the logistical and legal risks associated with sharing multimodal data. In particular, distributing images accompanying online content is significantly more prone to ethical and legal complications, as such material may contain copyrighted content, personal data, or context-specific sensitive information beyond what is present in the case of text-only data. Ensuring compliant sharing would either require extensive filtering and rights clearance or, alternatively, relying on participants to redownload the data themselves – which introduces substantial technical complexity (beyond what is practical – we would argue – in a shared task) and would likely reduce participation while also leading to teams working with inconsistent datasets, thus compromising fair evaluation.

Given these constraints, focusing the shared task on multilingual text-only retrieval allowed us to maintain high data quality, ensure legal and ethical compliance, and deliver a task that was accessible to a broad international audience while still being fully aligned with the project's scientific and capacity-building goals. We believe that the acceptance of the task to SemEval 2025 and the high participation demonstrate that this adjustment ultimately strengthened the impact of the activity, enabling wider community engagement and more meaningful scientific contributions than would have been feasible under the originally envisioned multimodal setup.

Information about the shared task was communicated via the DisAI project website[2], while the official results and proceedings were published in the ACL Anthology SemEval-2025 volume[3].

[2] https://disai.eu/semeval-2025
[3] https://aclanthology.org/volumes/2025.semeval-1/

2 Organization

The organization of the shared task started intensively in mid-2024, marking the beginning of a coordinated effort to prepare the dataset, refine evaluation procedures, and establish the technical infrastructure needed for the competition. During this period, the organizing team finalized the overall schedule and defined a four-phase structure intended to guide participants from initial familiarization to final system evaluation. These phases, **trial**, **development**, **test**, and **post-test,** were designed not only to ensure fairness and transparency but also to support participants in building and refining their models in a structured manner.

The **trial phase**, which ran from **1 July 2024 to 2 September 2024**, served as an introductory stage for both organizers and potential participants. For the organizers, it provided an opportunity to thoroughly test the CodaBench competition platform[4], verify the integrity of the evaluation pipeline, and ensure that all components of the platform performed reliably with the trial data. At the same time, this phase marked the official opening of the shared task to the research community. Calls for participation were circulated, inviting teams to register, explore the trial data, and begin familiarizing themselves with the task setup. The feedback collected during this period helped the organizers refine technical guidelines and address potential usability issues before the next phase began.

Following the trial phase, the **development phase** ran from **2 September 2024 to 10 January 2025**. This was the main working period for participants. All teams gained access to the official training and development data splits, enabling them to experiment with model architectures, optimize hyperparameters, and perform internal validation. The lengthy duration of this phase was intended to accommodate a wide range of methodological approaches, from traditional machine-learning pipelines to large neural architectures. Throughout this stage, the organizers monitored participant engagement, provided clarifications on the task formulation, and responded to technical inquiries to maintain a smooth development process.

The **test phase**, running from **10 January 2025 to 1 February 2025**, represented the core evaluation stage of the shared task. Participants were allowed to submit system predictions on the unseen test data up to **five times**, a limit designed to balance flexibility with fairness by preventing excessive trial-and-error testing. Only the **final submission** from each team during this period was considered for the official test leaderboard. This rule encouraged participants to carefully plan their submission strategies and ensured that the final ranking reflected each team's best-considered system rather than exploratory attempts.

After the official evaluation concluded, the competition entered the long-lasting **post-test phase**. Unlike the preceding phases, the post-test period did not contribute to

[4] https://www.codabench.org/competitions/3737/

leaderboard ranking or official outcomes. Instead, it served as a **community-oriented extension** of the shared task. Participants and others, who joined later out of interest, were able to continue experimenting with their systems, submit models for exploratory evaluation, or use the platform as a benchmark for ongoing research. While results submitted during this phase were not considered for the official ranking, they remained valuable for researchers seeking additional insight into model behavior, comparative analysis, or continued methodological development. The post-test phase thus helped maximize the long-term utility and impact of the shared task within the broader research community.

The shared task was officially featured at ACL 2025, held in Vienna from 31 July 2025 to 1 August 2025. During the event, we presented the outcomes of the task, including the final results, key insights from participating systems, and an overview of methodological trends observed throughout the competition.

The shared task was made possible through collaboration between researchers from the University of Copenhagen (UCPH), the Kempelen Institute of Intelligent Technologies (KInIT), and the German Research Center for Artificial Intelligence (DFKI). The organizing team consisted of:

Qiwei Peng (UCPH), Róbert Móro (KInIT), Michal Gregor (KInIT), Ivan Srba (KInIT), Simon Ostermann (DFKI), Marián Šimko (KInIT), Juraj Podroužek (KInIT), Matúš Mesarcík (KInIT), Jaroslav Kopčan (KInIT), Anders Søgaard (UCPH).

3 Task Description

The shared task is formulated as an information retrieval task. Given social media posts, the goal is to retrieve the most relevant claims from a list of claims that are previously fact-checked. We set up two tracks for the shared task. In the monolingual track, matched posts and claims are always in the same language and the search pools for candidate claims are language specific to the post. In the crosslingual track, the search pool of candidate claims is not restricted and the matched claims can be in languages that are different from the post.

3.1 Dataset

The dataset used in the shared task is available for research purposes at Zenodo[5]. The training and development sets are based on the MultiClaim (Pikuliak et al., 2023) dataset. The test set contains additional data, which are not part of the original MultiClaim dataset, but which were collected using the same methodology.

Training and Development Sets. The original MultiClaim dataset consists of two parts: **1. Fact-checked claims**. The dataset contains 205,751 fact-checked claims extracted from factchecks created by 142 fact-checking organizations. Claims are usually one sentence long summaries of the main idea that is being fact-checked. **2. Social media posts.** The dataset contains 28,092 SMPs spanning 27 different languages, collected from Facebook, Instagram, and Twitter by following the links to these platforms given in the fact-checking articles. This way, 31,305 SMPClaim pairs (each SMP has at least one claim assigned) were established. All posts in MultiClaim were published before 2023.

To make sure that only relevant SMPs are considered, two filtering techniques were used: (1) We use links from the ClaimReview json schema that the fact-checkers use. (2) We use the factchecking warnings provided by the Facebook and Instagram platforms. When a visitor visits SMPs that were flagged as disinformation, the warnings contain a link to relevant fact-checks. In both cases, we can be sure that the link between a claim and an SMP is correct, because a fact-checker left an explicit signal confirming it. Manual inspections of the connections (done by three authors on a random sample of 100 pairs; see Pikuliak et al., 2023 for more details) confirmed the absence of false positives. However, roughly 15% of the connections rely on visual information present in either attached images or videos, making it impossible to match the SMP with an appropriate claim based on the text data only.

An additional manual inspection (of claims retrieved for a sample of 87 posts done by three authors; see Pikuliak et al., 2023 for more details) also revealed that there are many potential connections between SMPs and claims that are missing in the dataset due to the employed collection procedure, especially crosslingual connections. To at least partially mitigate this, we added to the dataset additional SMP-Claim pairs created by following transitive connections: If two SMPs p_i and p_j share a link to the same fact-checked claim

[5] https://doi.org/10.5281/zenodo.14989176

and one of those (p_j) is also linked to a different claim c_k, we add the link (p_j, c_k), if missing. This way, we were able to identify 3,351 additional SMP-Claim pairs.

Table 2. The dataset statistics reporting the number of posts, claims, and pairs for the monolingual (shown also per individual languages) as well as for the crosslingual track.

Language	# of posts			# of claims		# of pairs		
	train	dev	test	train & dev	test	train	dev	test
Arabic	676	78	500	14,201	21,153	680	78	514
English	4,351	478	500	85,734	145,287	5,446	627	574
French	1,596	188	500	4,355	6,316	1,667	200	740
German	667	83	500	4,996	7,485	830	101	549
Malay	1,062	105	93	8,424	686	1,169	116	96
Portuguese	2,571	302	500	21,569	32,598	3,386	403	613
Spanish	5,628	615	500	14,082	25,440	6,313	692	576
Thai	465	42	183	382	583	465	42	183
Polish	-	-	500	-	8,796	-	-	564
Turkish	-	-	500	-	12,536	-	-	550
Monolingual (total)	17,016	1,891	4,276	153,743	260,880	19,956	2,259	4,959
Crosslingual	4,972	552	4,000	153,743	272,447	5,787	651	5,322

Table 3. Combinations of languages in SMP-Claim pairs for train and dev sets of the crosslingual track.

Claim language SMP language	ara	deu	eng	fra	msa	por	spa	tha
Arabic (ara)	94	0	17	19	3	0	2	0
German (deu)	1	84	4	3	0	4	19	1
English (eng)	15	50	699	90	82	135	225	17
French (fra)	5	0	13	218	2	12	6	0
Hindi (hin)	0	0	964	0	0	0	0	0
Korean (kor)	0	0	257	0	0	0	0	0
Malay (msa)	0	1	96	0	135	0	0	3
Portuguese (por)	1	0	22	3	2	392	36	1
Sinhala (sin)	0	0	504	0	0	0	0	0
Spanish (spa)	3	1	40	7	2	36	804	0
Tagalog (tgl)	0	0	258	0	0	0	0	0
Thai (tha)	0	0	88	0	0	0	0	50
Urdu (urd)	0	0	373	0	0	0	0	0
Chinese (zho)	0	0	539	0	0	0	0	0

To construct high-quality training and development sets covering different languages, we filter out languages that have less than 500 SMP-Claim pairs. After filtering, the dataset contains 8 languages (Arabic, English, French, German, Malay, Portuguese, Spanish, and Thai) for the monolingual track. For the crosslingual track, we include all SMP-Claim pairs, where the language of the claim is different than the language of the post and at the same time, both of the languages are already included in the monolingual pairs; additionally, all pairs with a language combination appearing at least 200 times are included even if these languages do not appear in the monolingual pairs. There are still only 8 claim languages, but there are 6 additional post languages as compared to the monolingual track (Hindi, Korean, Sinhala, Tagalog, Urdu, and Chinese). There are a total of 52 different combinations that are covered in the crosslingual track. To prevent an exploitation of the data design in the crosslingual task (e.g., by ignoring the language of the post, if the matched claim is always in a different language), we assign a randomly selected 10% of monolingual pairs into the crosslingual subset of the data. The development set was created by holding out a randomly selected 10% of pairs for the monolingual and the crosslingual tracks. The final numbers of selected posts, claims, and pairs are reported in Table 2. Table 3 shows the combinations of languages in SMP-Claim pairs for the combined training and development sets.

Table 4. Combinations of languages in SMP-Claim pairs for the test set of the crosslingual track.

Claim language SMP language	ara	deu	eng	fra	hin	msa	pol	por	spa	tha	tur
Arabic (ara)	226	2	39	4	0	0	0	1	8	0	3
Bengali (ben)	0	0	149	0	0	0	0	0	0	0	0
German (deu)	0	110	16	2	0	0	2	0	4	0	5
English (eng)	14	51	1,030	29	186	4	12	24	39	2	49
French (fra)	1	4	28	56	0	0	0	1	3	0	1
Hindi (hin)	0	0	2,337	0	0	0	0	0	0	0	0
Malay (msa)	0	0	8	0	0	5	0	0	0	0	0
Polish (pol)	0	0	14	2	0	0	55	1	0	0	0
Portuguese (por)	6	3	21	2	0	0	0	77	15	0	2
Spanish (spa)	1	0	40	1	0	0	0	11	211	0	3
Thai (tha)	0	0	18	0	0	0	0	0	0	13	0
Turkish (tur)	0	1	1	1	0	0	0	0	0	0	140
Urdu (urd)	0	0	147	0	0	0	0	0	0	0	0
Chinese (zho)	0	0	81	0	0	0	0	0	0	0	0

Test set. Our test set consists of additionally collected data containing posts that are not a part of the original MultiClaim dataset and which were published until March 2024. For fact-checked claims, the test set contains the ones already present in the training and development sets, as well as newly collected ones (ranging until May 2024). The same data collection methodology was applied as in the original MultiClaim dataset, while being extended to more sources (more fact-checking organizations and also including Telegram in case of SMPs). Similarly, the same data cleaning and pre-processing was applied as described in Pikuliak et al. (2023) to ensure consistency.

For the monolingual track, the same 8 languages were used as in the training and development sets. Additionally, we added two unannounced languages: Polish and Turkish, which both meet the criteria imposed on the number of SMP-Claim pairs and which did not appear in the prior phases of the shared task in either of the subtracks. We select a random set of 500 posts (where available; only 93 and 183 posts were available for Malay and Thai, resp.) for each language and all fact-checked claims available for that language.

To construct a test set for the crosslingual track, we again apply the same criteria as for the training and development sets, resulting in 11 languages for claims (10 from the monolingual track plus Hindi), 14 languages in SMPs – the ones not contained in the monolingual track being Hindi, Urdu, Chinese, and Bengali – and 60 language combinations in total. The complete statistics for the test set are reported in Table 2. Table 4 shows the combinations of languages in SMP-Claim pairs contained in the test set. It is worth noting that despite the language diversity, most crosslingual pairs still contain English as either a language of a post or of a claim. This is one of the limitations of the dataset and it needs to be taken into consideration when interpreting the results of the crosslingual track.

Details on Data Format. Each post has five fields as illustrated in Table 5. The *post_id* refers to the *id* of the post in our dataset. The *instances* field is used to indicate the timestamp of the post and its social platform. The *ocr* field is filled with transcribed texts if there are any associated images in the post. The *verdict* is assigned by professional fact-checkers. The *text* field contains the content of the post, its translation to English and a list of detected languages.

Table 5. An example of a social media post. There are five fields for each post.

post_id	27169
instances	(1620128767, 'fb')
ocr	[('Merche Gonzalez de Mingo-Sancho 1h. En mi colegio las papeletas de Vox al lado de las de Volt para liar a los abuelos... Vot vox xx', "Merche Gonzalez de Mingo-Sancho 1 hour. In my school the Vox ballots next to Volt's to mess with the grandparents... vote vox xx", [('spa', 0.6682217717170715), ('cat', 0.2810060381889343), ('eng', 0.01799275539815426)])]
verdicts	Partly false information
text	('Qué casualidad, no dejan ni un cabo suelto...., cuanto más confusión crean, cuanto más caos... mayor cosecha intenta sacar la siniestra. La izquierda, siempre con la mentira y la trampa por bandera.', "What a coincidence, they don't leave a single end loose...., the more confusion they create, the more chaos... the bigger harvest the sinister tries to get. The left, always with lies and cheating as a flag.", [('spa', 1.0)])

Each claim has three fields as illustrated in Table 6. The *fact_check_id* is the identification of the fact-check (claim) in the dataset. The *claim* field contains the content of the claim, its translation to English and a list of identified languages. Similar to the structure of posts, the *instances* field is used to indicate the timestamp of the claim and its relevant URL. The *title* field contains the original title of the fact-check, its translation to English and a list of identified languages.

We utilize the Google Vision API and Google Translate API to obtain the texts in the image associated to posts and the translations of non-English texts respectively.

Table 6. An example of a fact-checked claim. There are four fields for each claim.

fact_check_id	74855
claim	('Jarum suntik palsu dalam sebuah video yang diklaim telah disiapkan untuk vaksinasi Covid-19 para pemimpin dunia atau elite global', 'Fake syringe in a video that claims to have been prepared for the Covid-19 vaccination of world leaders or global elites', [('msa', 1.0)])
instances	[(1612137540, 'https://cekfakta.tempo.co/fakta/1221/keliru-jarum-suntik-palsu-di-video-ini-disiapkan-untuk-vaksinasi-covid-19-elite-global')]
title	('Keliru, Jarum Suntik Palsu di Video Ini Disiapkan untuk Vaksinasi Covid-19 Elite Global', 'Wrong, Fake Syringe in this Video Prepared for Global Elite Covid-19 Vaccination', [('msa', 1.0)])

3.2 Evaluation

The participants were asked to provide top 10 results for each SMP. The lists of fact-checked claims are provided as a search pool. We calculate *success@ 10* to evaluate the performance of all systems:

$$success@10 \ = \frac{1}{N}\sum_{i=1}^{N} 1\ (relevant\ claim\ in\ top\ 10)$$

where *N* is the total number of examples (SMPs), and 1(·) is an indicator function that returns 1 if the condition is met for the post *i*, otherwise 0. We count a retrieval as success if at least one relevant claim is ranked in the top 10.

4 Baselines

We selected four approaches as baselines:

BM25. BM25 (Robertson and Zaragoza, 2009) is used as a default retriever in many prior works on claim matching (see, e.g., Vo and Lee, 2020; Shaar et al., 2020, 2022) and it also achieved competitive results especially in monolingual evaluation performed in Pikuliak et al. (2023). We use it with the English-translated version of the dataset.

GTR-T5-Large. GTR-T5-Large (Ni et al., 2022) was the overall best performing model in Pikuliak et al. (2023) in monolingual as well as crosslingual settings. Since it is an English-only model, we use it with the English-translated version of the dataset.

Paraphrase-Multi-v2. We use the Paraphrase-Multilingual-MPNet-Base-v2 model, which is a part of the Sentence-BERT set of pre-trained models (Reimers and Gurevych, 2019). We selected this model as one of the baselines, since it was the best multilingual model in Pikuliak et al. (2023), although having lower performance than both BM25 and GTR-T5-Large (which, however, work on the English translations). The original multilingual version of the dataset is used with this model.

E5-Large. We use Multilingual-E5-Large model (Wang et al., 2024a), which belongs to the best performing multilingual text embedding models under 1B parameters based on public benchmarks (2nd for reranking task, 6th for retrieval and sentence similarity tasks) and it had competitive results in our initial experiments. Being a multilingual model, it works with the original multilingual version of the dataset.

5 Participating Systems and Results

The high-level overview of approaches utilized by individual systems is summarized in Table 7. The results presented in Tables 8 and 9 reflect the performance of test submissions by the end of the test phase. In their system papers, participants may provide additional post-test analysis that achieve further improvements.

Table 7. The overview of approaches utilized in the created systems as reported by individual teams. Note: Some teams are missing as they have not responded with a report about their system. Order of teams corresponds to the rank on the Monolingual Track.

Team Name	Pre-processing/Data Curation			Model Training				Post-processing/Reranking		
	Data augmentation	Data filtering/ cleaning	Other	Pre-trained model	Single model on English data	Single model on all languages	Ensemble of per-language models	Results reranking	Results filtering	Other
PALI				✓		✓				
TIFIN India (Devadiga et al., 2025)	✓		✓	✓	✓			✓		
RACAI (Chivereanu and Tufis, 2025)				✓						
QUST_NLP (Liu et al., 2025)			✓	✓			✓	✓		
UniBuc-AE (Enache, 2025)				✓	✓	✓				✓
UWBa (Lenc et al., 2025)				✓						
ipezoTU (Pezo et al., 2025)			✓	✓				✓		
fact check AI (Rastogi, 2025)		✓		✓	✓	✓			✓	
YNU-HPCC (Mao et al., 2025)	✓	✓		✓		✓				
DKE-Research (Wang and Wang, 2025)				✓						
CAIDAS (Benchert et al., 2025)	✓			✓		✓				
UPC-HLE (Becerra-Tome and Conesa, 2025)		✓		✓	✓	✓		✓		
ClaimCatchers (Panchendrarajan et al., 2025)				✓	✓			✓		
Shouth NLP (Harbin and Pérez, 2025)				✓		✓				
MultiMind (Abootorabi et al., 2025)	✓	✓		✓		✓		✓		
JU_NLP (Nayak et al., 2025)		✓		✓						
Duluth (Syed and Pedersen, 2025)				✓		✓				
Howard Univeristy-AI4PC (Rijal and Aryal, 2025)				✓						
CAISA (Haroon et al., 2025)		✓	✓	✓		✓			✓	

5.1 Monolingual Track

In the monolingual track, posts and claims are in the same language, and multilingual data cover ten different languages. Two (Turkish and Polish) out of ten are unannounced languages – i.e., no examples in these languages appear in the training and development sets. 21 out of 27 teams have submitted their system description papers on the monolingual track. The results are summarized in Table 8. Out of all teams that submitted system description papers, we describe the top-5 performing systems.

TIFIN India. Inspired by Khan et al. (2024), Devadiga et al. (2025) first translate all posts and claims into English and utilize English-focused models for the claim retrieval task. They employ the Aya-expanse-8B (Dang et al., 2024) model for the translation. Then, a two-stage pipeline is implemented. The Stella 400M embedding model (Zhang et al., 2024) is first used to retrieve top 50 claims using cosine similarity. These are fed into an LLM-based re-ranker, Qwen2.5-72B-Instruct (Yang et al., 2024). The re-ranking is performed in a prompting style. The top 10 candidates from the re-ranker are then combined with the top 10 results from the embedding model by Reciprocal Rank Fusion

(Cormack et al., 2009), producing the final top 10 results. They further fine-tune the embedding model by adding hard negatives, demonstrating that 20 negatives per post give the best performance.

*Table 8. The results (success@10) on Monolingual Track. Teams are ranked by their average performance across all languages. The best results for each language and the average are in bold. The rows with gray background and * indicate the performance of four baselines. Teams with reference have submitted system papers.*

Rank	Team Name	avg	eng	fra	deu	por	spa	tha	msa	ara	tur	pol
1	PINGAN AI	**0.9601**	**0.9160**	**0.9720**	**0.9580**	**0.9260**	**0.9740**	0.9945	**1.0000**	**0.9860**	**0.9480**	**0.9260**
2	PALI	0.9472	0.9040	0.9540	0.9360	0.9080	0.9700	**1.0000**	**1.0000**	0.9820	0.9300	0.8880
3	TIFIN India (Devadiga et al., 2025)	0.9383	0.8800	0.9540	0.9360	0.9020	0.9600	0.9945	**1.0000**	0.9660	0.9040	0.8860
4	RACAI (Chivereanu and Tufis, 2025)	0.9377	0.9000	0.9420	0.9280	0.8960	0.9520	0.9945	**1.0000**	0.9660	0.9160	0.8820
5	QUST_NLP (Liu et al., 2025)	0.9365	0.8940	0.9500	0.9020	0.8900	0.9480	0.9945	**1.0000**	0.9700	0.9300	0.8860
6	UniBuc-AE (Enache, 2025)	0.9336	0.8860	0.9200	0.9320	0.8800	0.9620	**1.0000**	**1.0000**	0.9700	0.9100	0.8760
7	UWBa (Lenc et al., 2025)	0.9270	0.8800	0.9440	0.9160	0.8540	0.9380	**1.0000**	**1.0000**	0.9540	0.9120	0.8720
8	ipezoTU (Pezo et al., 2025)	0.9259	0.8900	0.9440	0.9180	0.8800	0.9300	0.9836	0.9892	0.9400	0.9100	0.8740
9	kubapok	0.9245	0.8700	0.9420	0.9220	0.8680	0.9420	0.9945	0.9785	0.9440	0.9120	0.8720
10	fact check AI (Rastogi, 2025)	0.9232	0.8820	0.9440	0.9260	0.8660	0.9420	0.9945	0.9892	0.9400	0.8840	0.8640
11	YNU-HPCC (Mao et al., 2025)	0.9218	0.8520	0.9540	0.9040	0.8740	0.9400	0.9727	0.9892	0.9580	0.8960	0.8780
12	UWOB	0.9190	0.8800	0.9340	0.9060	0.8540	0.9380	0.9781	0.9677	0.9540	0.9060	0.8720
13	TM_Trek	0.9189	0.8440	0.9380	0.9180	0.8180	0.9480	0.9945	**1.0000**	0.9660	0.8840	0.8780
14	joeblack	0.9105	0.8160	0.9280	0.9160	0.8040	0.9340	0.9891	**1.0000**	0.9620	0.8800	0.8760
15	DKE-Research (Wang and Wang, 2025)	0.8979	0.8200	0.9240	0.8680	0.8340	0.9160	0.9508	**1.0000**	0.9360	0.8740	0.8560
16	CAIDAS (Benchert et al., 2025)	0.8953	0.8340	0.9100	0.9000	0.8340	0.8860	0.9836	0.9677	0.9340	0.8680	0.8360
17	UPC-HLE (Becerra-Tome and Conesa, 2025)	0.8915	0.7940	0.9460	0.9220	0.8340	0.9140	0.9781	0.9892	0.9460	0.7920	0.8000
18	ClaimCatchers (Panchendrarajan et al., 2025)	0.8780	0.8100	0.9100	0.8420	0.8000	0.8860	0.9727	0.9570	0.9320	0.8740	0.7960
19	NCL-AR (Robertson and Liang, 2025)	0.8716	0.8340	0.9180	0.8980	0.7980	0.8780	0.9454	0.9462	0.8840	0.8140	0.8000
20	Shouth NLP (Harbin and Pérez, 2025)	0.8674	0.7960	0.8940	0.8580	0.7960	0.8660	0.9399	0.9785	0.9120	0.8280	0.8060
*	E5-Large	0.8589	0.8180	0.8540	0.8720	0.8080	0.8560	0.9290	0.8602	0.9160	0.8780	0.7980
*	BM25	0.8123	0.7500	0.8220	0.8120	0.7540	0.7960	0.9071	0.9140	0.8460	0.7900	0.7320
21	MultiMind (Abootorabi et al., 2025)	0.8080	0.6740	0.8640	0.8000	0.7480	0.7760	0.9235	0.9570	0.8480	0.7460	0.7440
22	JU_NLP (Nayak et al., 2025)	0.7868	0.6540	0.8700	0.7320	0.6460	0.6840	0.9290	0.9570	0.8740	0.7800	0.7420
*	GTR-T5-Large	0.7299	0.7880	0.7300	0.7380	0.7140	0.7320	0.7049	0.7097	0.8620	0.7160	0.6040
23	Duluth (Syed and Pedersen, 2025)	0.6883	0.4520	0.8140	0.6900	0.5580	0.5460	0.8415	0.8495	0.8200	0.6860	0.6260
*	Paraphrase-Multi-v2	0.5683	0.6180	0.6040	0.5340	0.4840	0.5160	0.6175	0.5054	0.6940	0.5740	0.5360
24	Word2winners (Azadi et al., 2025)	0.5488	0.6160	0.5460	0.5320	0.6000	0.5480	0.4372	0.4731	0.7080	0.5220	0.5060
25	UMUTeam (Pan et al., 2025)	0.5445	0.4140	0.5640	0.3840	0.4700	0.4200	0.7869	0.6882	0.7160	0.5380	0.4640
26	FactDebug (Nikolaev et al., 2025)	0.2213	0.3620	0.2640	0.4240	0.2360	0.1820	0.2350	0.0645	0.4460	0.0000	0.0000
27	TransformerHHU	0.1499	0.1980	0.1080	0.1800	0.1880	0.2220	0.1148	0.2043	0.0660	0.1140	0.1040

RACAI. Chivereanu and Tufis (2025) adopt a single-step strategy to tackle the task without reranking. All posts and claims are used in their original languages. The BGE-Multilingual-Gemma2-9B model (Xiao et al., 2023; Chen et al., 2024), is used to provide embeddings for posts and claims. They explored different strategies to optimize and adapt the general-purpose embedding model on the given task. They first utilize LoRA (Hu et al., 2022) with a contrastive learning objective with in-batch negatives to fine-tune the base model. Additionally, they use prompt-tuning to adjust the embeddings for instructions. To reduce the computational requirements, Matryoshka representation learning (Kusupati et al., 2022) is employed to produce embeddings of different dimension sizes.

QUST_NLP. Liu et al. (2025) propose a threestage retrieval framework. A group of six different retrieval models is utilized in the retrieval stage to coarsely identify relevant claims for each post. In the re-ranking stage, six different re-ranking models are employed to provide top 10 candidates from the retrieved results. For each language, a set of re-ranker models are fine-tuned with training data in that language. Finally, the weighted voting stage combines the top 10 candidates from each re-ranker and weight them by their performance on the dev set to obtain the final top 10 results. They also find that the combination of the original text and corresponding English translations can further improve the overall performance.

UniBuc-AE. Enache (2025) utilises both the original text and English-translated text for retrieval. Two embedding models (English one: e5-largev2, Wang et al., 2022, and multilingual one: multilingual-e5-large-instruct, Wang et al., 2024b) are employed to provide corresponding embeddings after contrastive learning with in-batch negatives. They further fine-tune the two models with hard negatives, resulting in two hard fine-tuned models. All four models are used to vote for the final top 10 retrieval claims in a weighted manner. The optimal weights are discovered by their performance on the development set.

UWBa. Lenc et al. (2025) present a zero-shot claim retrieval approach with all posts and claims translated into English. Five embedding models without any further fine-tuning are utilized to provide embeddings for posts/claims. A model combination strategy is employed to produce the final top 10 candidates. The models are firstly ranked in term of performance on dev set. Their top 5 retrieval results are then combined to produce the top 10 results after potential de-duplication.

Discussion. We can see that most systems outperformed the baselines. On the other hand, a clear pattern is that the two unseen languages, Turkish and Polish, show a generally worse performance than the languages observed in the training data. This indicates that the generalization ability to unseen languages is still a big challenge. Approximately half of the submissions to the monolingual track did not apply any data filtering or pre-processing techniques, including many good-performing competitors. This does not mean that these techniques (e.g., using LLMs for data augmentation) do not improve performance, but others, such as fine-tuning, have a higher impact. On the other hand, a common and successful strategy (used by three out of top-5 described systems) was to improve base embedding models by finetuning them in a contrastive learning manner with in-batch negatives and using hard-negatives for further improvements. Many systems utilized a one-stage approach (retrieval with fine-tuned or vanilla embedding models) to tackle the retrieval task, while some systems built two- or three-stage pipelines (adding a re-ranking or a weighted voting stage) that demonstrated a better performance in some cases. However, the results do not provide a conclusive evidence whether adding them is in general better, especially since we do not compare the systems in terms of their computational requirements. In the monolingual track, there are subsets for different languages. Yet, most systems chose to train one model for all languages. This was achieved by either translating everything into English and utilizing an English-focused model, or, more often, utilizing a multilingual embedding model with the original data. When comparing the former (e.g., TIFIN India) with the latter (e.g., RACAI), we can see that the differences between the two approaches are negligible, which clearly shows the recent improvements in multilingual models compared to the situation reported in Pikuliak et al. (2023). A small number of systems trained a set of models for each language specifically. Some other models utilized both texts in original language and their English translations, reporting improved performance.

5.2 Crosslingual Track

20 out of 28 teams submitted system description papers for the crosslingual track. The results are summarized in Table 9. We describe the top 5 performing systems of those teams that submitted system description papers.

*Table 9. The results (success@10) on Crosslingual Track. Teams are ranked by their performance. The best result is in bold. The rows with gray background and * indicate the performance of baselines. Teams with reference have submitted system papers.*

Rank	Team Name	S@10
1	PINGAN AI	**0.85875**
2	PALI	0.82675
3	RACAI (Chivereanu and Tufis, 2025)	0.82450
4	TIFIN India (Devadiga et al., 2025)	0.81025
5	fact check AI (Rastogi, 2025)	0.79750
6	kubapok	0.79700
7	QUST_NLP (Liu et al., 2025)	0.79250
8	UniBuc-AE (Enache, 2025)	0.79000
9	UWBa (Lenc et al., 2025)	0.78250
10	UWOB	0.78250
11	YNU-HPCC (Mao et al., 2025)	0.77025
12	ipezoTU (Pezo et al., 2025)	0.74775
13	CAIDAS (Benchert et al., 2025)	0.74475
14	TM_Trek	0.73975
15	ClaimCatchers (Panchendrarajan et al., 2025)	0.73675
16	joeblack	0.71875
17	DKE-Research (Wang and Wang, 2025)	0.71325
*	GTR-T5-Large	0.70525
18	Team I2R	0.69725
19	UPC-HLE (Becerra-Tome and Conesa, 2025)	0.63850
*	E5-Large	0.63725
20	JU_NLP (Nayak et al., 2025)	0.61900
*	BM25	0.60275
21	Howard Univeristy-AI4PC (Rijal and Aryal, 2025)	0.59225
22	Word2winners (Azadi et al., 2025)	0.55425
23	CAISA (Haroon et al., 2025)	0.54475
24	MultiMind (Abootorabi et al., 2025)	0.48900
*	Paraphrase-Multi-v2	0.41075
25	NCL-AR (Robertson and Liang, 2025)	0.39775
26	FactDebug (Nikolaev et al., 2025)	0.32000
27	UMUTeam (Pan et al., 2025)	0.28025
28	DANGNT-SGU	0.00025

RACAI. The same pipeline is employed as for the monolingual track, where the base embedding model is improved by contrastive learning with in-batch negatives and prompt-tuning. They demonstrate that it brings much higher improvements in crosslingual compared to monolingual setup.

TIFIN India. They translate everything into English and use the same two-stage pipeline as in the monolingual track. Their high performance shows that this strategy is effective in both monolingual and crosslingual setups.

fact check AI. Rastogi (2025) uses the English translations. Base embedding models (mul-e5- large-instruct, Stella-400M, and mxbai-embedlarge- v1) are fine-tuned with contrastive learning. The fine-tuning is performed in a K-fold manner to increase the robustness and prevent over-fitting. The final 10 candidates are produced by models' voting. They also conduct pre-processing such as removing noise (e.g., url, emoji, extra space) and filtering out too-short texts.

QUST_NLP. The same three-stage framework is utilized for the crosslingual track. Different from the language-specific fine-tuning strategy for rerankers in the monolingual track, they use English translations and employ one set of re-rankers for refined re-ranking. They also demonstrate that though the combination of the original text and English translations are helpful in the monolingual track, such combination often causes performance decrease in the crosslingual track.

UniBuc-AE. They adopt the same strategy for the crosslingual track where four embedding models are used for weighted voting. They show that fine-tuning with hard negatives brings larger improvements in the crosslingual setup.

Discussion. We can see that most systems outperform the baselines, but they achieve a worse performance compared to the monolingual track (more than 10 percentage points difference), highlighting the challenge of this setup. Also, most systems that were submitted to both tracks do not consider a new approach but adopt the same pipeline, which seems to work quite well, given that it is the case for all top-5 best performing systems in the crosslingual track. However, they also highlight some performance inconsistencies of different techniques when applied across the tracks. For example, the improvements brought in by additional re-ranking and weighted-voting seem to be larger in the crosslingual setup. On the other hand, as indicated by Liu et al. (2025), the combination of original text with English translations instead has a negative impact on the performance in the crosslingual setup. This raises questions on the transferability of such techniques and models to the crosslingual scenario, which future research should examine more comprehensively.

6 Conclusions

This report presents an overview of SemEval 2025 Task 7 on multilingual and crosslingual factchecked claim retrieval. It provides a comprehensive analysis of the dataset and submitted systems. The task has attracted 179 participants. In the final test phase, 31 teams submitted test systems, of which 23 submitted system description papers. Our analysis reveals that the most common strategy to improve base embedding models is to finetune them in a contrastive learning manner with in-batch negatives. As demonstrated by many systems, further improvements can be obtained by adding hard negatives and training on diverse subsets. A multi-stage pipeline is adopted by a large number of systems that additionally include a re-ranking and a weighted voting stage, which can bring improvements in both crosslingual and monolingual setups. An effective strategy to tackle different languages is still to translate them into English and utilize an English-focused model, but using multilingual models with the texts in original languages already gives comparable results. Several systems have discovered that some techniques which work well in one track have a different impact in the other track, suggesting the difficulty of transferability. We believe our shared task along with participating systems provides valuable insights into multilingual and crosslingual claim retrieval, supporting future research in this field.

Considering the broader goals of DisAI, we may also conclude that the organisation of the shared task helped to provide hands-on experience with multilingual data preparation, evaluation design, and large-scale coordination – competencies central to the project's focus. Through the collaboration with leading European partners and the engagement of 179 participants worldwide, KInIT also increased its international visibility, expanded its research network, and enhanced its reputation in multilingual fact-checking research. The task also supported the development of early-stage researchers by involving them in dataset creation, evaluation, and community interaction, aligning with the project's objectives of improving scientific skills, fostering innovation, and strengthening KInIT's role in the European Research Area.

References

Mohammad Mahdi Abootorabi, Alireza Ghahramani Kure, Mohammadali Mohammadkhani, Sina Elahimanesh, and Mohammad ali Ali panah. 2025. MultiMind at SemEval-2025 Task 7: Crosslingual fact-checked claim retrieval via multi-source alignment. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors. 2023. Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER). Association for Computational Linguistics, Dubrovnik, Croatia.

Firoj Alam, Alberto Barrón-Cedeño, Gullal S Cheema, Gautam Kishore Shahi, Sherzod Hakimov, Maram Hasanain, Chengkai Li, Rubén Míguez, Hamdy Mubarak, Wajdi Zaghouani, et al. 2023. Overview of the CLEF-2023 CheckThat! Lab Task 1 on check-worthiness of multimodal and multigenre content.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. Proceedings of the International AAAI Conference on Web and Social Media, 14(1):821–829.

Amirmohammad Azadi, Sina Zamani, Mohammadmostafa Rostamkhani, and Sauleh Eetemadi. 2025. Word2winners at SemEval-2025 Task 7: Multilingual and crosslingual fact-checked claim retrieval. In Proceed- ings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Firoj Alam, Tanmoy Chakraborty, Tamer Elsayed, Preslav Nakov, Piotr Przybyła, Ju- lia Maria Struß, Fatima Haouari, Maram Hasanain, Federico Ruggeri, Xingyi Song, and Reem Suwaileh. 2024.

The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In Advances in Information Retrieval, pages 449–458, Cham. Springer Nature Switzerland.

Alberto Becerra-Tome and Agustín Conesa. 2025. UPC-HLE at SemEval-2025 Task 7: Multilingual fact-checked claim retrieval with text embedding models and cross-encoder re-ranking. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Dominik Benchert, Severin Meßlinger, Sven Goller, Jonas Kaiser, Jan Pfister, and Andreas Hotho. 2025. CAIDAS at SemEval-2025 Task 7: Enriching sparse datasets with LLM-generated content for improved information retrieval. In Proceedings of the 19th

International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Radu Chivereanu and Dan Tufis. 2025. RACAI at SemEval-2025 Task 7: Efficient adaptation of large language models for multilingual and crosslingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Christos Christodoulopoulos, James Thorne, Andreas Vlachos, Oana Cocarascu, and Arpit Mittal, editors. 2020. Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER). Association for Computational Linguistics, Online.

European Commission. 2022. Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/ec (digital services act).

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. 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, pages 758–759.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. arXiv preprint arXiv:2412.04261.

Prasanna Jagadeesh Devadiga, Arya Suneesh, Pawan Kumar Rajpoot, Bharatdeep Hazarika, and Aditya U. Baliga. 2025. TIFIN India at SemEval-2025 task 7: Harnessing translation to overcome multilingual IR challenges in fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Alexandru Enache. 2025. UniBuc-AE at SemEval-2025 Task 7: Training text embedding models for multilingual and crosslingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Ahmet Bahadir Eyuboglu, Bahadir Altun, Mustafa Bora Arslan, Ekrem Sonmezer, and Mucahid Kutlu. 2023. Fight against misinformation on social media: detecting attention-worthy and harmful tweets and verifiable and check-worthy claims. In International Conference of the Cross-Language Evaluation Forum for European Languages, pages 161–173. Springer.

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

Santiago Lares Harbin and Juan Manuel Pérez. 2025. Shouth NLP at SemEval-2025 Task 7: Multilingual fact- checking retrieval using contrastive learning. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Muqaddas Haroon, Shaina Ashraf, Ipek Baris, and Lucie Flek. 2025. CAISA at SemEval-2025 Task 7: Multilingual and crosslingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Andrea Hrckova, Robert Moro, Ivan Srba, Jakub Simko, and Maria Bielikova. 2024. Autonomation, not automation: Activities and needs of fact-checkers as a basis for designing human-centered ai systems.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

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), pages 4504–4517, Online. Association for Computational Linguistics.

Mohammed Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, B Suriyaprasaad, G Varun, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, et al. 2024. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15831–15879.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. Digital Threats, 2(2).

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5430–5443.

S Vinod Kumar, Guravana Jothsna, Mummina Poojitha, Kandula Deepika, Indubu Nutana, MVS Ajit, and Konna Lalitendra Swamy. 2024. Verdict prediction using artificial neural networks. In 2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), pages 1–5. IEEE.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. 2022. Matryoshka representation learning. Advances in Neural Information Processing Systems, 35:30233–30249.

Michelle Seng Ah Lee and Jatinder Singh. 2021. Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, page 704–714, New York, NY, USA. Association for Computing Machinery.

Ladislav Lenc, Daniel Cífka, Jiˇrí Martínek, Jakub Šmíd, and Pavel Král. 2025. UWBa at SemEval-2025 Task 7: Multilingual and crosslingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 4461–4472.

Youzheng Liu, Jiyan Liu, Xiaoman Xu, Taihang Wang, Yimin Wang, and Ye Jiang. 2025. QUST_NLP at SemEval- 2025 Task 7: A three-stage retrieval framework for monolingual and crosslingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 505–514.

Yuheng Mao, Jin Wang, and Xuejie Zhang. 2025. YNU-HPCC at SemEval-2025 Task 7: Multilingual and cross- lingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or false: Studying the work practices of professional fact-checkers. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1):1– 44.

Ndapandula Nakashole and Tom Mitchell. 2014. Language-aware truth assessment of fact candidates. In Proceed- ings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1009–1019.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, et al. 2022. Overview of the CLEF–2022 CheckThat! Lab on fighting the COVID-19 infodemic and fake news detection. In International conference of the cross-language evaluation forum for european languages, pages 495–520. Springer.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Atanu Nayak, Srijani Debnath, Arpan Majumdar, Pritam Pal, and Dipankar Das. 2025. JU_NLP at SemEval-2025 Task 7: Leveraging transformer-based models for multilingual & crosslingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Evgenii Nikolaev, Ivan Bondarenko, Islam Aushev, Vasilii Krikunov, Andrei Glinskii, Vasily Konovalov, and Julia Belikova. 2025. FactDebug at SemEval-2025 Task 7: Hybrid retrieval pipeline for identifying previously fact-checked claims across multiple languages. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, and Rafael Valencia-García. 2025. UMUTeam at SemEval-2025 Task 7: Multilingual fact-checked claim retrieval with XLM-RoBERTa and self-alignment pretraining strategy. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Rrubaa Panchendrarajan, Rafael Martins Frade, and Arkaitz Zubiaga. 2025. ClaimCatchers at SemEval-2025 Task 7: Sentence transformers for claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Iva Pezo, Allan Hanbury, and Moritz Staudinger. 2025. ipezoTU at SemEval-2025 Task 7: Hybrid ensemble retrieval for multilingual fact-checking. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleˇn, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16477–16500, Singapore. Association for Computational Linguistics.

Pranshu Rastogi. 2025. fact check AI at SemEval-2025 Task 7: Multilingual and crosslingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Suprabhat Rijal and Saurav K. Aryal. 2025. Howard University-AI4PC at SemEval-2025 Task 7: Crosslingual fact-checked claim retrieval-combining zero-shot claim extraction and knn-based classification for multilingual claim matching. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Alex Robertson and Huizhi Liang. 2025. NCL-AR at SemEval-2025 Task 7: A sieve filtering approach to refute the misinformation within harmful social media posts. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. Founda- tions and Trends® in Information Retrieval, 3(4):333–389.

Daniel Russo, Serra Sinem Tekiroˇglu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. Transactions of the Association for Computational Linguistics, 11:1250–1264.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The automated verification of textual claims (AVeriTeC) shared task. In Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER).

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. Advances in Neural Information Processing Systems, 36:65128–65167.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. Computational Linguistics, 46(2):499–510.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The role of context in detecting previously fact-checked claims. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1619–1631, Seattle, United States. Association for Computational Linguistics.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3607–3618, Online. Association for Computational Linguistics.

S Suryavardan, Shreyash Mishra, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Naresh Reganti, Amitava Das, Amit P Sheth, Manoj Chinnakotla, et al. 2023. Findings of factify 2: multimodal fake news detection. In DE-FACTIFY@ AAAI.

Shujauddin Syed and Ted Pedersen. 2025. Duluth at SemEval-2025 Task 7: TF-IDF with optimized vector dimensions for multilingual fact-checked claim retrieval. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In Proceedings of the ACL 2014 workshop on language technologies and computational social science, pages 18–22.

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), pages 7717–7731, Online. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. Science, 359(6380):1146–1151.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 317–326, Online. Association for Computational Linguistics.

Yuqi Wang and Kangshi Wang. 2025. DKE-Research at SemEval-2025 Task 7: A unified multilingual framework for cross-lingual and monolingual retrieval with efficient language-specific adaptation. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. Language and Linguistics Compass, 15(10):e12438.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. arXiv preprint arXiv:2412.19048.