Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, Fact-Checking, Claim Normalization, and Retrieval

Firoj Alam[1(✉)], Julia Maria Struß[2], Tanmoy Chakraborty[5], Stefan Dietze[6,7], Salim Hafid[11], Katerina Korre[1], Arianna Muti[12], Preslav Nakov[4], Federico Ruggeri[3], Sebastian Schellhammer[6], Vinay Setty[8], Megha Sundriyal[10], Konstantin Todorov[11], and V. Venktesh[9]

[1] Qatar Computing Research Institute, HBKU, Doha, Qatar
fialam@hbku.edu.qa
[2] University of Applied Sciences Potsdam, Potsdam, Germany
[3] DISI, University of Bologna, Bologna, Italy
[4] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[5] Indian Institute of Technology Delhi, New Delhi, India
[6] GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
[7] Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany
[8] University of Stavanger, Stavanger, Norway
[9] Delft University of Technology, Delft, The Netherlands
[10] Indraprastha Institute of Information Technology, New Delhi, India
[11] University of Montpellier, LIRMM, CNRS, Montpellier, France
[12] Bocconi University, Milan, Italy
https://checkthat.gitlab.io

Abstract. This paper presents the eighth edition of the CheckThat! lab, part of the 2025 Conference and Labs of the Evaluation Forum (CLEF). As in previous editions of CheckThat!, the lab offers tasks from the core of the verification pipeline, including check-worthiness, identifying previously fact-checked claims, supporting evidence retrieval, and claim verification as well as auxiliary tasks addressing different facets of individual steps of the pipeline: Task 1 is on identification of subjectivity (a follow-up of the CheckThat! 2024 edition), which is related to the check-worthiness task, Task 2 is on claim normalization, Task 3 addresses fact-checking numerical claims, and Task 4 focuses on scientific web discourse processing. These challenging classification and retrieval problems are offered in different mono-, multi- and crosslingual settings covering more than 20 languages. This year, CheckThat! was one of the most popular labs at CLEF-2025 in terms of team registrations: 177 teams registered, almost half of them actually participating (a total of 83 teams) and 54 submitted system description papers.

Keywords: Fact-Checking · Check-Worthiness · Subjectivity · Claim verification

J. Carrillo-de-Albornoz et al. (Eds.): CLEF 2025, LNCS 16089, pp. 199–223, 2026.
https://doi.org/10.1007/978-3-032-04354-2_13

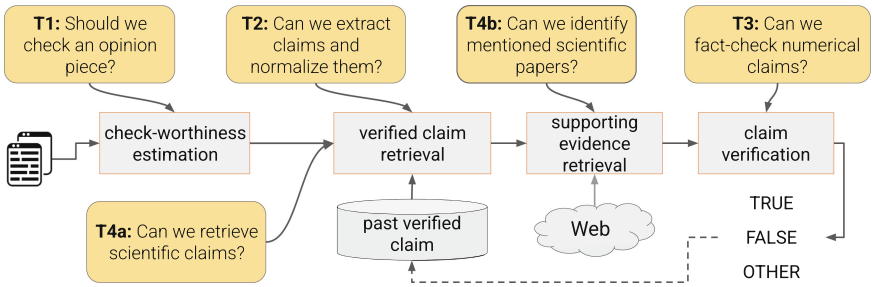

Fig. 1. The CheckThat! verification pipeline, featuring the four core tasks along with the CheckThat! 2025 tasks.

1 Introduction

The primary goal of CheckThat! is to promote the development of technology and resources to assist different tasks along the fact-checking verification pipeline, as well as auxiliary tasks that support the process. During the first five iterations of the lab the main focus was set on the core tasks of the verification pipeline (see Fig. 1). From the sixth edition [16] on, the lab has widened the focus and opened up for auxiliary tasks helping to address the different steps of the pipeline.

This year [6] we offered four tasks with multiple mono-, multi- and cross-lingual settings covering more than 20 languages. Task 1 [55] is a follow-up of CheckThat! 2023 and 2024 editions dealing with the subjectivity of sentences in news articles, in order to spot text that should be processed with specific strategies [52], potentially benefiting the fact checking pipeline [40,41,80]. Task 2 [74] addresses the challenge of claims buried within noisy, unstructured social media posts and asks to normalize the claim into unambiguous, check-worthy statements. Task 3 [79] tackles the challenge associated with verifying numerical and temporal claims. Task 4 [33] focuses on scientific web discourse offering two subtasks, firstly asking participants to classify different forms of science-related online discourse, and secondly asking participants to identify the source of an informal reference made in social media posts.

As in previous editions, CheckThat! was one of the most popular tasks at CLEF, attracting a total of 177 registrations and 83 actively participating teams, using a variety of approaches to the different tasks, mainly based on the encoding and decoding of large language models combined with different sources of information.

2 Previously on the CheckThat! Lab

In the past seven iterations, the CheckThat! lab have offered a variety of tasks from the verification pipeline, in a multitude of languages and in different domains. An overview is given in Table 1.

Table 1. Overview of the tasks offered in the previous editions of the lab.

tasks	years						domains					languages														
	2018	2019	2020	2021	2022	2023	2024	debates	speeches	tweets	web pages	news articles	Arabic	Bulgarian	Dutch	English	French	Georgian	German	Greek	Hind (code-mixed)	Italian	Polish	Portuguese	Russian	Slovenian	Spanish	Turkish
check-worthiness estimation	■	■	■	■	■	■	■	■	■	■			■	■	■	■											■	■
verified claim retrieval			■	■	■			■	■	■			■			■												
supporting evidence retrieval		■	■				■			■	■		■			■												
claim verification	■	■	■			■	■						■			■												
fake news detection				■	■							■				■			■									
bias						■						■				■												
subjectivity						■	■					■	■	■		■			■			■						■
topic identification				■												■												
authority finding						■				■			■															
adversarial robustness						■									■													
persuasion							■					■	■	■		■	■	■	■	■		■	■	■	■	■	■		

The first two editions of the lab in 2018 and 2019 focused on check-worthiness and claim verification of political debates and speeches in Arabic and English, with an additional focus on fact-checking by a task on classifying and ranking supporting evidence from the web in the second edition. The 2020 edition [17] covered the complete verification pipeline, with check-worthiness estimation, verified claim and supporting evidence retrieval, and claim verification, for the first time including social media data. The 2021 edition focused on multilinguality, offering tasks in five languages [47]. It also offered a fake news detection task, where the focus was on news articles, which was also continued in 2022.

The 2023 year's edition of the `CheckThat!` lab [16] paid special attention to the various sub-aspects of check-worthiness estimation, subjectivity of news articles, factuality, bias, authority findings, again in a multitude of languages. Transformer-based models were extensively used. This edition has also introduced multimodality for check-worthiness estimation.

3 Description of the 2025 Tasks

The 2025 edition of `CheckThat!` featured a total of four tasks in a variety of languages and modalities, three of which were run for the first time (cf. Sects. 3.1, 3.2, 3.3 and 3.4). Moreover, two of the tasks had two subtasks each (cf. Sects. 3.1 and 3.3).

3.1 Task 1: Subjectivity in News Articles

Verifiable claims are not only expressed through objective and neutral statements, but can also appear in subjectively framed ones. While objective sentences can be directly assessed for verification, subjective ones require additional processing, such as extracting their underlying objective content or any embedded claims. Consequently, the goal of this task is to determine whether a given sentence is subjective or objective. This is framed as a binary classification problem and is available in Arabic, Bulgarian, English, German, Italian, as well as in a multilingual setting. In this 2025 edition, the task of subjectivity in news articles also provides a zero-shot setting: a model is trained on data regarding certain languages and tested on data concerning unseen languages. In particular, we consider Greek, Romanian, Polish and Ukrainian as unseen languages and Arabic, Bulgarian, English, German, and Italian as training languages. A more detailed description and discussion of the task can be found in [72].

3.2 Task 2: Claim Normalization

Social media platforms impose minimal restrictions on writing, allowing users to post in vague and informal language. These posts often mix personal opinions, rhetorical questions, and incomplete information. This blend makes it difficult to identify clear claims – statements that assert something as true and can be verified or disproven [31]. As a result, fact-checkers face the difficult task of extracting concrete, check-worthy claims from noisy and unstructured content.

Claim Normalization addresses this challenge by transforming informal social media content into clear, concise, and verifiable statements, referred to as normalized claims [73]. These normalized claims capture the core factual assertion, making the fact-checking process more efficient and focused. This task is especially important in low-resource and multilingual settings, where identifying verifiable claims across language boundaries introduces additional complexity.

The task operates in two settings: monolingual and zero-shot. In the monolingual setting, training, development, and test datasets are provided for the same language. The model is trained, validated, and tested exclusively within this single language, allowing it to learn language-specific structures and patterns. Languages included in this setup are: English, German, French, Spanish, Portuguese, Hindi, Marathi, Punjabi, Tamil, Arabic, Thai, Indonesian, and Polish. In contrast, the zero-shot setting provides only the test data for the target language, without any corresponding training or development data. Participants may train their models using data from other languages or conduct zero-shot experiments with large language models (LLMs), evaluating performance on the target language without prior exposure. This setup tests the model's ability to generalize to unseen languages. Languages in this setting are: Dutch, Romanian, Bengali, Telugu, Korean, Greek, and Czech.

3.3 Task 3 Fact-Checking Numerical Claims

Task 3 is addressing the last task of the verification pipeline focusing on numerical claims (cf. Sect. 5.3).

There has been growing interest in developing tools [67], methods [30], and benchmarks [14,64] to enhance the fact-checking process. Automating fact-checking is challenging, as many claims are complex and require sophisticated reasoning for accurate validation, especially those involving numerical data. Numerical claims often appear more credible due to the *Numeric-Truth effect* [58], leading to uncritical acceptance. Recent studies show verifying numerical claims is more difficult than non-numerical ones [9,78]. For example, the social media claim that "CDC quietly deletes 6,000 COVID vaccine deaths from its website" exaggerates a clerical correction, causing unnecessary panic. This demonstrates the need for automated verification of such misleading claims.

This task focuses on verifying claims with numerical quantities and temporal expressions. Numerical claims are defined as those requiring validation of explicit or implicit quantitative or temporal details. Participants must classify each claim as *True, False, or Conflicting* based on a short list of evidence. Each claim is accompanied by the top-100 pieces of evidence retrieved using BM25 from our collection. These evidences can be used after re-ranking to perform claim verification with a classification or generative model that can perform the task of Natural language Inference (NLI). The objective here is to also evaluate the numerical reasoning capabilities of the claim verification model. The task is available in English, Spanish, and Arabic.

3.4 Task 4 Scientific Web Discourse Processing

Scientific web discourse, e.g., discourse about scientific claims or resources on the social web, has increased substantially throughout the past years [19,25]. However, scientific web discourse is usually informal, with examples such as *"covid vaccines just don't work on children"*, and displays fuzzy/incomplete citation habits, such as *"Stanford study shows that vaccines don't work"* where the actual study is never cited through explicit references. This poses challenges both from a computational perspective when mining social media or computing Altmetrics, but also from a societal perspective, leading to poorly informed online debates [53]. Based on this motivation, we introduce two tasks that are related to the second and third task of the verification pipeline:

- **Subtask 4a Scientific Web Discourse Detection**: Given a social media post (tweet), detect if it contains (1) a scientific claim, (2) a reference to a scientific study/publication, or (3) mentions of scientific entities, e.g. a university or scientist.
- **Subtask 4b Claim-source Retrieval**: Given an implicit reference to a scientific paper, i.e., a social media post (tweet) that mentions a research publication without a URL, retrieve the mentioned paper from a pool of candidate papers.

Refer to [33] for a detailed overview of **Task 4**, the dataset, and the participants' approaches.

Table 2. Task 1: Subjectivity in News Articles. Dataset statistics for all five languages for which we report training and development data splits.

	Training Languages									
	Arabic		**Bulgarian**		**English**		**German**		**Italian**	
	obj	**subj**	**obj**	**subj**	**obj**	**subj**	**obj**	**subj**	**obj**	**subj**
Train	1,391	1,055	379	312	532	298	492	308	1,231	382
Dev	266	201	167	139	240	222	317	174	490	177
Dev-test	425	323	134	107	362	122	153	71	334	128
Test	727	309	-	-	215	85	229	118	192	107
Total	2,809	1,888	689	558	1,349	727	1,191	671	2,247	794

	Unseen Languages							
	Greek		**Polish**		**Romanian**		**Ukrainian**	
	obj	**subj**	**obj**	**subj**	**obj**	**subj**	**obj**	**subj**
Test	236	48	161	154	154	52	219	78

4 Datasets

The following section describes the datasets developed for the individual tasks and distributed to the scientific community.[1]

4.1 Task 1: Subjectivity in News Articles

The dataset comprises sentences from news paper articles annotated with respect to their subjectivity. Information regarding the annotation guidelines can be found in [54]. The dataset included 4,697, 1,247, 2,076, 1,862 and 3,041 instances in Arabic (see [75] for more detail), Bulgarian, English, German, and Italian, respectively. Table 2 shows statistics. We provided a training set for the multilingual scenario, comprising the training data for all languages offered this year. The same holds for the dev and dev-test sets. The test set included only data from the languages offered in this edition. The participants were free to choose from the multilingual datasets, opening room for cross-lingual approaches. For the zero-shot setting, the unseen test sets statistics are as follows: 284 instances for Greek (236 OBJ, 48 SUBJ), 351 for Polish (161 OBJ, 154 SUBJ), 206 for Romanian (154 OBJ, 52 SUBJ), and 298 for Ukrainian (219 OBJ, 78 SUBJ).

[1] All datasets are available in the GitLab repository of the lab: https://gitlab.com/checkthat_lab/clef2025-checkthat-lab/.

4.2 Task 2: Claim Normalization

The posts originate from various social media platforms, such as Twitter, Reddit, Facebook, etc., and are sourced from the Google Fact-check Explorer API[2] and the Claim Review Schema.[3] Each post is paired with a corresponding normalized claim. We provide train, dev and test data for Arabic, German, English, French, Hindi, Marathi, Indonesian, Punjabi, Polish, Portuguese, Spanish, Tamil and Thai. While low-resource languages like Bengali, Czech, Greek, Korean, Romanian, Telugu, and Dutch are considered for zero-shot settings. The data statistics are provided in Table 3. The systems are evaluated using the METEOR score.

Table 3. Task 2: Claim Normalization. Dataset statistics for all 20 languages.

Split	Arabic	Bengali	Czech	German	Greek	English	French	Hindi	Korean	Marathi
Train	470	0	0	386	0	11,374	1,174	1,081	0	137
Dev	118	0	0	101	0	1,171	147	50	0	50
Test	100	81	123	100	156	1,285	148	100	274	100
Split	Indonesian	Dutch	Punjabi	Polish	Portugese	Romanian	Spanish	Tamil	Telugu	Thai
Train	540	0	445	163	1,735	0	3,458	102	0	244
Dev	137	0	50	41	223	0	439	50	0	61
Test	100	177	100	100	225	141	439	100	116	100

Example: Claim decomposition example

Claim: `Discretionary spending has increased over 20-some percent in two years if you dont include the stimulus. If you put in the stimulus, its over 80 percent.`

[Decomposition]: [Q1]: Has discretionary spending increased in the past two years?
[Q2]: Does the increase in discretionary spending exclude the stimulus?
[Q3]: Is there evidence to support the claim that

Fig. 2. Example for claim decomposition

4.3 Task 3: Fact-Checking Numerical Claims

The dataset is collected from various fact-checking domains through Google Fact-check Explorer API[4], complete with detailed metadata and an evidence

[2] https://toolbox.google.com/factcheck/apis.
[3] https://schema.org/ClaimReview.
[4] https://toolbox.google.com/factcheck/apis.

Table 4. Dataset statistics for task 3.

Split	English	Spanish	Arabic
Train	9,935	1,506	2,191
Dev	3,084	377	587
Test	3,656	1,806	482

Table 5. Task 4a: Scientific Web Discourse Detection Dataset statistics.

Split	Cat 1	Cat 2	Cat 3	Total
Train	333	224	306	1,229
Dev	26	26	34	137
Test	121	56	97	240
Total	480	306	437	1,606

corpus sourced from the web. Our pipeline filters out numerical claims for the task. An overview of dataset statistics is shown in Table 4. We use the train and validation sets from the English dataset released in [78], and also curate Arabic and Spanish claims. For the test set we collect new real-world English numerical claims additionally to the evaluation set released in [78] to avoid label leakage. The Arabic dataset only consists of claims belonging to the categories *True* and *False* for verification, as real-world distribution of conflicting claims for Arabic is too low.

Evidence for claims in all languages were obtained from search engines by excluding fact-checking websites to avoid leakage of fact-checker justification and verdict. For each claim, we decompose them to yes/no type subquestions as shown in Fig. 2, and issue the original claim and generated subquestions as queries to the search engines. Additionally, for English, evidences are also obtained by other decomposition approaches like sub-claim generation to increase diversity of evidence pool. All evidences are pooled to form the collection. Macro-averaged F1 and classwise F1 scores were employed as metrics for evaluating claim verification.

4.4 Task 4: Scientific Web Discourse Processing (SciWeb)

Task 4a: Scientific Web Discourse Detection. The dataset for subtask 4a is an extension of the SciTweets corpus [34] and consists of 1,606 posts from X (former Twitter) annotated with the different forms of science-related online discourse as introduced in [34], which are scientific claims (Cat 1), scientific references (Cat 2), and references to science contexts or entities (Cat 3). Table 5 shows the dataset statistics.

Task 4b: Scientific Claim Source Retrieval. The dataset for subtask 4b consists of two sets, a query set and a collection set. The query set contains 14,399 X (Twitter) posts with implicit references to scientific papers from CORD-19.

The collection set contains metadata, such as title, abstract, and affiliations of the 7,718 CORD-19 scientific papers, which the posts from the query set implicitly refer to. The dataset is divided into a train (14253 posts), dev (1400 posts), and test split (1446 posts).

5 Results and Overview of the Systems

5.1 Task 1: Subjectivity in News Articles

A total of 21 teams participated in the task, submitting 436 valid runs across all language tracks. 16 out of the 21 teams filled in the survey for the task, providing information about their systems and approaches. 12 teams participated in more than one subtask, while 5 teams opted for only the monolingual English subtask. Table 6 shows the results achieved by the top-3 ranking teams for each language.

Table 6. Task 1: results on subjectivity classification in news articles in terms of macro F1. Shown are the top-3 submissions per language.

Rank	Team	F1	Rank	Team	F1	Rank	Team	F1
Arabic			Italian			German		
1	CEA-LIST	0.6884	1	XplaiNLP	0.8104	1	SmolLab_SEU	0.8520
2	UmuTeam	0.5903	2	CEA-LIST	0.8075	2	UNAM	0.8280
3	Investigators	0.5880	3	SmolLab_SEU	0.7750	3	QU-NLP	0.8013
English			Multilingual			Polish		
1	QU-NLP	0.8052	1	TIFIN INDIA	0.7550	1	CEA-LIST	0.6922
2	TIFIN INDIA	0.7955	2	CEA-LIST	0.7396	2	IIIT Surat	0.6676
3	CEA-LIST	0.7739	3	CSECU-Learners	0.7321	3	CSECU-Learners	0.6558
Ukrainian			Romanian			Greek		
1	CSECU-Learners	0.6424	1	QU-NLP	0.8126	1	AI Wizards	0.5067
2	Investigators	0.6413	2	CSECU-Learners	0.7992	2	SmolLab_SEU	0.4945
3	ClimateSense	0.6395	3	XplaiNLP	0.7917	3	CSECU-Learners	0.4919

Most teams used a supervised binary classification approach, treating the task as classifying sentences into subjective (SUBJ) or objective (OBJ). The dominant strategy involved fine-tuning transformer-based models, with some using ensembles, data augmentation, or additional linguistic features. A few teams explored probabilistic thresholds, embedding-based classifiers, or LLM-based zero-shot and in-context learning methods. An overview of the approaches is given in Table 11 and a short description of the individual approaches for each team is given in the following.

Team **AI Wizards** [29] employed a probabilistic classifier with a decision threshold, fine-tuning DeBERTaV3 for the task.

Team **Investigators** [35] utilized decoder-based models including DeBERTa, BERT, Multilingual BERT, and Twitter RoBERTa.

Table 7. Task 1: Overview of the approaches.

Team	Language									Model															Misc			
	Arabic	Italian	German	English	Multilingual	Polish	Ukrainian	Romanian	Greek	DeBERTa	BERT	MBERT	RoBERTa	DistilRoBERTa	SentimentBERT	ModernBERT	MPNet	XLM-RoBERTa	SBERT	CT-BERT	Electra	InfoXLM	Llama	GPT	Zephyr	Qwen	Data Augmentation	Translating data	LLM Prompting	Feature Selection
AI Wizards [29]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓																				
Investigators [35]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓																	
DSGT-CheckThat [37]				✓									✓	✓	✓	✓			✓								✓			
CSECU-Learners [4]					✓	✓	✓	✓	✓	✓		✓					✓		✓											
CEA-LIST [26]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												✓	✓		✓			✓	
IIIT Surat [39]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓																		
TIFIN INDIA [32]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓			✓		✓									✓	✓		✓
ClimateSense [20]	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓			✓		✓	✓	✓					✓				✓	
CUET_KCRL [69]				✓								✓																		
nlu@utn [45]				✓							✓																			✓
XPlaiNLP [60]		✓	✓	✓	✓	✓	✓	✓	✓		✓							✓						✓					✓	
JU_NLP [23]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓																			✓
NapierNLP [7]				✓																				✓		✓			✓	
UmuTeam [18]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓					✓												✓
UGPLN [21]				✓															✓											✓
SmolLab_SEU [51]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓					✓			✓	✓								
Arcturus [3]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓																				
QU-NLP [5]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓											✓									✓
CheckMates [48]				✓							✓								✓											✓

Team **DSGT-CheckThat** [37] fine-tuned encoder models and explored data augmentation strategies. Their models included RoBERTa (emotion-large), DistilRoBERTa, Sentiment-BERT, ModernBERT, RoBERTa-large, and MiniLM. They further enhanced performance through Synthetic Data Generation and Data Augmentation.

Team **CSECU-Learners** [4] framed the task as multiclass classification with SUBJ (subjective) and OBJ (objective) as separate classes. Their transformer models included MPNet, mDeBERTa, and Multilingual BERT.

Team **CEA-LIST** [26] fine-tuned small language models (SLMs) and experimented with LLMs through techniques such as in-context learning, LLM-as-judge, and model debating. Their models included RoBERTa, UmBERTo, ALBERTo, Qwen 2.5 70B, Meta-LLaMA 3 70B, DeepSeek 67B, Aya-Expanse-32B, and GPT-4.1-mini.

Team **IIIT Surat** [39] employed a transformer-based model, specifically BERT, implemented via BertForSequenceClassification from Hugging Face, and fine-tuned it for binary classification (SUBJ/OBJ). They used the pre-trained BERT (English, uncased) for the monolingual classifier and Multilingual BERT (cased) for multilingual and other-language classification, fine-tuning both directly on the CLEF training data.

Team **TIFIN INDIA** [32] used a binary classification approach, where each input is classified as either subjective or objective. They used an ensemble of transformer-based models and combined their probability outputs to make the final prediction post data augmentation. To mitigate data imbalance, they applied back-translation as a data augmentation technique and used the label distribution ratio to monitor and address class imbalance. They sed deep learning models based on transformer encoder architectures, including BERT-Base, BERT-Large, RoBERTa-Base, RoBERTa-Large, XLM-RoBERTa-Base, XLM-RoBERTa-Large, Modern-BERT-Base, and Modern-BERT-Large. They applied probability-level averaging (soft voting) for model fusion to ensemble predictions across these models. Additionally, for some datasets, they used a traditional Support Vector Machine (SVM) classifier with TF-IDF features as a lightweight baseline and for comparative analysis. They used a feature-based approach using Support Vector Machines (SVMs) on selected datasets. The most important features included: TF-IDF vectors of unigrams and bigrams.

Team **ClimateSense** [20] used Embeddings and an MLP classifier. They experimented with various classifiers: SVC, Logistic Regression, MLP, etc. They also experimented with various transformers-based architectures for embedding the sentences: SBERT, RoBERTa-based models, ModernBERT-large, CT-BERT. Finally, they experimented with Zero-shot prompting some LLMs (such as Zephyr).

Team **CUET_KCRL** [69] pursued a supervised classification approach using an LSTM and fine-tuning mBERT.

Team **nlu@utn** [45] followed a Bert-based ensemble model approach, by also adapting the provided the training data with additional linguistic information before training, using persuasion techniques identified in the data and POS-counts. The models used were politicalBiasBERT and BERT-base-uncased.

Team **XPlaiNLP** [60] employed several transformer-based models, including XLM-RoBERTa-base, GPT o3-mini, and German-BERT. In particular, for monolingual tasks, German-BERT was fine-tuned on German and German-translated versions of English, Italian and Bulgarian train datasets.

Team **JU_NLP** [62] fine-tuned BERT model on available training data, formulating the task as a binary classification problem. In particular, they leverage hand-crafted features derived from knowledge bases and tools like SentiWordNet, WordNet, Opinion lexicon, POS taggers, and lemmatization.

Team **NapierNLP** [7] only tackled the English monolingual task by leveraging LLMs. More precisely, they employed GPT-2, GPTNeo-1.3B, and Qwen3-0.6B. The prompts provided instructions for addressing the task as a binary classification problem.

Team **UmuTeam** [18] employed a wide set of encoder-only transformers, each specific for a given language. In particular, they employed MARBERTv2 for Arabic data, GottBERT-base for German, BERTino for Italian, RoBERTa-base for English. Lastly, they used XLM-RoBERTa-base for multilingual and zero-shot tasks.

Team **UGPLN** [21] employed sentence transformers with hand-crafted linguistic features. A logistic regressor is then trained on top to perform the binary classification task. In particular, they employed MiniLM-L12-v2 and used the following hand-crafted features: presence of negation cues, sentence length (i.e., token count), punctuation marks, and lexical opinion indicators derived from the MPQA Subjectivity lexicon.

Team **SmolLab_SEU** [51] employed a vast set of encoder-only transformers, some of which are language-specific. The models are RoBERTa, DeBERTa-v3, AraBERTv2 and MARBERTv2 for Arabic, GBERT-large, GottBERT-base, and GElectra-large for German, UmBERTo-v1, and BERT-base-italian for Italian, MBERT, XLM-RoBERTa-large, InfoXLM-large, MT5-base, and MDeBERTa-v3 for multilingual. All models were fine-tuned by adding a sequence classification head on top of their pre-trained encoder layers.

Team **Arcturus** [3] fine-tuned the English-pretrained DeBERTa-v3 on monolingual datasets and evaluate it on all languages, including multilingual and zero-shot tasks.

Team **QU-NLP** [5] propose a feature-augmented transformer architecture that combines contextual embeddings from pre-trained language models with statistical and linguistic features. In particular, they employed AraElectra for Arabic, augmented with POS tags and TF-IDF features. For cross-lingual experiments, they employed DeBERTa-v3 with TF-IDF features through a gating mechanism.

Team **CheckMates** [48] explored various models such as logistic regression, Support Vector Machine, BERT, Sentence-BERT, and DistilBERT.

More details on the participants approaches can be found in [55].

5.2 Task 2: Claim Normalization

Task 2 received submissions from 18 teams, totalling 1,226 valid runs across all the languages. Table 9 presents the results for the monolingual setup, evaluated using METEOR scores, while Table 10 reports the outcomes for the zero-shot setup. These tables summarise team performance across the respective languages.

Most teams employed sequence-to-sequence generation strategies for claim normalization, typically relying on transformer-based models. The most prevalent approach involved fine-tuning pretrained models such as BART, T5, mBART, and LLaMA on monolingual training data. Common data preprocessing included de-duplication, emoji removal, hashtag normalization, multilingual data augmentation via translation, and prompt engineering tailored to each language. Some of the teams used LoRA-based adaptor fine-tuning to reduce resource needs, while others delved into ensemble solutions like embedding-based centroid voting or model-soup techniques. In the zero-shot setting, prompt-based generation took precedence, with models driven by structured instructions to extract factual, brief claims from informal posts. Others employed semantic similarity retrieval to choose in-context instances for prompting. To improve

Table 8. Task 2: Overview of the participating teams per language and their respective rankings.

Team	English	Arabic	German	French	Hindi	Marathi	Indonesian	Punjabi	Polish	Portugese	Spanish	Tamil	Thai	Bengali	Telugu	Dutch	Czech	Greek	Romanian	Korean
dfkinit2b [12]	1	1	2	2	1	1	2	1	2	2	2	1	3	1	1	1	1	1	1	1
DS@GT CheckThat! [50]	2	2	1	1	2	4	1	5	1	1	1	3	1	4	5	5	3	4	4	3
TIFIN [68]	3	5	5	6	7	6		4	5		5	5		5	6	4				
AKCIT-FN [8]	4	6	3	3	5	5	3	2	3	3	3	2	2	3	2	2	4	2	2	2
Factiverse [10]	5	7	4	4	8	9	4	7	6	6	6	8	4	6	7		5	5	5	
rohan_shankar	6																			
manan-tifin	7			7	9	7			5					5	6					
MMA [57]	8	3	7	8	6	3	5	6	7	4	4	6								
UNH [81]	9																			
Investigators [35]	10										8									5
teamopenfact [63]	11	4	6	5	4	2	6	3	4	5	7	4	5	2	3	3	2	3	3	4
Nikhil_Kadapala	12																			
aryasuneesh	13		5	6	7	6		4			5	5	6							
JUNLP_M&S [46]	14																			
uhh_dem4ai	15																			
tomasbernal01	16	8	8	9	10	8	7	8	8	7	9	7	7	7	8	6	6	6	6	6
VSE	17																			
saivineetha [15]					3										4					

the relevance and structure of claims, some teams used reinforcement learning, instruction tuning, or Chain-of-Thought prompting.

Out of all the participating teams, dfkinit2b [12], DS@GT CheckThat! [50], TIFIN [68], and AKCIT-FN [8] consistently ranked among the top-performers across most languages. More details on the other participating approaches can be found in [74].

Team **dfkinit2b** [12] performed comprehensive experiments in both monolingual and zero-shot settings, testing zero- and few-shot prompting with models like Gemma-3, Qwen-3, Qwen-2.5, Llama-3.3, and Mistral. They explored various prompt formulations and used cosine similarity to select demonstrations for few-shot learning. Experiments also included adapter fine-tuning, data preprocessing with language checks and emoji removal, and data augmentation via translation. For the final submission, they ensembled top-performing model outputs by computing embedding centroids with multilingual SentenceTransformers and selecting claims closest to these centroids, achieving strong results across languages.

Table 9. Task 2: Scores (METEOR) for languages with training data. Ranks across languages are in brackets.

Team	English	Arabic	German	French	Hindi	Marathi	Thai
dfkinit2b	0.4569 (1)	0.5037 (1)	0.3469 (2)	0.4703 (2)	0.3275 (1)	0.3888 (1)	0.2999 (3)
DS@GT CheckThat!	0.4521 (2)	0.5035 (2)	0.3859 (1)	0.5273 (1)	0.3001 (2)	0.2608 (4)	0.5859 (1)
TIFIN	0.4114 (3)	0.3705 (5)	0.2642 (5)	0.3441 (6)	0.2604 (7)	0.1521 (6)	-
AKCIT-FN	0.4058 (4)	0.3277 (6)	0.2652 (3)	0.3811 (3)	0.2706 (5)	0.2181 (5)	0.3179 (2)
Factiverse	0.4049 (5)	0.2457 (7)	0.2644 (4)	0.3750 (4)	0.2125 (8)	0.0847 (9)	0.0965 (4)
rohan_shankar	0.3920 (6)	-	-	-	-	-	-
manan-tifin	0.3881 (7)	-	-	0.2768 (7)	0.2080 (9)	0.1230 (7)	-
MMA	0.3841 (8)	0.4584 (3)	0.1556 (7)	0.2469 (8)	0.2641 (6)	0.2793 (3)	-
UNH	0.3737 (9)	-	-	-	-	-	-
Investigators	0.3565 (10)	-	-	-	-	-	-
teamopenfact	0.3370 (11)	0.4175 (4)	0.2319 (6)	0.3605 (5)	0.2722 (4)	0.3048 (2)	0.0872 (5)
Nikhil_Kadapala	0.3321 (12)	-	-	-	-	-	-
aryasuneesh	0.3153 (13)	-	0.2642 (5)	0.3441 (6)	0.2604 (7)	0.1521 (6)	0.0464 (6)
JUNLP_M&S	0.3098 (14)	-	-	-	-	-	-
uhh_dem4ai	0.2612 (15)	-	-	-	-	-	-
tomasbernal01	0.1660 (16)	0.0003 (8)	0.1039 (8)	0.1649 (9)	0.0132 (10)	0.0877 (8)	0.0147 (7)
VSE	0.0070 (17)	-	-	-	-	-	-

Team	Indonesian	Punjabi	Polish	Portugese	Spanish	Tamil
dfkinit2b	0.5021 (2)	0.3307 (1)	0.3961 (2)	0.5744 (2)	0.5539 (2)	0.6316 (1)
DS@GT CheckThat!	0.5650 (1)	0.2567 (5)	0.4065 (1)	0.5770 (1)	0.6077 (1)	0.4702 (3)
TIFIN	-	0.2685 (4)	0.2331 (5)	-	0.3906 (5)	0.3676 (5)
AKCIT-FN	0.3866 (3)	0.3038 (2)	0.2798 (3)	0.5290 (3)	0.5213 (3)	0.5197 (2)
Factiverse	0.3099 (4)	0.1251 (7)	0.1964 (6)	0.3381 (6)	0.3821 (6)	0.0043 (8)
manan-tifin	-	-	0.2331 (5)	-	-	-
MMA	0.3089 (5)	0.1834 (6)	0.1243 (7)	0.4719 (4)	0.5094 (4)	0.3468 (6)
Investigators	-	-	-	-	0.3447 (8)	-
teamopenfact	0.2445 (6)	0.2696 (3)	0.2666 (4)	0.3779 (5)	0.3710 (7)	0.4681 (4)
aryasuneesh	-	0.2685 (4)	-	-	0.3906 (5)	0.3676 (5)
tomasbernal01	0.1305 (7)	0.0097 (8)	0.0742 (8)	0.1898 (7)	0.2048 (9)	0.0196 (7)

Team **DS@GT CheckThat!** [50] embedded the non-normalized claims from the pooled train and development datasets, as well as from the test set, using state-of-the-art embedding models tailored to each language. For every test claim, a GPT-4o mini model was prompted following the approach discussed in [73], utilising the top-3 most cosine-similar examples from the train and development sets as in-context examples. The final response for the monolingual task was derived by combining the best-matching answer from the train and development sets, based on cosine similarity, and the output of the GPT-4 model. For

Table 10. Task 2: Scores (METEOR) for languages without training data. Ranks across languages are in brackets.

Team Name	Bengali	Telugu	Dutch	Czech	Greek	Romanian	Korean
dfkinit2b	0.3777 (1)	0.5257 (1)	0.2001 (1)	0.2519 (1)	0.2619 (1)	0.2950 (1)	0.1339 (1)
teamopenfact	0.2959 (2)	0.4559 (3)	0.1866 (3)	0.2144 (2)	0.2333 (3)	0.2350 (3)	0.1050 (4)
AKCIT-FN	0.2916 (3)	0.5176 (2)	0.1922 (2)	0.1734 (4)	0.2567 (2)	0.2516 (2)	0.1209 (2)
DS@GT CheckThat!	0.2435 (4)	0.3171 (5)	0.1608 (5)	0.1959 (3)	0.2250 (4)	0.2220 (4)	0.1156 (3)
TIFIN	0.2030 (5)	0.2502 (6)	0.1720 (4)	-	-	-	-
manan-tifin	0.2030 (5)	0.2502 (6)	-	-	-	-	-
Factiverse	0.1068 (6)	0.0802 (7)	-	0.1571 (5)	0.1455 (5)	0.2097 (5)	-
tomasbernal01	0.0451 (7)	0.0269 (8)	0.0817 (6)	0.0544 (6)	0.0062 (6)	0.0779 (6)	0.0014 (6)
Investigators	-	-	-	-	-	-	0.0149 (5)

Table 11. Task 3: Overview of the approaches for fact-checking numerical claims.

Team	**Language**			**Model**											**Macro-F1**		
	Arabic	**Spanish**	**English**	**BM25**	**cross-encoder**	**gpt-4o-mini**	**Qwen**	**Llama**	**DeepSeek**	**ModernBERT**	**Math-Roberta**	**RoBERTa-base**	**QWQ-32B**	**Qwen-8B**	**Deberta-Large-MNLI**	**mxbai-rerank-large-v1**	**granite-3.3-8b-instruct**	**Arabic**	**Spanish**	**English**
LIS [42]	☑	☑	☑										☑					**50.34**	**96.15**	**59.54**
DS@GT-CheckThat! [38]			☑															-	-	52.10
TIFIN [36]	☑		☑	☑											☑		☑	55.36		55.70
ClaimIQ [11]			☑					☑										-	-	42.43
FraunhoferSIT [56]			☑		☑			☑				☑						-	-	51.00
NGU_Research [1]	☑	☑		☑	☑				☑									63.52	24.41	-
JU_NLP [22]	☑		☑	☑	☑													36.38	-	48.83
CornellNLP [24]			☑	☑		☑		☑										-	-	48.57
UGLPN [77]			☑	☑								☑						-	-	45.53
UCOM_UNAM_PLN [2]		☑		☑		☑												-	35.95	-
News-polygraph*		☑		☑	☑				☑									-	-	42.86

zero-shot languages, they utilised a modified version of CACN [73], essentially using the prompting method with standard examples.

Team **TIFIN** [68] fine-tuned the Qwen-14B model using LoRA with 4-bit precision for efficiency. They preprocessed data by filtering meaningful post-claim pairs, removing duplicates, and creating a unified multilingual dataset. Instruction-based fine-tuning incorporated Chain-of-Thought prompting with 5W1H questions to guide claim extraction. During inference, context resolution replaced partial posts with complete ones, and few-shot prompting with

similar examples improved claim structure. This approach aimed to boost claim extraction accuracy and multilingual performance.

Team **AKCIT-FN** [8] fine-tuned sequence-to-sequence Transformer models, including monolingual and multilingual variants like unicamp-dl/monoptt5-large, unicamp-dl/ptt5-v2-large, and t5-large, focusing on monolingual training to capture language-specific features. They performed hyperparameter tuning across learning rates, optimizers, batch sizes, and epochs to optimize performance. Evaluation combined METEOR, BERTScore, and semantic similarity metrics to better assess claim fidelity. For zero-shot tasks, they leveraged OpenAI's LLMs with carefully crafted prompts to generate concise, factual claims from informal posts in unseen languages, testing the models' generalization capabilities.

5.3 Task 3: Fact-Checking Numerical Claims

A total of 258 valid runs were submitted by 13 unique teams across languages, with 4 participants in Spanish and Arabic. 11 teams participated in fact-checking English numerical claims.

Among all participating teams, LIS was the top performer across all languages. TIFIN, NGU_Research, DS@GT-CheckThat! performed well in the respective languages the teams participated. Most teams employed generative models like gpt-40-mini or Qwen LLMs to decompose claims, followed by BM25 based retrieval for retrieving evidence and transformer based cross-encoder models for re-ranking the evidences. For claim verification fine-tuned transformer based NLI models were employed by some teams where transformers were trained as discriminative models on the training sets provided. Some teams employed prompting based approaches to leverage large Language Models (LLMs) like gpt-4o-mini or reasoning models like deepseek-r1 to perform claim verification. The authors observe that fine-tuning LLMs for claim verification coupled with claim decomposition using recent reasoning models helps outperform the best baselines reported in [78].

Team **LIS** [42] used QwQ-32B to generate question followed by Linq-Embed-Mistral to retrieve evidence from the corpus by combining the questions and claims. Mistral-Small-24B-Instruct-2501 was fine-tuned to obtain the final veracity labels. The Qwen model seem to overcome certain limitations associated with gpt-3.5 and gpt-4 series models used in baselines [78].

Team **DS@GT-CheckThat!** [38] performed pre-processing to normalize the number and dates of the claims and decomposed questions from these claims. They employed gpt-4o-mini to decompose the claims. BM25 was employed for first stage retrieval to prioritize documents relevant to the claim and sub-questions. This is followed by re-ranking the documents using cross-encoder/ms-marco-MiniLM-L-12-v2 or mixedbread-ai/mxbai-rerank-large-v1. The main workhorse model for the veracity classification was ModernBERT - an optimized model based on the BERT architecture, that can natively support longer sequence length.

Team **TIFIN** [36] employed inverse class weighting to handle class imbalance in claim verification step and to give more importance to minority classes. they also employed other strategies such as oversampling to balance training examples, and label smoothing to prevent the model from becoming overconfident in its predictions. The authors also incorporated Focal Loss, for fine-tuning the verification model *microsoft/deberta-large-mnl* with LoRA, to focus training on the harder examples. They also employed *ibm-granite/granite-3.3-8b-instruct model* to summarize contexts before feeding them to the verification model.

Team **NGU_Research** [1] employed hybrid retrieval techniques ranging from pretrained encoder-based models to BGE, E5, Gemini as embedding models and finally settled on pretrained embeddings from openai's **text-embedding-3-large** model together with bm25 filtering via Qdrant database collections for each language. Then finally Deepseek and gpt-4o-mini were employed for performing claim verification using the retrieved evidence.

5.4 Task 4: Scientific Web Discourse Processing (SciWeb)

Task 4a: Scientific Web Discourse Detection. Task 4a is a multilabel classification task and was evaluated through the macro-averaged F1-score. The baseline is a DeBERTaV3-base model trained on the train set for 10 epochs with a learning rate of $2e^{-5}$ and a batch size of 16. For the final test set predictions, we used the checkpoint with the best dev set performance, resulting in a test set macro F1-score of 0.7668 (rank 7).

In total, ten teams participated in subtask 4a. Table 12 provides an overview of the different approaches and their performances for those teams that submitted a paper description of their work. The F1-score and rank indicate the performance and position on the final test set leaderboard. Most teams relied on Transformer-based models such as DeBERTa-v3, SciBERT, and Twitter-Roberta, while team DS@GT CheckThat! [49] and TurQUaz [61] also used LLMs. In addition, different techniques such as LLM-based data augmentation, ensemble methods, and other optimizations were employed.

Team **ClimateSense** [20] fine-tuned a twitter-roberta-base-2022-154 m model with a weighted loss. For each category, the best-performing checkpoint was identified based on the dev set performance. Using the embeddings of these checkpoints, a traditional classifier was trained for each category.

Team **UTB-CEDNAV** [70] fine-tuned a DeBERTa-v3-base model using hyperparameters found with 5-fold cross-validation. To improve performance, they employed class weighting and threshold-tuning and used an ensemble of their two strongest model (with and without class weight) for their final submission.

Team **SBU-SCIRE** [76] augmented the training data with paraphrases using DeepSeek-R1. They trained a DeBERTa-v3-large model with a Focal Loss on the train set and performed a grid search over the per-class threshold to maximize the performance on the dev set.

Team **DS@GT** [49] trained different transformer-based models and used zero-shot and few-shot classification with GPT-4o and GPT-4o mini. For their

final submission, they used DeBERTa-v3-base for categories one and three and GPT-4o mini with few-shot (five examples based on semantic similarity) for category two.

Team **TurQUaz** [61] employed various LLMs in different collaborative settings. The setting for their final submission includes five models discussing a post together to reach an agreement, with another model acting as a chairperson.

Team **JU_NLP** [43] generated tweet embeddings using SciBERT and Twitter-RoBERTa models to capture both scientific and social media discourse characteristics of tweets. The embeddings were used to train a two-layer classification head.

Task 4b: Scientific Claim Source Retrieval. Task 4b is a retrieval task and was evaluated by the MRR@5 (Mean Reciprocal Rank) score. BM25 ranking using the title and abstract of the papers and the text of the X posts serves as the baseline with an MMR@5 of 0.43. The best-performing team reached an MMR@5 of 0.68.

Table 12. Task 4a: Overview of the approaches

Team	Models				Misc.			Perf.	
	DeBERTa-v3	SciBERT	Twitter-RoBERTa	LLMs	Others	Data Augmentation	Ensemble	Other Optimizations	Macro-avg. F1-Score	Rank
ClimateSense [20]			■		■		■		0.7998	1
UTB-CEDNAV [70]	■						■	■	0.7983	2
SBU-SCIRE [76]	■					■		■	0.7917	4
DS@GT CheckThat! [49]	■			■	■		■		0.7685	6
DeBERTa-v3 Baseline	■								0.7668	7
TurQUaz [61]				■			■		0.7615	8
JU_NLP [43]		■	■				■		0.7347	9

In total, 30 teams participated in subtask 4b. Table 13 provides an overview of the different approaches and their performance for teams that submitted a paper description of their work. Most teams relied on a combination of retrieval methods (dense, sparse, or both) and re-ranking models. Retrieval methods included both lexical and semantic methods. LLMs such as ChatGPT, LLaMa

and Gemma were mainly used as re-rankers, but did not always outperform fine-tuned Transformer-based models. Additionally, some teams experimented with data augmentation and style transfer techniques.

Team **AIRwaves** [13] employed a two-stage pipeline using neural representation learning for candidate generation with a fine-tuned E5-large model, followed by neural re-ranking with a SciBERT cross-encoder to re-order the top predictions.

Team **Deep Retrieval** [59] combined lexical BM25-based and semantic search-based approaches to generate candidates, which were re-ranked using LLMs.

Team **ATOM** [71] used a GTR-T5-Large model to retrieve candidates, followed by a neural re-ranking with MXBAI-base-v2.

Team **SBU-SCIRE** [76] used a Snowflake/snowflake-arctic-embed-l-v2.0 for dense retrieval, followed by ms-marco-MiniLM-L4-v2 re-ranking.

Team **SeRRa** [44] used a multi-step pipeline including dense retrieval for candidate generation with a Sentence-BERT model, re-ranking using a binary classification model, and a final ranking through pairwise comparisons of the top 10 re-ranked documents with the input claim.

Team **Claim2Source** [66] first applied style transfer techniques to both claims and source documents using LLaMa 3.3-70B-Instruct (e.g., enhancing readability, adopting a scientific tone, or reformulating the abstract as a tweet). They then combined BM25 with dense retrieval models such as SPECTER, all-Mini-LM-L6-v2, and GritLM-7B.

Table 13. Task 4b: Overview of the approaches

Team	Models				Misc.		Perf.	
	Dense Retrieval	Sparse Retrieval	Re-ranking	LLMs	Data Augmentation	Style transfer	MRR@5	Rank
AIRwaves [13]		■	■				0.67	2
Deep Retrieval [59]	■	■	■	■			0.66	3
ATOM [71]	■		■				0.66	4
SBU-SCIRE [76]	■		■	■			0.65	5
SeRRa [44]	■			■			0.61	8
Claim2Source [66]	■	■		■		■	0.59	12
DS@GT [65]		■	■	■	■	■	0.58	16
BM25 Baseline		■					0.43	28

Team **DS@GT** [65] used data-augmentation and style transfer techniques on tweets using ChatGPT. They then implemented a two-stage retrieval pipeline based on bi-encoder and cross-encoder approaches for retrieval and reranking using zero-shot and fine-tuned Sentence-Transformers.

6 Conclusion and Future Work

This paper presents the eighth edition of the CheckThat!, one of the most popular labs at CLEF 2025. This year, 177 teams registered, of which 83 actively participated and 54 submitted working notes. The number of languages covered also increased to 20, spanning four tasks—surpassing all previous years and establishing a truly multilingual task setup.

In this edition, Task 1 focused on predicting the subjectivity or objectivity of sentences; Task 2 addressed claim normalization; Task 3 targeted numerical factual claims; and Task 4 examined scientific web discourse. Among the tasks, Task 1 was particularly popular, with a total of 21 teams participating. Most teams relied on fine-tuning transformer models for binary classification. Some teams also utilized and fine-tuned Large Language Models (LLMs) such as Llama and Qwen. For the claim normalization task (Task 2), most teams employed sequence-to-sequence generation approaches. This task received participation from 18 teams, with English attracting the highest number of participants. In Task 3, which focused on fact-checking numerical claims, 13 teams took part. Most systems used LLMs to decompose the claims and employed BM25 for retrieval. For the scientific web discourse detection task (Task 4a), ten teams participated, primarily using transformer-based models. In Task 4b, which focused on claim source retrieval, 30 teams participated and predominantly used various sparse and dense retrieval-based approaches.

Acknowledgments. The work of F. Alam is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund, part of Qatar Research Development and Innovation Council (QRDI). The work of J. Struß is partially supported by the BMBF (German Federal Ministry of Education and Research) under the grant no. 01FP20031J. The work of Stefan Dietze, Konstantin Todorov, Salim Hafid and Sebastian Schellhammer is partially funded under the AI4Sci grant, co-funded by MESRI (France, grant UM-211745), BMBF (Germany, grant 01IS21086), and the French National Research Agency (ANR). The responsibility for the contents of this publication lies with the authors.

References

1. Abdallah, M.A., Fekry, R.M., El-Beltagy, S.R.: NGU_Research at CheckThat! 2025: an LLM based hybrid fact-checking pipeline for numerical claims. In: Faggioli et al. (2025)
2. Acosta, G., Morales, E., Gómez-Adorno, H.: UCOM_UNAM_PLN at Checkthat 2025: evaluating LLMs in a two-step architecture for numerical fact checking. In: Faggioli et al. (2025)

3. Aditya, A., Jambulkar, R., Pal, S.: Arcturus at CheckThat! 2025: Deberta-v3-base for multilingual subjectivity detection in news articles. In: Faggioli et al. (2025)
4. Ahmad, M., Chy, A.N.: CSECU-Learners at CheckThat! 2025: multilingual transformer-based approach for subjectivity detection in news articles across multilingual and zero-shot settings. In: Faggioli et al. (2025)
5. Al-Smadi, M.: QU-NLP at CheckThat! 2025: Multilingual subjectivity in news articles detection using feature-augmented transformer models with sequential cross-lingual fine-tuning. In: Faggioli et al. (2025)
6. Alam, F., et al.: The CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval. In: Hauff, C., et al. (eds.) Advances in Information Retrieval, pp. 467–478. Springer Nature Switzerland, Cham (2025)
7. Alexander, K., Ullah, M.Z., Gkatzia, D.: NapierNLP at CheckThat! 2025: detecting subjectivity with LLMs and model fusion. In: Faggioli et al. (2025)
8. Almada, F.L.N., et al.: Akcit-FN at CheckThat!2025: switching fine-tuned SLMs and LLM prompting for multilingual claim normalization. In: Faggioli et al. (2025)
9. Aly, R., et al.: FEVEROUS: fact extraction and verification over unstructured and structured information. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual (2021)
10. Amatya, P., Setty, V.: Factiverse and IAI at CheckThat! 2025: adaptive ICL for claim extraction. In: Faggioli et al. (2025)
11. Anik, A.S., Chowdhury, M.F.K., Wyckoff, A., Choudhury, S.R.: ClaimIQ at CheckThat! 2025: comparing prompted and fine-tuned language models for verifying numerical claims. In: Faggioli et al. (2025)
12. Anikina, T., et al.: dfkinit2b at CheckThat! 2025: leveraging LLMs and ensemble of methods for multilingual claim normalization. In: Faggioli et al. (2025)
13. Ashbaugh, C., Baumgärtner, L., Greß, T., Sidorov, N., Werner, D.: AIRwaves at CheckThat! 2025: retrieving scientific sources for implicit claims on social media with dual encoders and neural re-ranking. In: Faggioli et al. (2025)
14. Augenstein, I., et al.: MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims. 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), pp. 4685–4697. Association for Computational Linguistics, Hong Kong, China (2019)
15. Baddepudi Venkata Naga Sri, S.V.: Saivineetha at CheckThat! 2025: exploring fine-tuning and zero-shot approaches for claim normalization. In: Faggioli et al. (2025)
16. Barrón-Cedeño, A., et al.: Overview of the CLEF–2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source. In: Arampatzis, A., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023) (2023)
17. Barrón-Cedeño, A., et al.: Overview of CheckThat! 2020: automatic identification and verification of claims in social media. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 215–236. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_17
18. Beltrán, T.B., Pan, R., Díaz, J.A.G., García, R.V.: UmuTeam at CheckThat! 2025: language-specific versus multilingual models for fact-checking. In: Faggioli et al. (2025)

19. Brüggemann, M., Lörcher, I., Walter, S.: Post-normal science communication: exploring the blurring boundaries of science and journalism. J. Sci. Commun. **19**(3), A02 (2020)
20. Burel, G., Lisena, P., Daga, E., Troncy, R., Alani, H.: ClimateSense at CheckThat! 2025: combining fine-tuned large language models and conventional machine learning models for subjectivity and scientific web discourse analysis. In: Faggioli et al. (2025)
21. del Carmen Toapanta-Bernabé, M., Ángel Garcia-Cumbreras, M., Ureña-López, L.A., Intriago, D.D.M., Holguín-Reyes, J.S.: Sinai-UGPLN at CheckThat! 2025: a hybrid sbert–logistic regression framework for segment-level subjectivity detection in English news. In: Faggioli et al. (2025)
22. Das, R., Pal, P., Das, D.: JU_NLP at CheckThat! 2025: utilizing transformer models to fact-check numerical claims. In: Faggioli et al. (2025)
23. Debnath, S., Das, D.: JU_NLP at CheckThat! 2025: a confidence-guided transformer-based approach for multilingual subjectivity classification. In: Faggioli et al. (2025)
24. Duesterwald, L., Arora, A., Cardie, C.: CornellNLP at CheckThat! 2025:hybrid llama–GPT-4 ensembles with confidence filtering for numerical claim verification. In: Faggioli et al. (2025)
25. Dunwoody, S.: Science journalism: prospects in the digital age. In: Routledge Handbook of Public Communication of Science and Technology, pp. 14–32. Routledge (2021)
26. Elbouanani, A., Dufraisse, E., Tuo, A., Popescu, A.: CEA-LIST at CheckThat! 2025: evaluating LLMs as detectors of bias and opinion in text. In: Faggioli et al. (2025)
27. Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.): Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. CLEF 2024 (2024)
28. Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.): Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum. CLEF 2025, Madrid, Spain (2025)
29. Fasulo, M., Babboni, L., Tedeschini, L.: AI Wizards at CheckThat! 2025: enhancing transformer-based embeddings with sentiment for subjectivity detection in news articles. In: Faggioli et al. (2025)
30. Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Trans. Assoc. Comput. Linguist. **10**, 178–206 (2022)
31. Gupta, S., Singh, P., Sundriyal, M., Akhtar, M.S., Chakraborty, T.: Lesa: linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 3178–3188 (2021)
32. Gurumurthy, K., et al.: TIFIN at CheckThat! 2025: cross-lingual subjectivity classification in news through monolingual, multilingual, and zero-shot learning. In: Faggioli et al. (2025)
33. Hafid, S., et al.: Overview of the CLEF-2025 CheckThat! lab task 4 on scientific web discourse. In: Faggioli et al. (2025)
34. Hafid, S., Schellhammer, S., Bringay, S., Todorov, K., Dietze, S.: Scitweets-a dataset and annotation framework for detecting scientific online discourse. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 3988–3992 (2022)

35. Hashmi, S.M.A., Aamir, S., Anas, M., Usmani, T., Alvi, F., Samad, A.: Investigators at CheckThat! 2025: using LLMs to improve fact-checking. In: Faggioli et al. (2025)
36. Hazarika, B., et al.: TIFIN at CheckThat! 2025: X-VERIFY - multi-lingual nli-based fact checking with condensed evidence. In: Faggioli et al. (2025)
37. Heil, M., Bang, D.: DS@GT at CheckThat! 2025: detecting subjectivity via transfer-learning and corrective data augmentation. In: Faggioli et al. (2025)
38. Heil, M., Pramov, A.: DS@GT at CheckThat! 2025: evaluating context and tokenization strategies for numerical fact verification. In: Faggioli et al. (2025)
39. Jaiswal, S.C., Kumar, R.: IIIT Surat at CheckThat! 2025: identifying subjectivity from multilingual text sequence. In: Faggioli et al. (2025)
40. Jerônimo, C.L.M., Marinho, L.B., Campelo, C.E.C., Veloso, A., da Costa Melo, A.S.: Fake news classification based on subjective language. In: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, pp. 15–24 (2019)
41. Kasnesis, P., Toumanidis, L., Patrikakis, C.Z.: Combating fake news with transformers: a comparative analysis of stance detection and subjectivity analysis. Inf. **12**(10), 409 (2021)
42. Le, Q.T., Badache, I., Yacoub, A., Hamri, M.E.A.: LIS at CheckThat! 2025: multi-stage open-source large language models for fact-checking numerical claims. In: Faggioli et al. (2025)
43. Majumdar, A., Das, D., Pal, P.: JU_NLP at CheckThat! 2025: leveraging hybrid embeddings for multi-label classification in scientific social media discourse. In: Faggioli et al. (2025)
44. Marchetti, G., Rocha, G., Cardoso, H.L.: Team SeRRa at CheckThat! 2025: Sequential re-ranking in a scientific claim source retrieval pipeline. In: Faggioli et al. (2025)
45. Meyer, S., Roth, M.: nlu@utn at CheckThat! 2025: combining bias sensitivity, linguistic features, and persuasion cues in an ensemble for subjectivity detection. In: Faggioli et al. (2025)
46. Mondal, M., Saha, S., Saha, D., Das, D.: JU_NLP@M&S at CheckThat! 2025: automated claim extraction and normalization for misinformation detection in social media content. In: Faggioli et al. (2025)
47. Nakov, P., et al.: Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Candan, K., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association. LNCS (12880) (2021)
48. Padmashri, R., V., K., Srikumar, V., Thenmozhi, D.: CheckMates At CheckThat! 2025: transformer-based models for subjectivity classification. In: Faggioli et al. (2025)
49. Parikh, A., Truong, H., Schofield, J., Heil, M.: DS@GT at CheckThat! 2025: ensemble methods for detection of scientific discourse on social media. In: Faggioli et al. (2025)
50. Pramov, A., Ma, J., Patel, B.: DS@GT at CheckThat! 2025: a simple retrieval-first, LLM-backed framework for claim normalization. In: Faggioli et al. (2025)
51. Rahman, M.A., Amin, M.A., Dewan, M.S., Hasan, M.J., Rahman, M.A.: SmolLab_SEU at CheckThat! 2025: how well do multilingual transformers transfer across news domains for cross-lingual subjectivity detection? In: Faggioli et al. (2025)

52. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 105–112. EMNLP '03 (2003)
53. Rocha, Y.M., de Moura, G.A., Desidério, G.A., de Oliveira, C.H., Lourenço, F.D., de Figueiredo Nicolete, L.D.: The impact of fake news on social media and its influence on health during the covid-19 pandemic: a systematic review. J. Public Health, 1–10 (2021)
54. Ruggeri, F., Antici, F., Galassi, A., Korre, K., Muti, A., Barrón-Cedeño, A.: On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection. In: Proceedings the Sixth Workshop on Narrative Extraction From Texts (at ECIR), pp. 103–111 (2023)
55. Ruggeri, F., et al.: Overview of the CLEF-2025 CheckThat! lab task 1 on subjectivity in news article. In: Faggioli et al. (2025)
56. Runewicz, A., Ranly, P.M., Vogel, I., Steinebach, M.: Fraunhofer SIT at CheckThat! 2025: multi-instance evidence pooling for numerical claim verification. In: Faggioli et al. (2025)
57. Saeed, M., Yasser, M., Torki, M., Elmakky, N.: MMA at CheckThat! 2025: multilingual claim normalization of social-media posts. In: Faggioli et al. (2025)
58. Sagara, N.: Consumer understanding and use of numeric information in product claims. University of Oregon (2009)
59. Sager, P.J., Kamaraj, A., Grewe, B.F., Stadelmann, T.: Deep Retrieval at CheckThat! 2025: identifying scientific papers from implicit social media mentions via hybrid retrieval and re-ranking. In: Faggioli et al. (2025)
60. Sahitaj, A., Li, J., Neves, P.W., Fedor Splitt, P.S., Jakob, C., Solopova, V., Schmitt, V.: XplaiNLP at CheckThat! 2025: multilingual subjectivity detection with fine-tuned transformers and prompt-based inference with large language models. In: Faggioli et al. (2025)
61. Saraç, T., Mergen, S., Kutlu, M.: TurQUaz at CheckThat! 2025: debating large language models for scientific web discourse detection. In: Faggioli et al. (2025)
62. Sardar, A.A.M., Fatema, K., Islam, M.A.: JUNLP at CheckThat! 2024: enhancing check-worthiness and subjectivity detection through model optimization. In: Faggioli et al. (2025)
63. Sawiński, M., Węcel, K., Księżniak, E.: OpenFact at CheckThat! 2025: application of self-reflecting and reasoning LLMs for fact-checking claim normalization. In: Faggioli et al. (2025)
64. Schlichtkrull, M., Guo, Z., Vlachos, A.: AVeriTeC: a dataset for real-world claim verification with evidence from the web. arXiv preprint arXiv:2305.13117 (2023)
65. Schofield, J., Tian, S., Truong, H., Heil, M.: DS@GT at CheckThat! 2025: exploring retrieval and reranking pipelines for scientific claim source retrieval on social media discourse. In: Faggioli et al. (2025)
66. Schreieder, T., Färber, M.: Claim2Source at CheckThat! 2025: zero-shot style transfer for scientific claim-source retrieval. In: Faggioli et al. (2025)
67. Setty, V.: Factcheck editor: multilingual text editor with end-to-end fact-checking. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2744–2748 (2024)
68. Sharma, M., et al.: TIFIN at CheckThat! 2025: reasoning-guided claim normalization for noisy multilingual social media posts. In: Faggioli et al. (2025)
69. Shawon, M.T.A., Haq, F., Mia, M.A., Mursalin, G.S.M., Khan, M.I.: CUET_KCRL at CheckThat!2025: ensemblenet with roberta-large for subjectivity detection in news articles. In: Faggioli et al. (2025)

70. Sosa, M., Serrano, J., Martinez Santos, J.C., Puertas, E.: VerbaNexAI Lab at CheckThat! 2025: fine-tuning DeBERTa for multi-label scientific discourse detection in tweets. In: Faggioli et al. (2025)
71. Staudinger, M., El-Ebshihy, A., Kusa, W., Piroi, F., Hanbury, A.: ATOM at CheckThat! 2025: retrieve the implicit - scientific evidence retrieval. In: Faggioli et al. (2025)
72. Struß, J.M., et al.: Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles. In: Faggioli et al. (2024)
73. Sundriyal, M., Chakraborty, T., Nakov, P.: From chaos to clarity: claim normalization to empower fact-checking. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 6594–6609. Association for Computational Linguistics, Singapore (2023). https://aclanthology.org/2023.findings-emnlp.439/
74. Sundriyal, M., Chakraborty, T., Nakov, P.: Overview of the CLEF-2025 CheckThat! lab task 2 on claim normalization. In: Faggioli et al. (2025)
75. Suwaileh, R., Hasanain, M., Hubail, F., Zaghouani, W., Alam, F.: ThatiAR: subjectivity detection in Arabic news sentences. arXiv:2406.05559 (2024)
76. Thapliyal, P., Chavan, R., Samridh, S., Zuo, C., Banerjee, R.: SBU-SCIRE at CheckThat! 2025: bridging social media, scientific discourse, and scientific literature. In: Faggioli et al. (2025)
77. Toapanta Bernabé, M.d.C., García Cumbreras, M.A., Ureña López, L.A., Mora, D.: UGPLN at CheckThat! 2025: meta-ensemble transformers for numerical claim verification in Spanish. In: Faggioli et al. (2025)
78. Venktesh, V., Anand, A., Anand, A., Setty, V.: Quantemp: a real-world open-domain benchmark for fact-checking numerical claims. In: 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, pp. 650–660. Association for Computing Machinery (ACM) (2024)
79. Venktesh, V., et al.: Overview of the CLEF-2025 CheckThat! lab task 3 on fact-checking numerical claims. In: Faggioli et al. (2025)
80. Vieira, L.L., Jerônimo, C.L.M., Campelo, C.E.C., Marinho, L.B.: Analysis of the subjectivity level in fake news fragments. In: Proceedings of the Brazillian Symposium on Multimedia and the Web, pp. 233–240. WebMedia '20, ACM (2020)
81. Wilder, J., et al.: UNH at Check That! 2025: fine-tuning vs prompting. In: Faggioli et al. (2025)