CheckThat! 2025

8th edition

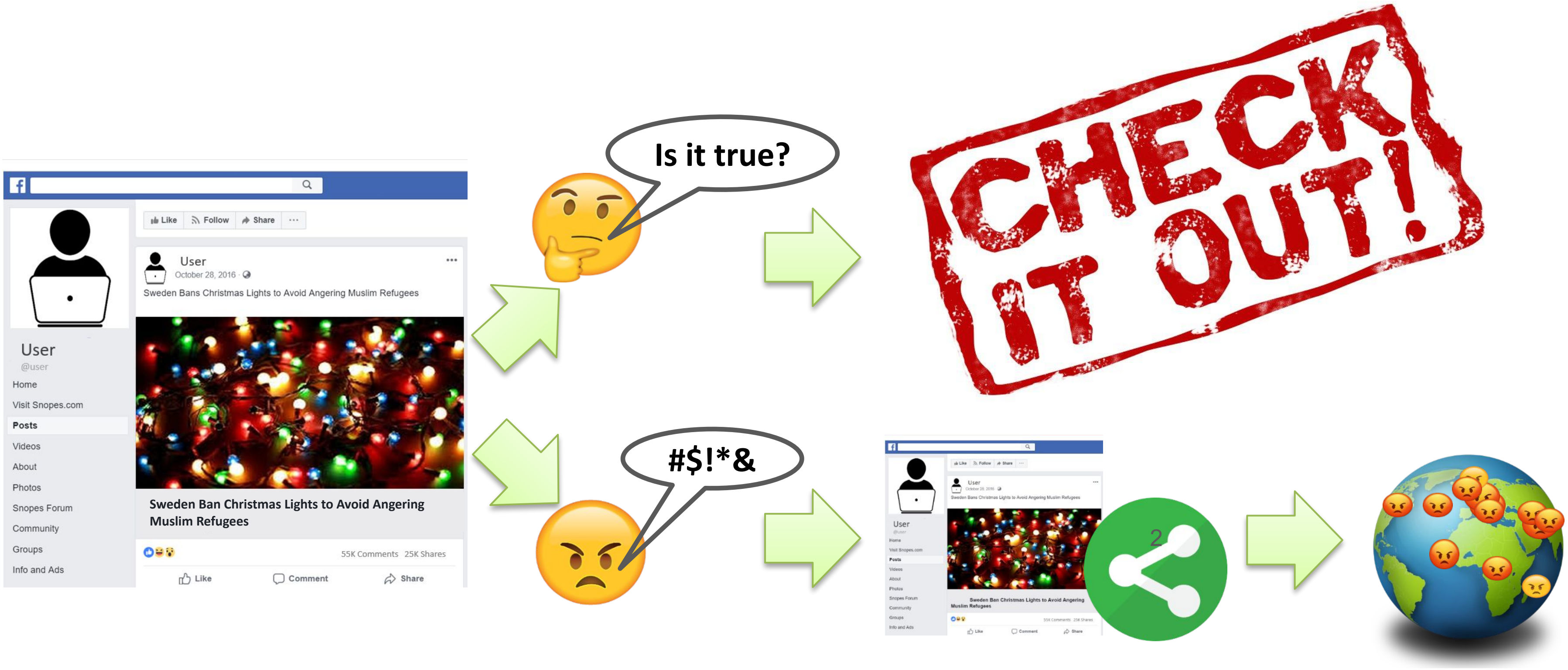
Subjectivity, Fact-Checking, Claim Extraction & Normalization, and Retrieval

http://checkthat.gitlab.io

https://gitlab.com/checkthat_lab/clef2025-checkthat-lab

CLEF 2025 Extended Lab Overview

How?


Like
Follow
Share
User
October 28, 2016
Sweden Bans Christmas Lights to Avoid Angering Muslim Refugees
User
@user
Home
Visit Snopes.com
Posts
Videos
About
Photos
Snopes Forum
Community
Groups
Info and Ads
Sweden Ban Christmas Lights to Avoid Angering Muslim Refugees
55K Comments 25K Shares
Like
Comment
Share
Is it true?
CHECK IT OUT!
#$!*&
2

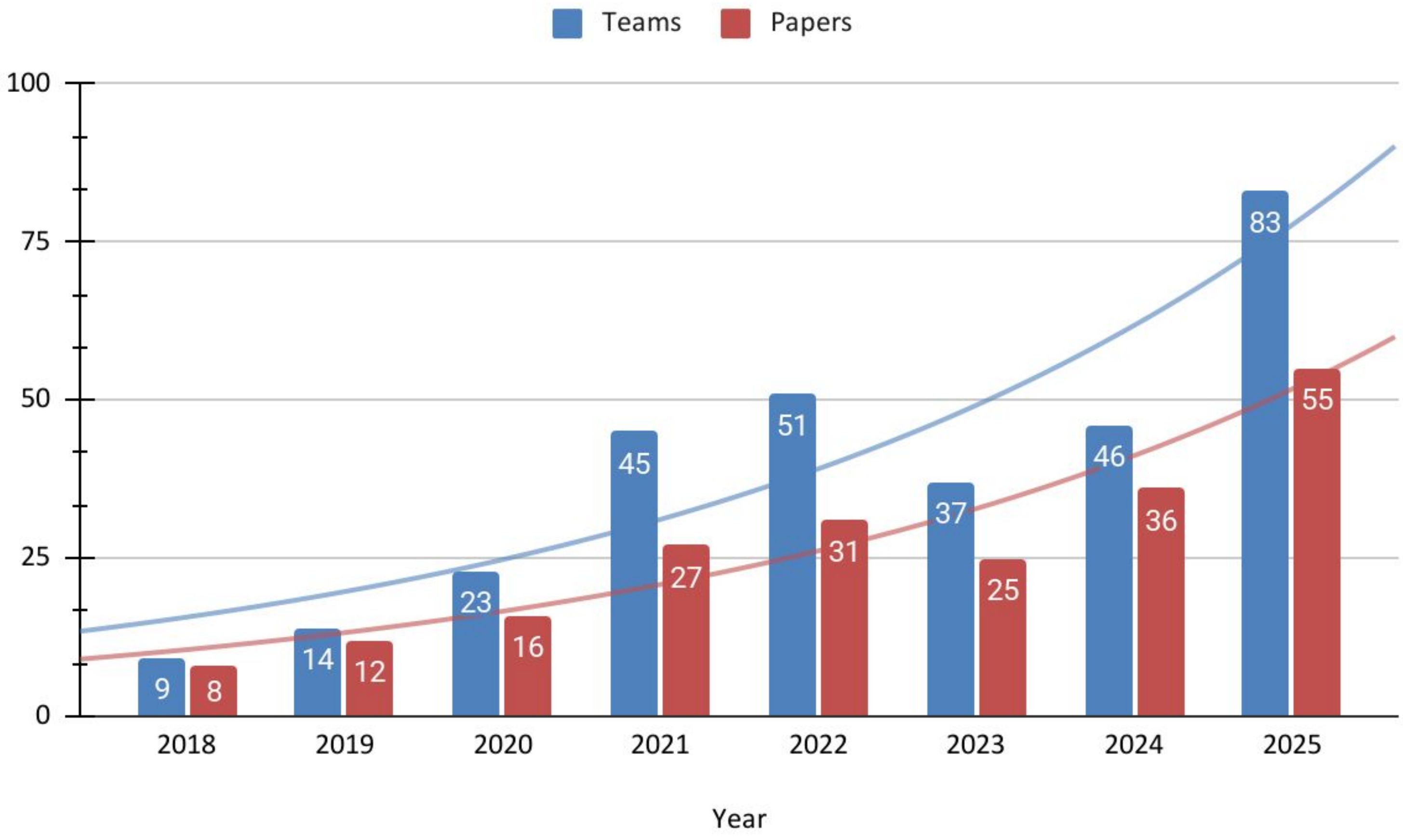
The CheckThat! Lab @ CLEF

Participation

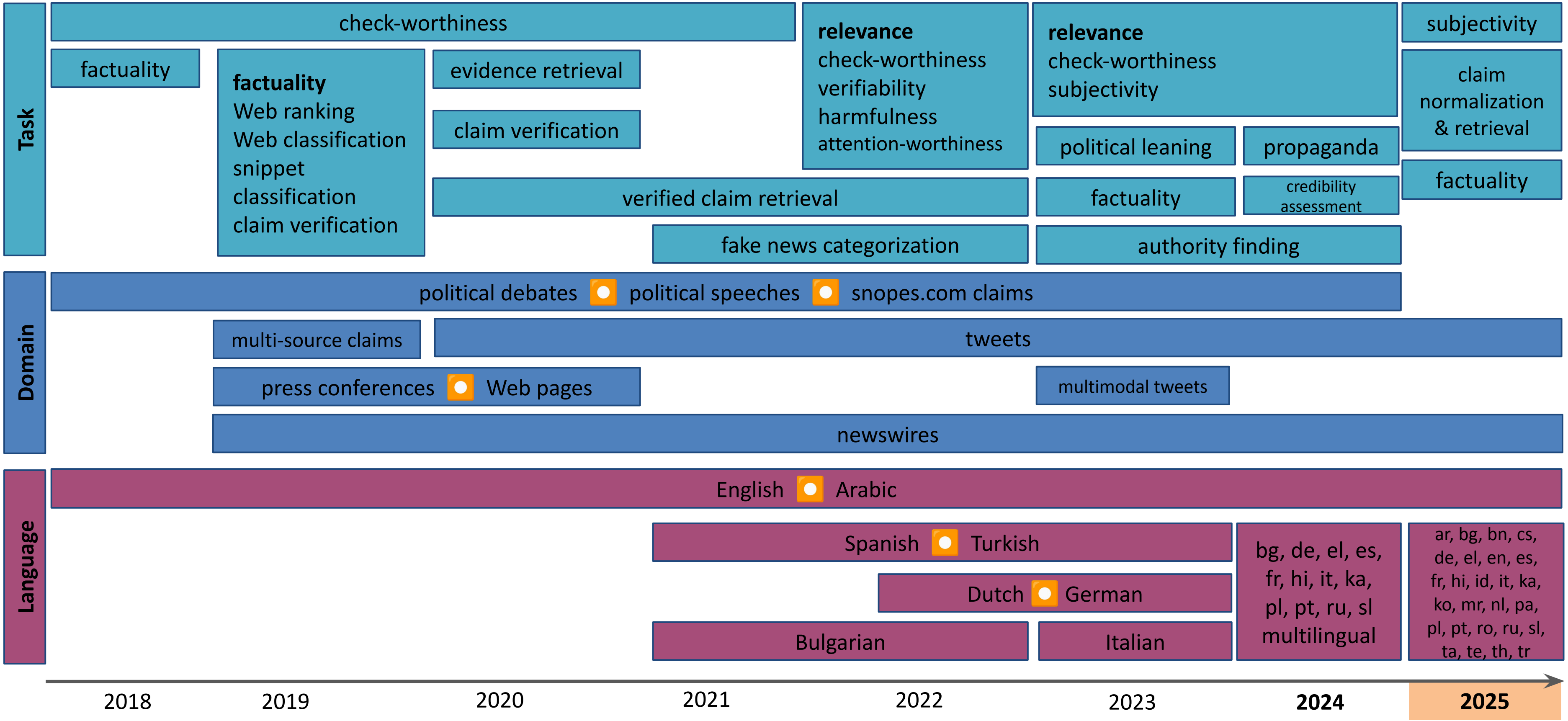
Year	Tasks	Teams	Runs	Papers
2018	Check-worthiness	7	21	5
	Fact-checking	5	14	4
	Total	**9**	**35**	**8**
2019	Check-worthiness	12	21	8
	Evidence & Factuality	4	36	4
	Total	**14**	**57**	**12**
2020	Check-worthiness	15	54	10
	Verified claim retrieval	9	20	5
	Evidence retrieval	1	2	1
	Claim verification	1	2	1
	Total	**23**	**86**	**16**
2021	Check-worthiness	15	74	10
	Verified claim retrieval	5	16	4
	Fake news detection	27	139	13
	Total	**47**	**229**	**27**
2022	Check-worthiness	18	210	13
	Verified claim retrieval	7		3
	Fake news detection	26	126	15
	Total	**51**	**373**	**31**

Year	Tasks	Teams	Runs	Papers
2023	Check-worthiness	19	155	12
	Subjectivity	12	88	10
	Bias	6	41	4
	Factuality	6	28	4
	Authority	2	4	1
	Total	**45**	**316**	**31**
2024	Check-worthiness	28	236	19
	Subjectivity	15	113	11
	Persuasion Techniques	2	-	2
	Hero, villain, and victim	-	-	-
	Authority	5	16	3
	Adversarial Robustness	6	6	6
	Total	**46**	**294**	**36**
2025	Subjectivity	22	436	22
	Claims Normalization	18	1,226	12
	Numerical Claims	13	258	11
	Scientific Web Discourse	40	114	13
	Total	**83**	**2,034**	**55**

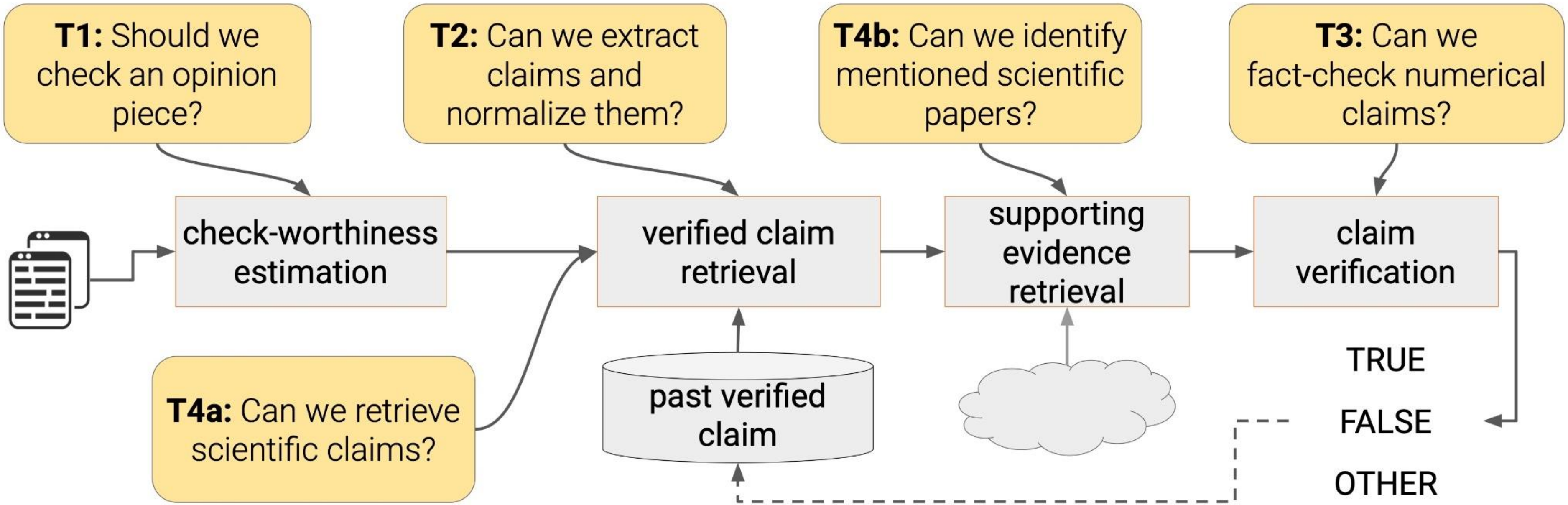
Evolution in Terms of Participation



The CLEF CheckThat! Lab:Tasks, Lang & Data



The Verification Pipeline and 2025 Tasks



Task 1: Subjectivity in News Articles

Motivation

As the influence of digital media has grown, so has the importance of distinguishing between subjective and objective language.

Objective sentences => Fact-checking pipeline

Subjective sentences => Additional processing

- Opinion piece: discard information
- Contains fact:
  - extract the objective version
  - flag it as a feature?

The event, which organisers had envisaged as a celebration of a new, progressive era, turned into a chaotic nightmare.

There is yet everywhere a deficit in the public revenue because the shrinkage in everything taxable was so sudden and violent.

Task Description

Given a sentence, extracted either from a news article, determine whether it is influenced by the subjective view of its author (class **SUBJ**) or presents an objective view of the covered topic (class **OBJ**).

Offered in **nine** languages:

- **Train & Test:** Arabic, Bulgarian, English, German, and Italian
- **Zero-shot:** Greek, Polish, Ukrainian, and Romanian

Also offered in a **multilingual setting**.

Examples

Language	Sentence	Class
Arabic	وجدت بوحيرد نفسها بين يدي ضباط المستعمر الفرنسي فريسة ينهش لحمها بكل الطرق.	SUBJ
	كما تدخل نترات الأمونيوم في صناعة المتفجرات خاصة في مجال التعدين والمناجم.	OBJ
Bulgarian	*Думите на Тръмп са просто думи, докато тези на Обама означават война.*	SUBJ
	Аз се почувствах се глупаво, когато разбрах фактите.	OBJ
English	*But the state's budget is nothing like a credit card.*	SUBJ
	The plan incorporates cash payments supplemented by contingent contributions.	OBJ
German	*Den Grünen bleibt nur, immer wieder darauf hinzuweisen, dass sie selbst gerne ein bisschen großzügiger wären -sich damit aber leider nicht durchsetzen können.*	SUBJ
	Mitte November kündigte die Ampel-Koalition an, das zu ändern.	OBJ
Italian	*Inoltre paragonare immagini di attori paparazzati per strada a foto di studio photo-shoppate non ha senso.*	SUBJ
	Il presidente russo, Vladimir Putin, ha visitato Kaliningrad per incontrare gli studenti dell'Università Kant e tenere un incontro sullo sviluppo della regione.	OBJ

Data

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

Results - monolingual

Rank	Team	F1
Arabic		
1	CEA-LIST	0.6884
2	UmuTeam	0.5903
3	Investigators	0.5880
4	QU-NLP	0.5771
5	AI Wizards	0.5646
6	IIIT Surat	0.5456
7	Arcturus	0.5376
8	Baseline	0.5133
9	ClimateSense	0.5120
10	SmolLab_SEU	0.5053
11	hazemAbdelsalam	0.5038
12	TIFIN INDIA	0.4427
13	JU_NLP	0.4328

Rank	Team	F1
Italian		
1	XplaiNLP	0.8104
2	CEA-LIST	0.8075
3	SmolLab_SEU	0.7750
4	UmuTeam	0.7703
5	Investigators	0.7468
6	Arcturus	0.7282
7	QU-NLP	0.7139
8	AI Wizards	0.7130
9	UNAM	0.7086
10	JU_NLP	0.6991
11	Baseline	0.6941
12	ClimateSense	0.6839
13	TIFIN INDIA	0.5808
14	IIIT Surat	0.4612

Rank	Team	F1
German		
1	SmolLab_SEU	0.8520
2	UNAM	0.8280
3	QU-NLP	0.8013
4	CEA-LIST	0.7733
5	AI Wizards	0.7718
6	Investigators	0.7583
7	TIFIN INDIA	0.7375
8	JU_NLP	0.7356
9	UmuTeam	0.7324
10	XplaiNLP	0.7269
11	ClimateSense	0.7213
12	Arcturus	0.7115
13	duckLingua	0.7114
14	Baseline	0.6960
15	IIIT Surat	0.6342

English		
1	QU-NLP	0.8052
2	TIFIN INDIA	0.7955
3	CEA-LIST	0.7739
4	UmuTeam	0.7604
5	Investigators	0.7544
6	Arcturus	0.7522
7	nlu@utn	0.7486
8	JU_NLP	0.7334
9	SmolLab_SEU	0.7328
10	XplaiNLP	0.7228
11	ClimateSense	0.7226
12	NLP-UTB	0.7130
13	UNAM	0.7075
14	CheckMates	0.7009
15	DSGT-CheckThat	0.6830
16	CUET_KCRL	0.6783
17	CSECU-Learners	0.6777
18	NapierNLP	0.6724
19	AI Wizards	0.6600
20	IIIT Surat	0.6492
21	TIFIN India	0.5756
22	UGPLN	0.5531
23	Baseline	0.5370

Results - unseen languages

Ukrainian		
1	CSECU-Learners	0.6424
2	Investigators	0.6413
3	ClimateSense	0.6395
4	AI Wizards	0.6383
5	Baseline	0.6296
6	SmolLab_SEU	0.6238
7	UmuTeam	0.6210
8	QU-NLP	0.6168
9	XplaiNLP	0.6124
10	CEA-LIST	0.6061
11	JU_NLP	0.5802
12	Arcturus	0.5553
13	IIIT Surat	0.5125
14	TIFIN INDIA	0.4731

Romanian		
1	QU-NLP	0.8126
2	CSECU-Learners	0.7992
3	XplaiNLP	0.7917
4	SmolLab_SEU	0.7892
5	UmuTeam	0.7793
6	CEA-LIST	0.7659
7	AI Wizards	0.7507
8	JU_NLP	0.7442
9	ClimateSense	0.7396
10	Arcturus	0.7366
11	Investigators	0.7133
12	IIIT Surat	0.6496
13	Baseline	0.6461
14	TIFIN INDIA	0.5181

Polish		
1	CEA-LIST	0.6922
2	IIIT Surat	0.6676
3	CSECU-Learners	0.6558
4	AI Wizards	0.6322
5	Arcturus	0.6298
6	Investigators	0.6055
7	UmuTeam	0.5763
8	SmolLab_SEU	0.5738
9	Baseline	0.5719
10	XplaiNLP	0.5665
11	JU_NLP	0.5603
12	ClimateSense	0.5525
13	QU-NLP	0.5165
14	TIFIN INDIA	0.3811

Greek		
1	AI Wizards	0.5067
2	SmolLab_SEU	0.4945
3	CSECU-Learners	0.4919
4	UmuTeam	0.4831
5	XplaiNLP	0.4750
6	Investigators	0.4539
7	CEA-LIST	0.4492
8	JU_NLP	0.4351
9	Baseline	0.4159
10	ClimateSense	0.4137
11	QU-NLP	0.4057
12	Arcturus	0.3905
13	IIIT Surat	0.3733
14	TIFIN India	0.3337

Results - multilingual

Multilingual		
1	TIFIN INDIA	0.7550
2	CEA-LIST	0.7396
3	CSECU-Learners	0.7321
4	XplaiNLP	0.7186
5	SmolLab_SEU	0.7115
6	UmuTeam	0.7074
7	QU-NLP	0.6692
8	JU_NLP	0.6536
9	Arcturus	0.6484
10	ClimateSense	0.6453
11	Baseline	0.6390
12	Investigators	0.6292
13	IIIT Surat	0.5411
14	AI Wizards	0.2380

Results

Arabic

Rank	Team	F1
1	CEA-LIST	0.6884
2	UmuTeam	0.5903
3	Investigators	0.5880
4	QU-NLP	0.5771
5	AI Wizards	0.5646
6	IIIT Surat	0.5456
7	Arcturus	0.5376
8	Baseline	0.5133
9	ClimateSense	0.5120
10	SmolLab_SEU	0.5053
11	hazemAbdelsalam	0.5038
12	TIFIN INDIA	0.4427
13	JU_NLP	0.4328

English

Rank	Team	F1
1	QU-NLP	0.8052
2	TIFIN INDIA	0.7955
3	CEA-LIST	0.7739
4	UmuTeam	0.7604
5	Investigators	0.7544
6	Arcturus	0.7522
7	nlu@utn	0.7486
8	JU_NLP	0.7334
9	SmolLab_SEU	0.7328
10	XplaiNLP	0.7228
11	ClimateSense	0.7226
12	NLP-UTB	0.7130
13	UNAM	0.7075
14	CheckMates	0.7009
15	DSGT-CheckThat	0.6830
16	CUET_KCRL	0.6783
17	CSECU-Learners	0.6777
18	NapierNLP	0.6724
19	AI Wizards	0.6600
20	IIIT Surat	0.6492
21	TIFIN India	0.5756
22	UGPLN	0.5531
23	Baseline	0.5370

Ukrainian

Rank	Team	F1
1	CSECU-Learners	0.6424
2	Investigators	0.6413
3	ClimateSense	0.6395
4	AI Wizards	0.6383
5	Baseline	0.6296
6	SmolLab_SEU	0.6238
7	UmuTeam	0.6210
8	QU-NLP	0.6168
9	XplaiNLP	0.6124
10	CEA-LIST	0.6061
11	JU_NLP	0.5802
12	Arcturus	0.5553
13	IIIT Surat	0.5125
14	TIFIN INDIA	0.4731

Italian

Rank	Team	F1
1	XplaiNLP	0.8104
2	CEA-LIST	0.8075
3	SmolLab_SEU	0.7750
4	UmuTeam	0.7703
5	Investigators	0.7468
6	Arcturus	0.7282
7	QU-NLP	0.7139
8	AI Wizards	0.7130
9	UNAM	0.7086
10	JU_NLP	0.6991
11	Baseline	0.6941
12	ClimateSense	0.6839
13	TIFIN INDIA	0.5808
14	IIIT Surat	0.4612

Multilingual

Rank	Team	F1
1	TIFIN INDIA	0.7550
2	CEA-LIST	0.7396
3	CSECU-Learners	0.7321
4	XplaiNLP	0.7186
5	SmolLab_SEU	0.7115
6	UmuTeam	0.7074
7	QU-NLP	0.6692
8	JU_NLP	0.6536
9	Arcturus	0.6484
10	ClimateSense	0.6453
11	Baseline	0.6390
12	Investigators	0.6292
13	IIIT Surat	0.5411
14	AI Wizards	0.2380

Romanian

Rank	Team	F1
1	QU-NLP	0.8126
2	CSECU-Learners	0.7992
3	XplaiNLP	0.7917
4	SmolLab_SEU	0.7892
5	UmuTeam	0.7793
6	CEA-LIST	0.7659
7	AI Wizards	0.7507
8	JU_NLP	0.7442
9	ClimateSense	0.7396
10	Arcturus	0.7366
11	Investigators	0.7133
12	IIIT Surat	0.6496
13	Baseline	0.6461
14	TIFIN INDIA	0.5181

German

Rank	Team	F1
1	SmolLab_SEU	0.8520
2	UNAM	0.8280
3	QU-NLP	0.8013
4	CEA-LIST	0.7733
5	AI Wizards	0.7718
6	Investigators	0.7583
7	TIFIN INDIA	0.7375
8	JU_NLP	0.7356
9	UmuTeam	0.7324
10	XplaiNLP	0.7269
11	ClimateSense	0.7213
12	Arcturus	0.7115
13	duckLingua	0.7114
14	Baseline	0.6960
15	IIIT Surat	0.6342

Polish

Rank	Team	F1
1	CEA-LIST	0.6922
2	IIIT Surat	0.6676
3	CSECU-Learners	0.6558
4	AI Wizards	0.6322
5	Arcturus	0.6298
6	Investigators	0.6055
7	UmuTeam	0.5763
8	SmolLab_SEU	0.5738
9	Baseline	0.5719
10	XplaiNLP	0.5665
11	JU_NLP	0.5603
12	ClimateSense	0.5525
13	QU-NLP	0.5165
14	TIFIN INDIA	0.3811

Greek

Rank	Team	F1
1	AI Wizards	0.5067
2	SmolLab_SEU	0.4945
3	CSECU-Learners	0.4919
4	UmuTeam	0.4831
5	XplaiNLP	0.4750
6	Investigators	0.4539
7	CEA-LIST	0.4492
8	JU_NLP	0.4351
9	Baseline	0.4159
10	ClimateSense	0.4137
11	QU-NLP	0.4057
12	Arcturus	0.3905
13	IIIT Surat	0.3733
14	TIFIN India	0.3337

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 [33]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓																				
Investigators [34]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓																	
DSGT-CheckThat [35]				✓									✓	✓	✓	✓			✓								✓			
CSECU-Learners [36]					✓	✓	✓	✓	✓	✓		✓					✓		✓											
CEA-LIST [37]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												✓	✓		✓			✓	
IIIT Surat [38]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓																		
TIFIN INDIA [39]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓			✓		✓									✓	✓		✓
ClimateSense [40]	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓			✓		✓	✓	✓					✓				✓	
CUET_KCRL [41]				✓								✓																		
nlu@utn [42]				✓							✓																			✓
XPlaiNLP [43]		✓	✓	✓	✓	✓	✓	✓	✓		✓							✓						✓					✓	
JU_NLP [44]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓																			✓
NapierNLP [45]				✓																				✓		✓			✓	
UmuTeam [46]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓					✓												✓
UGPLN [47]				✓															✓											✓
SmolLab_SEU [48]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓					✓			✓	✓								
Arcturus [49]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓																				
QU-NLP [50]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓											✓									✓
CheckMates [51]				✓							✓								✓											✓
UNAM [52]		✓	✓	✓							✓																			

Summary

- Transformers were most popular, both monolingual and multilingual.
- Many performed feature selection for improvement
- Few approaches relied on LLM-based translation and data augmentation

Task 2: Claims Extraction & Normalization

CheckThat! 2025 Task 2

Claim Normalization

Given noisy social media posts the task is to transform them into clear, concise, and verifiable statements known as normalized claims, which capture the core factual assertion of a post.

Task 2 was offered in 20 languages: English, German, French, Spanish, Portuguese, Hindi, Marathi, Punjabi, Tamil, Arabic, etc.

	Social Media Post	Normalized Claim
English	Something to #consider don't you #think ? Something to #consider, don't you #think? Something to #consider, don't you #think? 40 years worth of research...*no vaccine for HIV *At least 100 years of research...no vaccine for cancer Ongoing research... no vaccine for the common cold Less than a year for a Covid vaccine?	Vaccines for HIV, cold, and cancer should deter you from getting the Covid-19 vaccine.
German	Das reiche Deutschland, wir haben das geringste Durchschnittseinkommen, die geringsten Renten und die dümmsten Wähler. *(Translation: Rich Germany, we have the lowest average income, the lowest pensions and the stupidest voters.)*	Deutschland hat geringste Durchschnittseinkommen und Renten. *(Translation: Germany has the lowest average incomes and pensions.)*

CheckThat! 2025 Task 2

Datasets

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

CheckThat! 2025 Task 2

Model Highlights

Team	Setting		Data			Approach					Model Family					
	Monolingual	Zero-Shot	Content Filtering	Deduplication	Data Augmentation	Fine-Tuning (Full/PEFT)	Zero-Shot Prompting	Retrieval-Augmented ICL	Self-Reflection/Reasoning	Ensemble/Hybrid System	T5-family (T5, Flan-T5)	BART	GPT-family	Llama-family	Qwen	Gemma
dfkinit2b [39]	✓	✓	✓	✓		✓	✓	✓		✓				✓	✓	✓
DS@GT [40]	✓	✓		✓			✓	✓		✓			✓			
TIFIN [41]	✓	✓	✓	✓	✓	✓	✓	✓	✓						✓	
AKCIT-FN [42]	✓	✓		✓		✓	✓	✓			✓	✓	✓		✓	
Factiverse and IAI [43]	✓				✓	✓	✓	✓			✓			✓		
MMA [44]	✓				✓	✓	✓				✓				✓	
UNH [45]	✓					✓	✓	✓	✓		✓		✓	✓		
Investigators [46]	✓	✓	✓			✓					✓	✓		✓		
OpenFact [47]	✓	✓	✓			✓			✓				✓	✓		
JU_NLP@M&S [48]	✓			✓		✓						✓				
Saivineetha [50]	✓	✓				✓	✓									✓
UmuTeam [49]	✓	✓				✓					✓					

CheckThat! 2025 Task 2

Teams participated in different languages

Team	English	Arabic	German	French	Hindi	Marathi	Indonesian	Punjabi	Polish	Portugese	Spanish	Tamil	Thai	Bengali	Telugu	Dutch	Czech	Greek	Romanian	Korean
dfkinit2b [39]	1	1	2	2	1	1	2	1	2	2	2	1	3	1	1	1	1	1	1	1
DS@GT [40]	2	2	1	1	2	4	1	5	1	1	1	3	1	4	5	5	3	4	4	3
TIFIN [41]	3	5	5	6	7	6		4	5		5	5		5	6	4				
AKCIT-FN [42]	4	6	3	3	5	5	3	2	3	3	3	2	2	3	2	2	4	2	2	2
Factiverse and IAI [43]	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 [44]	8	3	7	8	6	3	5	6	7	4	4	6								
UNH [45]	9																			
Investigators [46]	10										8									5
OpenFact [47]	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							
JU_NLP@M&S [48]	14																			
uhh_dem4ai^{+}	15																			
UmuTeam [49]	16	8	8	9	10	8	7	8	8	7	9	7	7	7	8	6	6	6	6	6
VSE^{+}	17																			
saivineetha [50]					3										4					

CheckThat! 2025 Task 2

Results - with seen languages

Scores (METEOR) for languages with training data.

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	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)	-	-	-	-	-	-
OpenFact	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)
JU_NLP@M&S	0.3098 (14)	-	-	-	-	-	-
uhh_dem4ai	0.2612 (15)	-	-	-	-	-	-
UmuTeam	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)	-	-	-	-	-	-

CheckThat! 2025 Task 2

Results - with seen languages

Scores (METEOR) for languages with training data.

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	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)	-
OpenFact	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)
UmuTeam	0.1305 (7)	0.0097 (8)	0.0742 (8)	0.1898 (7)	0.2048 (9)	0.0196 (7)

CheckThat! 2025 Task 2

Results - with unseen languages

Scores (METEOR) for languages with training data.

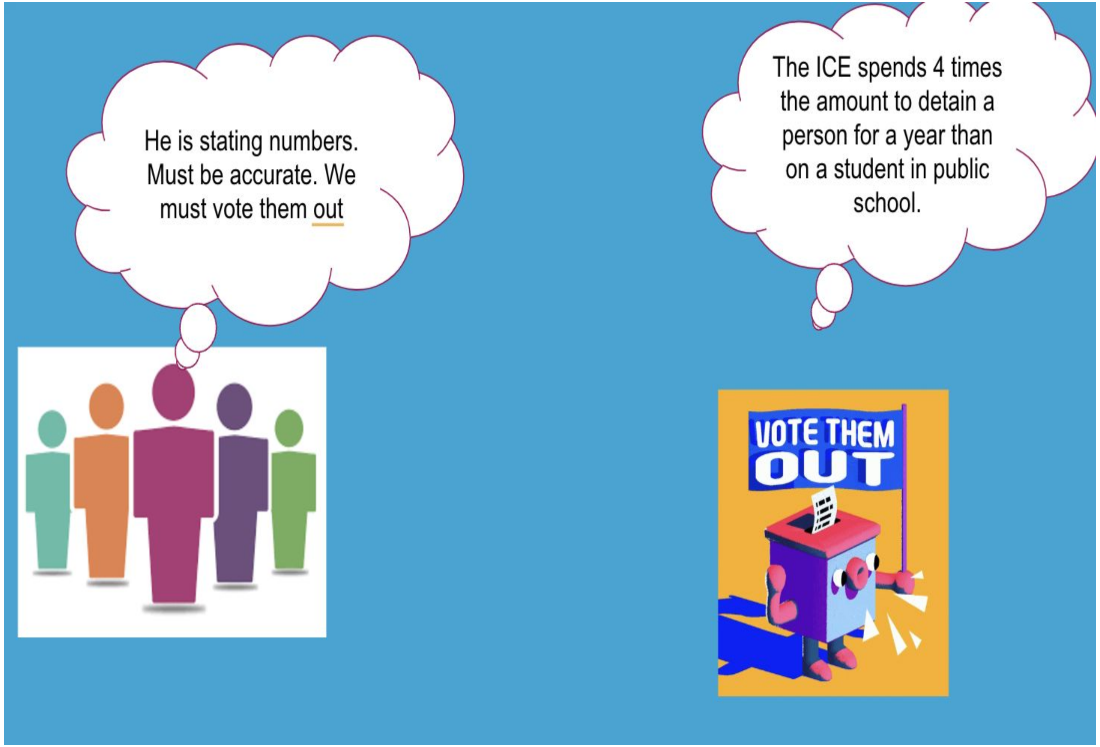
Team Name	Bengali	Telugu	Dutch	Czech	Greek	Romanian	Korean
dfkinit2b	0.3777 (1)	0.5257 (1)	0.2001 (1)	0.2519 (1)	0.2619 (1)	0.2950 (1)	0.1339 (1)
OpenFact	0.2959 (2)	0.4559 (3)	0.1866 (3)	0.2144 (2)	0.2333 (3)	0.2350 (3)	0.1050 (4)
AKCIT-FN	0.2916 (3)	0.5176 (2)	0.1922 (2)	0.1734 (4)	0.2567 (2)	0.2516 (2)	0.1209 (2)
DS@GT	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)

CheckThat! 2025 Task 2 Summary/Findings

- Sequence-to-sequence generation strategies
- Most prevalent approach involved fine-tuning pretrained models such as BART, T5, mBART, and LLaMA
- Preprocessing include emoji removal, hashtag normalization, multilingual data augmentation via translation, and prompt engineering tailored to each language
- Semantic similarity retrieval to choose in-context instances for prompting

Task 3: Fact-Checking Numerical Claims

Motivation - Illusion that numbers indicate truth

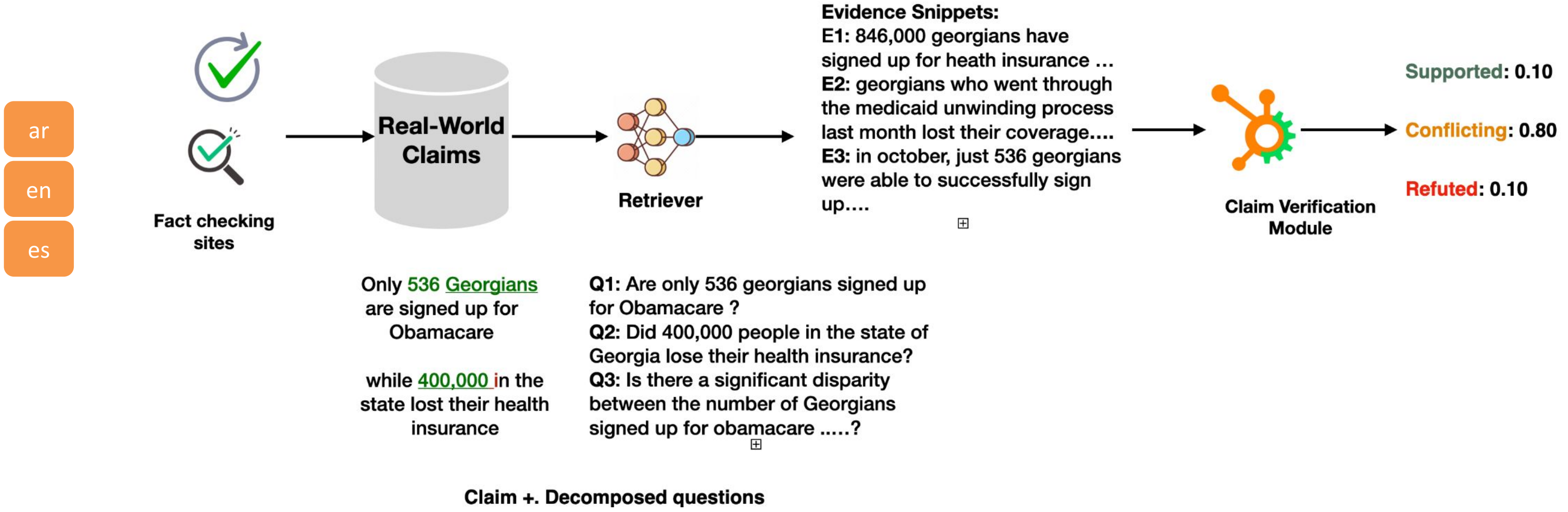


No open-domain Benchmark existed before for fact-checking numerical claims.

Closest works focused a bit on a small sub-category of simple statistical claims.

Task & Data Collection Pipeline

Task: : *Given a numerical claim and the retrieved evidence snippets, the goal is to predict if the evidence supports, refutes, conflicting or is unrelated to the numerical claim.*



Diversity and Coverage of QuanTemp

Table 2: Top fact-checking domains

Claim Source	#Occurences
Politifact	3,840
Snopes	1,648
AfP	412
Africacheck	410
Fullfact	349
Factly	330
Boomlive_in	318
Logically	276
Reuters	235
Lead Stories	223

Table 3: Top claim source countries.

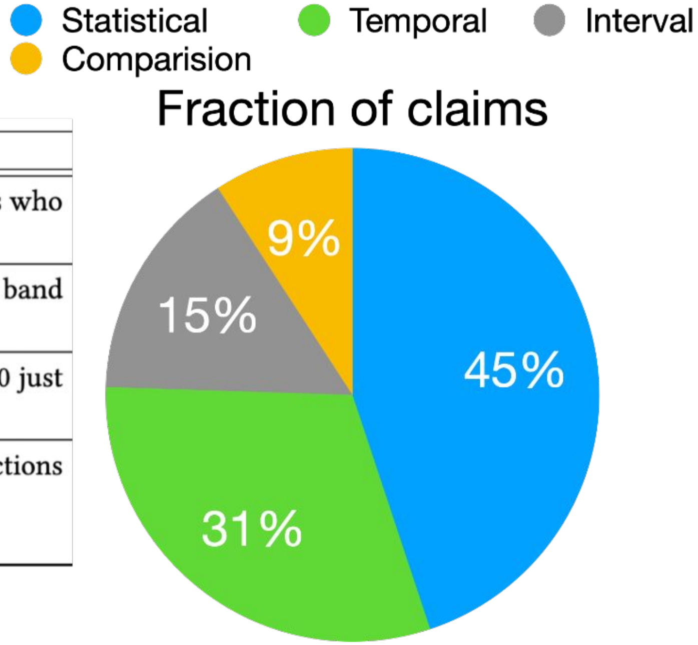
Country	#Occurences
USA	6,215
India	1,356
UK	596
France	503
South Africa	410
Germany	124
Philippines	103
Australia	65
Ukraine	35
Nigeria	17

Table 4: Top evidence domains.

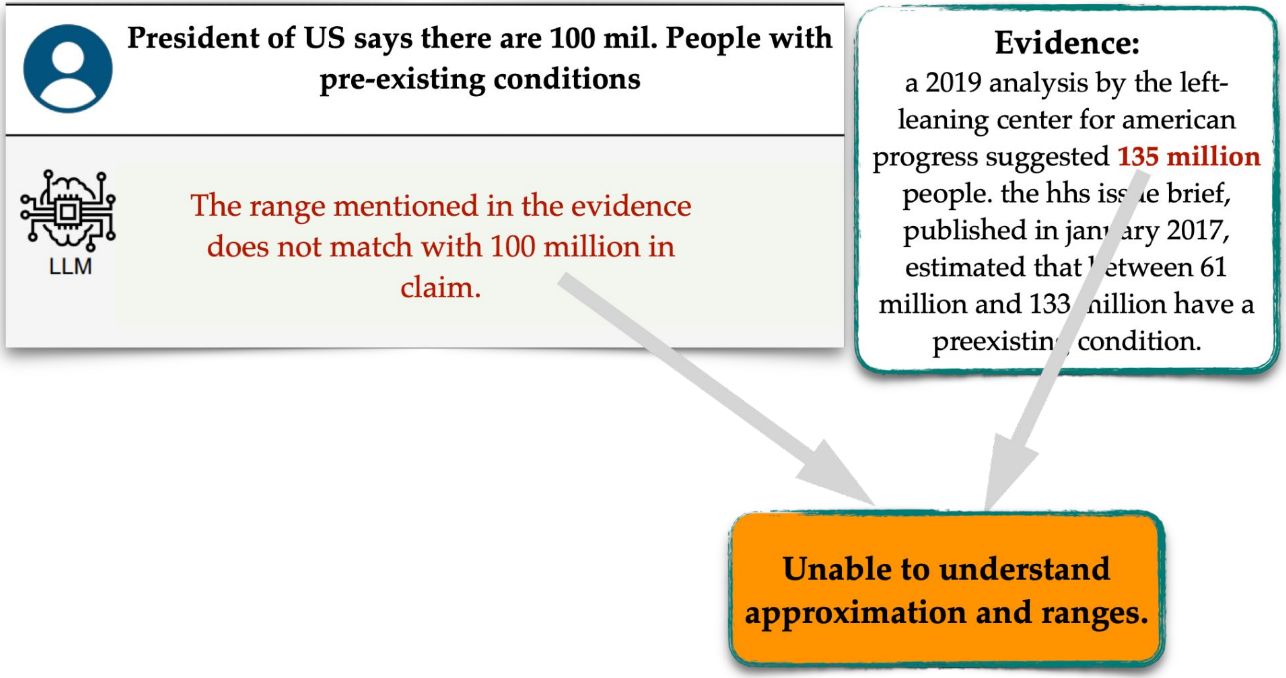
Category	#Occurences
en.wikipedia.org	28,124
nytimes.com	8,430
ncbi.nlm.nih.gov	8,417
quora.com	4,967
cdc.gov	3,987
statista.com	3,106
youtube.com	2,889
who.int	2,557
cnbc.com	2,448
investopedia.com	1977

Claim Categories

Category	Examples
Statistical	We've got 7.2% unemployment (in Ohio), but when you include the folks who have stopped looking for work, it's actually over 10%.
Temporal	The 1974 comedy young frankenstein directly inspired the title for rock band aerosmiths song walk this way
Interval	In Austin, Texas, the average homeowner is paying about $1,300 to $1,400 just for recapture, meaning funds spent in non-Austin school districts
Comparison	A vaccine safety body has recorded 20 times more COVID jab adverse reactions than the government's Therapeutic Goods Administration.



Limitations of just asking LLMs



Some baseline evaluations

- LLMs including GPT-4 struggle with numerical claims
- Models specialized for numerical data FinQA and NumT5 perform best!

Model type	Method	True class (F1)	False class (F1)	Conflicting class (F1)	Full data Macro-F1	Full data Micro-F1
Standard classifier	Roberta-large (fine-tuned)	50.58	77.23	35.50	54.43	62.16
Small generative classifiers	T5-small (fine-tuned)	19.65	77.22	38.02	44.96	56.89
	BART (fine-tuned)	**51.23**	**79.56**	39.37	56.71	64.54
Specialised models with numerical reasoning	NumT5 (fine-tuned)	36.56	78.45	35.76	50.26	60.26
	FinFact (fine-tuned)	49.72	77.91	**47.33**	**58.32**†	**65.23**†
LLM few-shot	FlanT5 (few-shot)	33.90	54.73	20.92	36.52	42.67
	GPT4 (few-shot)	14.38	52.82	42.31	36.50	42.99
	GPT3.5T (few-shot)	44.41	64.26	32.35	47.00	50.98

CheckThat! 2025 Task 3

Datasets

Split	English				Spanish				Arabic			
	T	F	C	Total	T	F	C	Total	T	F	C	Total
Train	1,824	5,770	2,341	9,935	127	1,200	179	1,506	975	1,216	-	2,191
Dev	617	1,795	672	3,084	30	299	48	377	274	313	-	587
Test	717	2,275	664	3,656	115	1,539	152	1,806	206	276	-	482

CheckThat! 2025 Task 3

Datasets

Category	Examples	#of claims
Statistical	We've got 7.2% unemployment (in Ohio), but when you include the folks who have stopped looking for work, it's actually over 10%.	7302 (47.07%)
Temporal	The 1974 comedy young frankenstein directly inspired the title for rock band aerosmiths song walk this way	4193 (27.03%)
Interval	In Austin, Texas, the average home-owner is paying about \$1,300 to \$1,400 just for recapture, meaning funds spent in non-Austin school districts	2357 (15.19%)
Comparison	A vaccine safety body has recorded 20 times more COVID jab adverse reactions than the government's Therapeutic Goods Administration.	1645 (10.60%)

CheckThat! 2025 Task 3

Model Highlights

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 [45]	✓	✓	✓										✓					**50.34**	**96.15**	**59.54**
DS@GT-CheckThat! [41]			✓															-	-	52.10
TIFIN [39]	✓		✓	✓											✓		✓	55.36		55.70
ClaimIQ [11]			✓					✓										-	-	42.43
FraunhoferSIT [59]			✓		✓			✓				✓						-	-	51.00
NGU_Research [1]	✓	✓		✓	✓				✓									63.52	24.41	-
JU_NLP [23]	✓		✓	✓	✓													36.38	-	48.83
CornellNLP [26]			✓	✓		✓		✓										-	-	48.57
UGLPN [80]			✓	✓								✓						-	-	45.53
UCOM_UNAM_PLN [2]		✓		✓		✓												-	35.95	-
News-polygraph*		✓		✓	✓				✓									-	-	42.86

CheckThat! 2025 Task 3 Summary/Findings

- While fine-tuning LLMs for verification helps improve performance, even the best performing solution falls short of upper bound.
- This demonstrates that LLMs struggle to contextualize and accurately interpret numerical information in claims and evidence.
- The task requires reasoning over mixed
  - modalities of numerical and textual data,
  - the ability to contextualize and compare numerical values,
  - and performing numerical reasoning for claim verification.
- Task is far from **being solved.**

Task 4: Scientific Web Discourse

Motivation

Robust methods for the processing of scientific discourse on social media

➢ Scientific topics, claims and resources are **increasingly debated online** *(Fig.)*

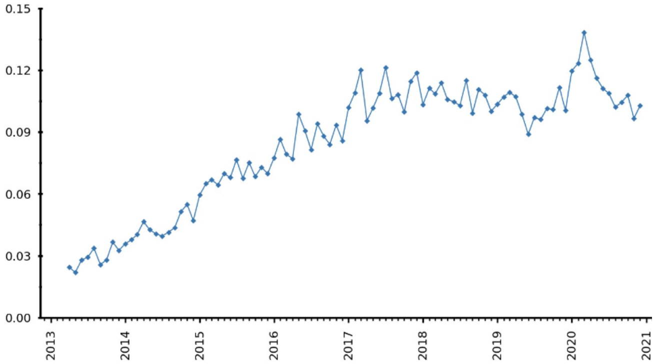


Fig. *Proportion of science-related tweets between 2013 and 2020*

➢ Yet scientific discourse on the Web is often **decontextualized,**[1] making it **difficult to assess the validity and the original sources of scientific claims** around important societal topics (e.g., COVID-19, climate change)

Example:

"stanford study says masks are totally inefficient"

1. scientific claim
   => no scientific context (e.g., population size, statistical significance)
2. scientific reference
   => no links/identifiers (e.g., DOI) to the actual study

[1] *Hafid et al., "Disambiguation of Implicit Scientific References on X", ACM HyperText, 2025*

Task Description & Data

Task 4a: Scientific Web Discourse Detection

Objective: Detecting different forms of Scientific Web Discourse (e.g., claims, references)

Task 4b: Scientific Claim Source Retrieval

Objective: Retrieving source publications from which claims and references originate

Task Description & Data

Task 4a: Scientific Web Discourse Detection

Detect different forms of Scientific Web Discourse in a given set of social media posts (tweets). Scientific Web Discourse is categorised as posts that contain:

1. **a scientific claim** that may be verified or refuted using primary scientific publications
2. **a reference to a scientific study**/publication
3. **a reference to scientific contexts or entities**, e.g., a university, a scientist or a scientific conference

Dataset*

- ➢ 1,606 posts from X (Twitter)
- ➢ Manual annotation for each of the 3 categories of scientific web discourse

*(*): **Definitions, Categories** and **Examples** are extracted from our **previous work**, see Hafid et al., "SciTweets- a dataset and annotation framework for detecting scientific online discourse", CIKM 2022*

Task Description & Data

Task 4a: Scientific Web Discourse Detection

Detect different forms of Scientific Web Discourse in a given set of social media posts (tweets). Scientific Web Discourse is categorised as posts that contain:

1. **a scientific claim** that may be verified or refuted using primary scientific publications
2. **a reference to a scientific study**/publication
3. **a reference to scientific contexts or entities**, e.g., a university, a scientist or a scientific conference

Examples*

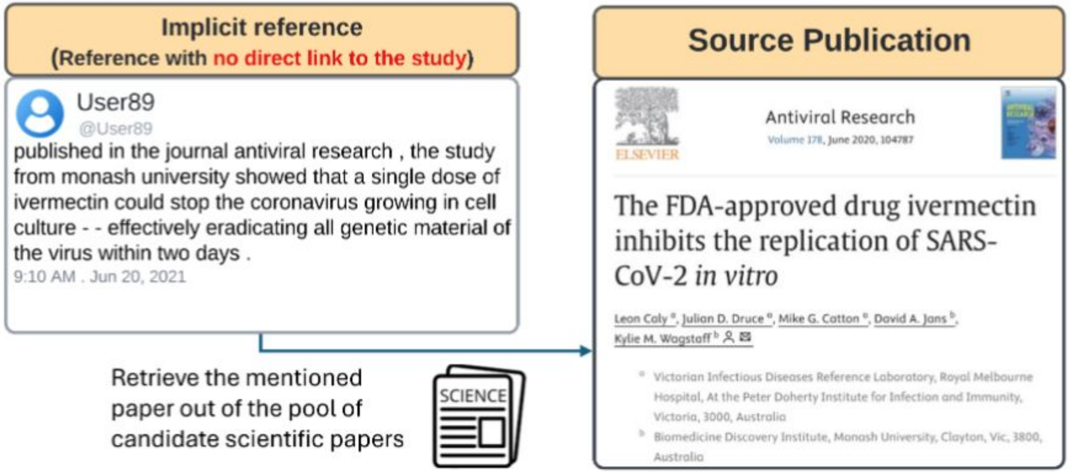
1) Science related	1.1 Scientific Claim	***Donating blood*** *not only helps others, but* ***reduces the rate of cancer and heart disease*** *in the donor.*
	1.2 Scientific Reference	*via @medical_xpress* ***A new in vitro*** *(test tube)* ***study****, "Dietary functional benefits of Bartlet http://t.co/Qv1C1GjQin #UFO4UBlogHealth*
	1.3 Scientific Research Context	*How is @UChicagoIME* ***shaping the future or science*** *? Find out on April 6!*
2) Not science related		***My father*** *got COVID-19.*

*(*): **Definitions, Categories** and **Examples** are extracted from our **previous work**, see Hafid et al., "SciTweets- a dataset and annotation framework for detecting scientific online discourse", CIKM 2022*

Task Description & Data

Task 4b: Scientific Claim Source Retrieval

Given a tweet referring to a scientific study in an informal way, identify the correct study out of a pool of candidate scientific papers.*



(*): For more details on the definitions, task formulation, and annotation protocol, *see Hafid et al., "Disambiguation of Implicit Scientific References on X", ACM HyperText, 2025*

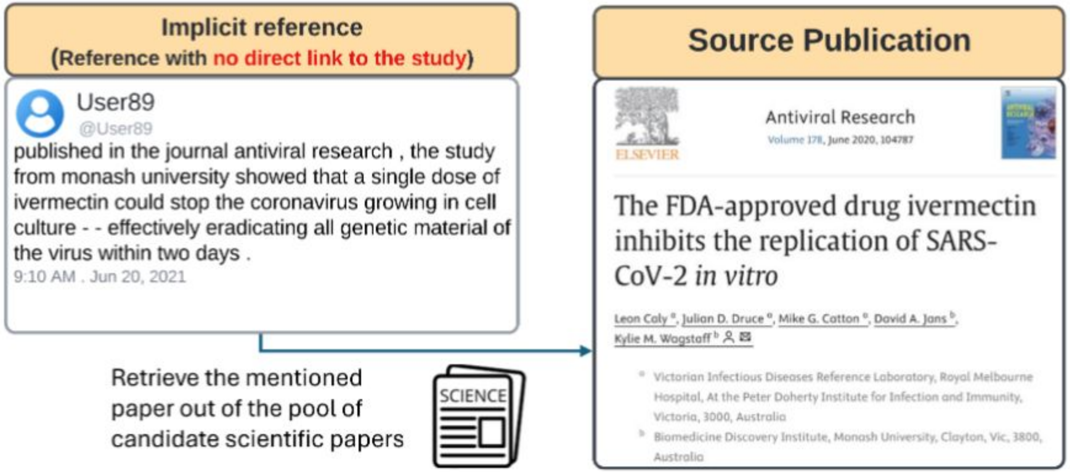
Task Description & Data

Task 4b: Scientific Claim Source Retrieval

Dataset:

➢ Query set: 15,699 posts from X with implicit references to scientific papers from CORD-19 [1]

➢ Collection set: metadata (e.g., title, abstract, affiliations) of the 7,718 CORD-19 scientific papers which the query set posts implicitly refer to



[1]: Wang et al., "CORD-19: The COVID-19 open research dataset", 1st Workshop on NLP for COVID-19, ACL 2020

Approaches

- **Task 4a (multi-class classification task)**
  - Mostly Transformer-based models (e.g., SciBERT, DeBERTa-v3, Twitter-Roberta) and LLMs
  - Additional approaches
    - data augmentation
    - ensemble methods
    - optimization techniques
- **Task 4b (IR task)**
  - Mostly a two-stage approach: Dense retrieval + Neural re-ranking
  - Additional approaches
    - strategic sampling of hard negatives
    - style transfer techniques

Results

Total participation (Task 4a): **10 teams**

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 [18]			■		■		■		0.7998	1
VerbaNexAI [20]	■						■	■	0.7983	2
SBU-SCIRE [21]	■					■		■	0.7917	4
DS@GT [23]	■			■	■		■		0.7685	6
DeBERTa-v3 Baseline	■								0.7668	7
TurQUaz [24]				■			■		0.7615	8
JU_NLP [25]		■	■				■		0.7347	9

Results

Total participation (Task 4b): **30 teams**

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 [26]	■		■				0.67	2
Deep Retrieval [27]	■	■	■	■			0.66	3
ATOM [28]	■		■				0.66	4
SBU-SCIRE [21]	■		■				0.65	5
SeRRa [29]	■		■				0.61	8
Claim2Source [30]	■	■		■		■	0.59	12
DS@GT [31]		■	■	■	■	■	0.58	16
BM25 Baseline		■					0.43	28

Summary/Main Takeaways/Highlights

- **Task 4a (multi-class classification task)**
  - Overall, **fine-tuning existing pre-trained language models works best in terms of avg F1-score**
  - LLM approaches perform better for the subtask of identifying scientific references (category 2)
- **Task 4b (IR task)**
  - 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
  - Re-rankers included **LLMs** (ChatGPT, LLaMa, Gemma) but **did not always outperform transformer-based models**
  - Style-transfer techniques showed mixed results
  - Strategic sampling of hard negative samples led to clear performance gains

⟶ With best-performing scores at **F1=0.80 for Task 4a** and **MRR@5=0.67 for Task 4b**, both tasks still show clear room for improvement

CheckThat! Program

Programme (Madrid time)

CT! oral session 1: Thursday 11th September, 14:15 to 15:45

14:15 - Introduction to the CheckThat! Lab

15:00 - **Task 1 & 2**: Three talks on Subjectivity and Claim Normalization

CLEF poster session 3: Thursday 11th September, 15:45 to 16:30

CT! oral session 2: Thursday 11th September, 16:30 to 18:00

16:30 - **Task 2**: One talk on Claim Normalization

16:45 - **Task 3**: Three talks on Numerical Claim Verification

17:30 - **Task 4**: Two talks on Numerical Claim Verification

CT! oral session 3: Friday 12th September, 11:30 to 13:00

11:30 - **Invited talk**. Rubén Míguez Pérez

Details on the CheckThat! website:

http://checkthat.gitlab.io/clef2025/#lab-programme

Our Organization Team


Firoj Alam


Julia Maria Struß


Tanmoy Chakraborty


Stefan Dietze


Salim Hafid


Katerina Korre


Arianna Muti


Preslav Nakov


Federico Ruggeri


Sebastian Schellhammer


Vinay Setty


Megha Sundriyal


Konstantin Todorov


Venktesh Viswanathan

CheckThat! 2025

https://checkthat.gitlab.io

Acknowledgements

مجلس قطر للبحوث والتطوير والابتكار
QRDI COUNCIL

Federal Ministry of Education and Research

République du Sénégal
Un Peuple - Un But - Une Foi
MESRI
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION

AGENCE NATIONALE DE LA RECHERCHE
ANR