Overview of the CLEF-2025 CheckThat! Lab Task 1 on Subjectivity in News Articles

Notebook for the CheckThat! Lab at CLEF 2025

Federico Ruggeri[1,*], Arianna Muti[2], Katerina Korre[3], Julia Maria Struß[4], Melanie Siegel[5], Michael Wiegand[6], Firoj Alam[7], Md. Rafiul Biswas[8], Wajdi Zaghouani[9], Maria Nawrocka[10], Bogdan Ivasiuk[1], Gogu Razvan[11] and Andreiana Mihail[11]

[1]*University of Bologna, Italy*
[2]*Bocconi University, Milan, Italy*
[3]*Athena RC, Athens, Greece*
[4]*University of Applied Sciences Postdam, Germany*
[5]*Darmstadt University of Applied Sciences, Germany*
[6]*University of Vienna, Austria*
[7]*Qatar Computing Research Institute, Doha, Qatar*
[8]*Hamad Bin Khalifa University, Qatar*
[9]*Northwestern University in Qatar, Doha, Qatar*
[10]*Warsaw University, Poland*
[11]*Bucharest University, Romania*

Abstract

We present an overview of Task 1 of the eighth edition of the CheckThat! lab at the 2025 edition of the Conference and Labs of the Evaluation Forum (CLEF). The task required participants to determine whether individual sentences from news articles expressed subjective viewpoints, such as opinions or personal bias, or presented objective, fact-based information. The task was offered in nine languages: Arabic, Bulgarian, English, German, Italian, Greek, Polish, Romanian, and Ukrainian, as well as in a multilingual setting. We curated datasets for each language, comprising roughly 14,000 sentences sourced from diverse news outlets. Participants were tasked with developing classification systems to identify subjectivity (personal opinions or biases) and objectivity (factual information) at the sentence level. A total of 22 teams participated in the task, submitting 436 valid runs across all language tracks. Most systems were based on transformer models, with approaches ranging from fine-tuning language-specific and multilingual encoders to applying English-centric models in combination with machine translation. Several teams also experimented with ensemble techniques, handcrafted features, and in-context learning using large language models. Systems were evaluated using macro-averaged F1 score to ensure equal weighting of subjective and objective classes. Performance varied considerably by language: German, Italian, English and Romanian yielded the highest results. In contrast, Greek and Ukrainian emerged as the most challenging languages, with no team surpassing the 0.65 and 0.51 F1 score marks, respectively. Task 1 offers a valuable benchmark for the development and evaluation of multilingual subjectivity detection systems. This paper presents an overview of Task 1, including datasets, system strategies, and outcomes, contributing to broader research efforts aimed at improving the transparency and trustworthiness of automated content analysis.

Keywords

subjectivity classification, fact-checking, misinformation detection

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.
federico.ruggeri6@unibo.it (F. Ruggeri); arianna.muti@unibocconi.it (A. Muti); k.korre@athenarc.gr (K. Korre); julia.struss@fh-potsdam.de (J. M. Struß); melanie.siegel@h-da.de (M. Siegel); michael.wiegand@univie.ac.at (M. Wiegand); firojalam@gmail.com (F. Alam); mbiswas@hbku.edu.qa (Md. R. Biswas); wajdi.zaghouani@northwestern.edu (W. Zaghouani); m.narwocka6@uw.edu.pl (M. Nawrocka); bogdan.ivasiuk@studio.unibo.it (B. Ivasiuk); gogu.razvan2001@gmail.com (G. Razvan); andreianna@yahoo.com (A. Mihail)
0000-0002-1697-8586 (F. Ruggeri); 0000-0002-3387-6557 (A. Muti); 0000-0002-9349-9554 (K. Korre); 0000-0001-9133-4978 (J. M. Struß); 0000-0002-5064-5750 (M. Siegel); 0000-0002-5403-1078 (M. Wiegand); 0000-0001-7172-1997 (F. Alam); 0000-0002-5145-1990 (Md. R. Biswas); 0000-0003-1521-5568 (W. Zaghouani); 0009-0003-1658-0804 (B. Ivasiuk)

1. Introduction

The CheckThat! lab is organized for the 8th time within CLEF 2025. This paper presents an overview of Task 1, which covers the challenge of identifying subjectivity in news articles — a task introduced in the 2023 edition [1] and now held for the third time.

As the influence of digital media has grown, so has the importance of distinguishing between subjective and objective language. This distinction is paramount in Natural Language Processing (NLP), especially in domains such as sentiment analysis, opinion mining, and, crucially, fact-checking. Subjective statements often convey personal judgments, emotions, or implicit bias, whereas objective ones aim to report verifiable facts. Automatically recognizing this difference is essential for building systems that can assess the trustworthiness and neutrality of textual information.

Task 1 is designed to foster research in this direction by providing a multilingual benchmark for sentence-level subjectivity classification. Participants are asked to determine whether a sentence taken from a news article reflects the author's personal viewpoint or offers a neutral, fact-based perspective. This binary classification task is especially relevant in the current media landscape, where biased reporting and misinformation pose ongoing challenges to public discourse and information integrity.

The task includes datasets in nine languages: Arabic, Bulgarian, English, German, Italian, Polish, Ukrainian, Romanian, and Greek. In particular, the subjectivity task is organized to cover three distinct settings: monolingual, where the focus is on a specific language; multilingual, where the contribution of multiple languages is evaluated; and zero-shot, where the generalization capabilities of models trained on seen languages are tested on unseen ones. All datasets were annotated using a prescriptive framework designed to support cross-lingual comparability and high annotation quality. System performance was evaluated using macro-averaged F1 score to ensure a balanced treatment of both subjective and objective classes. This approach provides a fair and comprehensive measure of system effectiveness across diverse languages and content.

The remainder of the paper is as follows. We first describe the dataset construction process, evaluation criteria, and submission protocols. We then analyze the submitted systems, comparing their methodologies and results to assess current progress and identify key challenges. Task 1 contributes to the broader effort to improve automated understanding of subjectivity in online content — an increasingly critical component of trustworthy AI applications in the digital era.

2. Related Work

Research on subjectivity detection spans a wide array of contexts and has evolved significantly over time. While early developments were closely tied to sentiment analysis in English-language texts [2, 3], subsequent efforts extended to multilingual domains [4, 5], paving the way for cross-lingual approaches. Over the years, the task has also found relevance in detecting bias [6, 7], identifying claims [8], and supporting fact-checking workflows [9, 10], which directly motivates the present work.

The criteria for identifying subjectivity often differ depending on the application, and so do the methodological approaches. Some studies employ lexical heuristics tailored to specific domains or tasks [2, 11, 12], while others rely on statistical modeling techniques [13]. A more rigorous path involves manually curated datasets developed through detailed annotation protocols [14, 15, 16]. As noted by Chaturvedi et al. [17], these approaches can be grouped into syntactic methods—primarily rule-based and surface-oriented—and semantic methods, which involve deeper linguistic and contextual understanding.

Syntactic methods, although efficient in certain settings, typically suffer from portability issues due to their dependence on language- or domain-specific indicators. Semantic methods have become more prevalent as they tend to generalize better, especially when built on systematic annotation schemes. Still, annotation-driven approaches are not without limitations: disagreements among annotators, vague or context-sensitive cases, and subjective interpretation introduce inconsistency and noise [15, 18]. Recent work has attempted to mitigate these issues through prescriptive annotation strategies [19],

Table 1
Dataset statistics for the five languages for which we report training and development data splits. Additionally, we report unseen language test split statistics.

	Arabic		Bulgarian		English		German		Italian	
Training Languages	**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

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

particularly in the setting of fact verification, where subjective cues are often indicative of unverifiable or misleading information [20].

Our work incorporates annotation at multiple textual levels—ranging from isolated sentences [8, 21], to text segments [22], and full documents [23]. Although English dominates in terms of available annotated resources, the field has seen growing interest in developing datasets for other languages, including Arabic [24, 25], German [24], French [22], Italian [23], Romanian [26], and Spanish [24]. Nevertheless, many of these efforts rely on machine translation and ontology-driven methods for scalability, which can introduce labeling errors and annotation inconsistencies across languages.

This framing has been formalized in recent shared tasks. For instance, the sixth edition of the CheckThat! lab included a dedicated task on subjectivity detection [1], which serves as the foundation for our current efforts. The language coverage has changed slightly since then: due to resource limitations, Dutch and Turkish were removed, and Bulgarian was added as a new language in the 2024 iteration [27]. In particular, the CheckThat!lab 2024 Task 2 edition [27] also covered multilingual subjectivity detection, covering five languages: Arabic, English, German, Italian, and Bulgarian. Our work builds on this task, extending the set of covered languages to nine, including Polish, Ukrainian, Romanian and Greek, and exploring zero-shot learning on these unseen languages.

3. Datasets

The task offered datasets in nine different languages with a total of more than 14k sentences manually annotated following the guidelines in [20]. Table 1 presents details on the dataset statistics. Some sample instances for each language are given in Table 2.

3.1. Arabic

For this edition, we used the released dataset from [28] and developed a new test set for the final evaluation. The dataset consists of manually annotated sentences from news articles, including sources such as AraFact [29]. The complete data collection and annotation process involved several phases. In the article selection phase, 1,159 news articles were selected from AraFact [29]. Additionally, opinionated articles were manually searched from various Arabic news outlets, resulting in the selection of 221 articles. These articles were parsed and segmented into sentences for annotation.

The annotation was conducted using the MTurk platform. To ensure annotation quality, standard qualification tests were applied, and the final label for each sentence was determined using majority

Table 2
Examples of subjective and objective sentences in the annotated datasets.

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
Greek	*Πάντως η Τουρκία συνεχίζει, παρά το καλό κλίμα στη συνάντηση Γεραπετρίτη – Φιντάν στην Ντόχα, να θέτει ζητήματα που αν μη τι άλλο αμφισβητούν κυριαρχικά δικαιώματα και συνθήκες*	SUBJ
	Η επικινδυνότητά τους για την υγεία, διευκρινίζει ο Οργανισμός, είναι ιδιαίτερα υψηλή	OBJ
Polish	*Co ciekawe, w obu wypadkach wściekłość Tuska wywołało to samo.*	SUBJ
	W 2023 r. lekarze psychiatrzy wystawili 1,4 mln zwolnień lekarskich.	OBJ
Romanian	*Societatea noastră se prăbușește din cauză că nu le respectăm.*	SUBJ
	Locuia într-o vilă în centrul Bucureștiului și tot acolo făcuse redacția.	OBJ
Ukrainian	*Однак віцеканцлер вважає, що європейське регулювання не справляється з цим завданням*	SUBJ
	Європа має велику проблему з імміграцією, — додав Трамп, захищаючи свого заступника	OBJ

agreement. A label was assigned to a sentence if at least two annotators agreed. The inter-annotator agreement (pairwise Cohen's kappa) was $\kappa = 0.538$. More details about the data collection and annotation process can be found in [28].

3.2. English

For training, we used NewsSD-ENG [20], a corpus of 1,049 sentences labeled by seven annotators following guidelines for subjectivity detection tailored to an information retrieval setting [30]. We merged the dev and dev-test partitions of the CheckThat! lab 2024 Task 2 edition [27] and re-declared its test split as the new dev-test split. We further collected a novel test set following the same data collection methodology for NewsSD-ENG. In particular, we retrieved 11 news articles on controversial topics and randomly sampled 301 sentences. Then, seven annotators labeled the sentences as subjective or objective. We organized annotators such that each sentence was annotated by three annotators. The inter-annotator agreement on the new test set measured with Krippendorff's alpha was 0.43.

3.3. German

The German dataset was assembled by randomly selecting sentences from the CT 2022 FAN-Corpus [31] consisting of news articles that have been annotated according to the factuality of their main claim, originally. The 800 manually annotated sentences for training and the 491 and 337 instances of the development and develpment-test sets are from the 2023 and 2024 editions of the task [27]. A new test set has been annotated following the guidelines outlined in [20]. We excluded all incomplete sentences

as well as non German ones. We also reduced instances consisting of more than one sentence due to wrong sentence splitting to one sentence. Each sentence has been annotated by the same three native speakers as in previous iterations of the task, all co-authors of this paper. As the agreement between the annotators was substantially lower compared to previous years, the annotators discussed every sentence with deviating labels reaching a consensus (Fleiss' kappa on the 2025 test set: $kappa = 0.547$ ($p < 0.0001$, $z = 17.7$), Fleiss' kappa on the 2024 test set: $kappa = 0.696$ ($p < 0.0001$, $z = 22.1$)).

3.4. Italian

For training, we used the re-annotated version of SubjectivITA [23] introduced in the CheckThat! lab 2024 Task 2 edition [27]. SubjectivITA is a corpus of news articles annotated for subjectivity detection, containing 1,841 sentences. We merged the dev and dev-test partitions of the CheckThat! lab 2024 Task 2 edition and re-declared its test split as the new dev-test split. We eventually collected a novel test split following the same methodology used for the English dataset. In particular, we collected 13 news articles targeting controversial topics and randomly sampled 300 sentences. The inter-annotator agreement on the new test set measured with Krippendorff's alpha [32] was 0.53.

3.5. Romanian

We built the Romanian zero-shot test set from multiple online news websites. In particular, we collected 4 news articles covering controversial topics and randomly sampled 300 instances. Each instance was labeled as objective or subjective by two native Romanian speakers. The inter-annotator agreement for the zero-shot Romanian test set, measured using Cohen's, is 0.30.

3.6. Polish

We built the Polish zero-shot test set from multiple online news websites. In particular, we collected 11 news articles covering controversial topics and randomly sampled 350 instances. Each instance was labeled as objective or subjective by one native Polish speaker.

3.7. Ukrainian

We built the Ukrainian zero-shot test set from multiple online news websites. In particular, we collected 17 news articles covering controversial topics and randomly sampled 297 instances. Each instance was labeled as objective or subjective by one native Ukrainian speaker.

3.8. Greek

We built the Greek zero-shot test set from multiple online news websites. In particular, we collected 11 news articles covering controversial topics and randomly sampled 300 instances. Each instance was labeled as objective or subjective by six native Greek speakers. The inter-annotator agreement for the zero-shot Greek test set, measured using Krippendorff's alpha, is 0.36.

4. Overview of the Systems and Results

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 3 shows the results achieved by the individual teams for each language. 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

Table 3
Results for subjectivity classification of news articles. The F_1-measure is macro-averaged.

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
4	QU-NLP	0.5771	4	UmuTeam	0.7703	4	CEA-LIST	0.7733
5	AI Wizards	0.5646	5	Investigators	0.7468	5	AI Wizards	0.7718
6	IIIT Surat	0.5456	6	Arcturus	0.7282	6	Investigators	0.7583
7	Arcturus	0.5376	7	QU-NLP	0.7139	7	TIFIN INDIA	0.7375
8	Baseline	0.5133	8	AI Wizards	0.7130	8	JU_NLP	0.7356
9	ClimateSense	0.5120	9	UNAM	0.7086	9	UmuTeam	0.7324
10	SmolLab_SEU	0.5053	10	JU_NLP	0.6991	10	XplaiNLP	0.7269
11	hazemAbdelsalam	0.5038	11	Baseline	0.6941	11	ClimateSense	0.7213
12	TIFIN INDIA	0.4427	12	ClimateSense	0.6839	12	Arcturus	0.7115
13	JU_NLP	0.4328	13	TIFIN INDIA	0.5808	13	duckLingua	0.7114
	English		14	IIIT Surat	0.4612	14	Baseline	0.6960
1	QU-NLP	0.8052		**Multilingual**		15	IIIT Surat	0.6342
2	TIFIN INDIA	0.7955	1	TIFIN INDIA	0.7550		**Polish**	
3	CEA-LIST	0.7739	2	CEA-LIST	0.7396	1	CEA-LIST	0.6922
4	UmuTeam	0.7604	3	CSECU-Learners	0.7321	2	IIIT Surat	0.6676
5	Investigators	0.7544	4	XplaiNLP	0.7186	3	CSECU-Learners	0.6558
6	Arcturus	0.7522	5	SmolLab_SEU	0.7115	4	AI Wizards	0.6322
7	nlu@utn	0.7486	6	UmuTeam	0.7074	5	Arcturus	0.6298
8	JU_NLP	0.7334	7	QU-NLP	0.6692	6	Investigators	0.6055
9	SmolLab_SEU	0.7328	8	JU_NLP	0.6536	7	UmuTeam	0.5763
10	XplaiNLP	0.7228	9	Arcturus	0.6484	8	SmolLab_SEU	0.5738
11	ClimateSense	0.7226	10	ClimateSense	0.6453	9	Baseline	0.5719
12	NLP-UTB	0.7130	11	Baseline	0.6390	10	XplaiNLP	0.5665
13	UNAM	0.7075	12	Investigators	0.6292	11	JU_NLP	0.5603
14	CheckMates	0.7009	13	IIIT Surat	0.5411	12	ClimateSense	0.5525
15	DSGT-CheckThat	0.6830	14	AI Wizards	0.2380	13	QU-NLP	0.5165
16	CUET_KCRL	0.6783		**Romanian**		14	TIFIN INDIA	0.3811
17	CSECU-Learners	0.6777	1	QU-NLP	0.8126		**Greek**	
18	NapierNLP	0.6724	2	CSECU-Learners	0.7992	1	AI Wizards	0.5067
19	AI Wizards	0.6600	3	XplaiNLP	0.7917	2	SmolLab_SEU	0.4945
20	IIIT Surat	0.6492	4	SmolLab_SEU	0.7892	3	CSECU-Learners	0.4919
21	TIFIN India	0.5756	5	UmuTeam	0.7793	4	UmuTeam	0.4831
22	UGPLN	0.5531	6	CEA-LIST	0.7659	5	XplaiNLP	0.4750
23	Baseline	0.5370	7	AI Wizards	0.7507	6	Investigators	0.4539
	Ukrainian		8	JU_NLP	0.7442	7	CEA-LIST	0.4492
1	CSECU-Learners	0.6424	9	ClimateSense	0.7396	8	JU_NLP	0.4351
2	Investigators	0.6413	10	Arcturus	0.7366	9	Baseline	0.4159
3	ClimateSense	0.6395	11	Investigators	0.7133	10	ClimateSense	0.4137
4	AI Wizards	0.6383	12	IIIT Surat	0.6496	11	QU-NLP	0.4057
5	Baseline	0.6296	13	Baseline	0.6461	12	Arcturus	0.3905
6	SmolLab_SEU	0.6238	14	TIFIN INDIA	0.5181	13	IIIT Surat	0.3733
7	UmuTeam	0.6210				14	TIFIN India	0.3337
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						

Table 4
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 [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]		☑	☑	☑							☑																			

probabilistic thresholds, embedding-based classifiers, or LLM-based zero-shot and in-context learning methods. An overview of the approaches is given in Table 4 and a short description of the individual approaches for each team is given in the following.

4.1. Baselines

We used the same baseline introduced in the CheckThat! lab 2024 Task 2 edition [27]. In particular, the baseline was a multilingual SentenceBERT [53] model with a logistic regression classifier on top of it. We considered paraphrase-multilingual-MiniLM-L12-v2 model card as one of the current top-performing models for semantic similarity. We regularized the logistic regression classifier by applying class re-weighting to account for class imbalance. We trained the baseline model on individual language-specific training data and we evaluated it on the corresponding test set. In the case of zero-shot languages, we trained the baseline on the multilingual dataset, comprising Arabic, Bulgarian, English, German and Italian training splits.

4.2. Results per Language

Arabic. A total of 12 teams participated in the Arabic subtask, with five of them not surpassing the baseline score of 0.5133. Top submissions largely outperformed the baseline, setting a new high score. In particular, **CEA-LIST** [37] achieved a macro F1 score of 0.6884 using an ensemble of small language models and standard encoder-based transformers. The second-ranked team **UmuTeam** [46] reports a considerably lower score using MARBERTv2. Likewise **Investigators** [34], using general-purpose transformer models like DeBERTa and Multilingual BERT.

Italian. A total of 13 teams participated in the Italian subtask, with only three of them not surpassing the baseline score of 0.6941. Team **XplaiNLP** [43] ranks first with a F1 score of 0.8104, closely fdollowed by team **CEA-LIST** [37]. Team **SmolLab_SEU** [48] follows with a difference of 3%-points, still surpassing the baseline by a large margin.

German. A total of 14 teams participated in the German subtask, with only one team reporting performance below the baseline score of 0.6960. Team **SmolLab_SEU** [48] achieved the first place with a score of 0.8520. Team **UNAM** [52] follows with an F1 score of 0.8280. Lastly, team **QU-NLP** [50] ranks third with a F1 score of 0.8013. All three top teams largely outperform the baseline with an improvement of around 10-15%-points.

English. A total of 22 teams participated in the English subtask, all of them reporting classification performance above the baseline score of 0.5370. Team **QU-NLP** [50] ranked first with a F1 score of 0.8052 using a feature-augmented transformer model. A similar performance is reported by team **TIFIN INDIA** [39] with a F1 score of 0.7955. Team **CEA-LIST** [37] achieves third place with a F1 socre of 0.7739. The large majority of remaining submissions achieved similar results in the range [0.76 - 0.70].

Multilingual. A total of 13 teams participated in the multilingual subtasks, with only three teams reporting performance below the baseline score of 0.6390. Team **TIFIN INDIA** [39] ranked first (0.7550) with their ensemble of transformer-based models. Teams **CEA-LIST** [37] and **CSECU-Learners** [36] follow with similar classification performance of around $\sim$0.73.

Polish. A total of 13 teams participated in the Polish subtask, where more than half of the submissions outperformed the baseline score of 0.5719. In particular, team **CEA-LIST** [37] ranked first with a F1 score of 0.6922. Team **IIIT Surat** [38] reports a $\sim$3-points performance difference using multilingual BERT. Similarly, Team **CSECU-Learners** ranks third with a F1 score of 0.6676.

Ukrainian. A total of 13 teams participated in the Ukrainian subtask. Only four teams managed to outperform the baseline score of 0.6296, while reporting slightly superior performance. In particular, team **CSECU-Learners** [36] achieved first place with a F1 score of 0.6424. Team **Investigators** [34] follows with a F1 score of 0.6413 using a combination of encoder-based models like DeBERTa, BERT, multilingual BERT and Twitter RoBERTa. Team **ClimateSense** [40] reports a similar performance to the top-two teams.

Romanian. A total of 13 teams participated in the Romanian subtask, with only one team (i.e., **TIFIN INDIA** [39]) not surpassing the baseline score of 0.6461. Team **QU-NLP** [50] ranks first with a F1 score of 0.8126. There is a $\sim$2-points difference between the first-ranked team and the second- and third-ranked teams, namely team **CSECU-Learners** [36] and team XplaiNLP [43].

Greek. A total of 13 teams participated in the Greek subtask, with around half of the submissions not surpassing the baseline score of 0.4159. Team **AI Wizards** [33] ranks first by fine-tuning a probabilistic classifier on top of DeBERTaV3 model. Similar performance is reported by team **SmolLab_SEU** [48] (0.4945) and team **CSECU-Learners** [36] (0.4919).

4.3. Detailed Description of the Participating Systems

Below, we describe the approaches of all participating systems; see also Table 4 for an overview.

Team **AI Wizards** [33] employed a probabilistic classifier with a decision threshold, fine-tuning DeBERTaV3 for the task.

Team **Investigators** [34] utilized encoder-based models including DeBERTa, BERT, Multilingual BERT, and Twitter RoBERTa.

Team **DSGT-CheckThat** [35] 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** [36] 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** [37] 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** [38] 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** [39] 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 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** [40] 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** [41] pursued a supervised classification approach using an LSTM and fine-tuning mBERT.

Team **nlu@utn** [42] 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** [43] 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** [54] 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** [45] 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** [46] 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** [47] 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** [48] 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** [49] 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** [50] 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** [51] explored various models such as logistic regression, Support Vector Machine, BERT, Sentence-BERT, and DistilBERT.

Team **UNAM** [52] used different language-specific versions of the BERT model and focused on monolingual subtasks.

5. Conclusion and Future Work

We presented an overview of Task 1 from the CheckThat! lab at CLEF 2025. The task concerned the detection of subjective sentences in controversial news articles. The task was offered in nine different languages, four of which were addressed in a zero-shot setting.

In alignment with the previous edition of the task [27], the majority of the submissions relied on encoder-only transformer-based architectures, either tailored to a specific language or covering multilingualism. Some approaches also evaluated popular large language models like GPT with instruction tuning to detect subjectivity, data augmentation, and automatic translation. The most successful solutions coupled transformer-based classifiers with domain knowledge in the form of feature extraction or large language models in an ensemble fashion. The best macro F_1 scores ranged between 0.50 and 0.85, showing that annotating and detecting subjectivity present different challenges that are specific of the given language. Overall, there is still ample room for improvement in all subtasks. More precisely, in many cases, we observed that more than half of the teams did not surpass our baseline model.

As future work, we plan to collect more data concerning existing languages and to expand the set of covered languages to gather more insights about the task.

Acknowledgments

We are thankful to the volunteers that helped with the annotation such Ploutarchos Iliadis, Angeliki Dimopoulou, Fotini Giannopoulou, Foivos Andrianopoulos, Evangelos Sinos, Panagiotis Reppas who helped with the annotation of the zero-shot Greek test set.

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 F. Ruggeri is partially supported by the project European Commission's NextGeneration EU programme, PNRR – M4C2 – Investimento 1.3, Partenariato Esteso, PE00000013 - "FAIR - Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI" and by the European Union's Justice Programme under Grant Agreement No. 101087342 for the project "Principles Of Law In National and European VAT". K. Korre's research is carried out under the project *RACHS: Rilevazione e Analisi Computazionale dell'Hate Speech in rete*, in the framework of the PON programme

FSE REACT-EU, Ref. DOT1303118. A. Muti's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE).

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly for grammar and spelling check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Galassi, F. Ruggeri, A. Barrón-Cedeño, F. Alam, T. Caselli, M. Kutlu, J. Struss, F. Antici, M. Hasanain, J. Köhler, K. Korre, F. Leistra, A. Muti, M. Siegel, T. Mehmet Deniz, M. Wiegand, W. Zaghouani, Overview of the CLEF-2023 CheckThat! lab: Task 2 on subjectivity in news articles, in: M. Aliannejadi, G. Faggioli, N. Ferro, Vlachos, Michalis (Eds.), Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF 2023, Thessaloniki, Greece, 2023.

[2] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: A. F. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, volume 3406, Springer, Berlin and Heidelberg, 2005, pp. 486–497. doi:10.1007/978-3-540-30586-6_53.

[3] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: R. Mooney, C. Brew, L.-F. Chien, K. Kirchhoff (Eds.), Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Morristown, NJ and USA, 2005, pp. 347–354.

[4] R. Mihalcea, C. Banea, J. Wiebe, Learning multilingual subjective language via cross-lingual projections, in: A. Zaenen, A. van den Bosch (Eds.), ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic, The Association for Computer Linguistics, 2007, pp. 976–983.

[5] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan, Multilingual subjectivity analysis using machine translation, in: M. Lapata, H. T. Ng (Eds.), Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA and USA, 2008, pp. 127–135.

[6] D. Aleksandrova, F. Lareau, P.-A. Ménard, Multilingual sentence-level bias detection in wikipedia, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing, Incoma Ltd., Shoumen, Bulgaria, 2019, pp. 42–51. doi:10.26615/978-954-452-056-4_006.

[7] C. Hube, B. Fetahu, Neural based statement classification for biased language, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, New York, NY, USA, 2019, pp. 195–203. doi:10.1145/3289600.3291018.

[8] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, pp. 105–112.

[9] L. L. Vieira, C. L. M. Jeronimo, C. E. C. Campelo, L. B. Marinho, Analysis of the subjectivity level in fake news fragments, in: Proceedings of the Brazilian Symposium on Multimedia and the Web, Association for Computing Machinery, New York, NY, USA, 2020, pp. 233–240. doi:10.1145/3428658.3430978.

[10] C. L. M. Jerônimo, L. B. Marinho, C. E. C. Campelo, A. Veloso, A. S. Da Costa Melo, Fake News Classification Based on Subjective Language, in: M. Indrawan-Santiago, E. Pardede, I. L. Salvadori, M. Steinbauer, I. Khalil, G. Anderst-Kotsis (Eds.), Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services, iiWAS2019, Association for Computing Machinery, New York, NY, USA, 2020, pp. 15–24. doi:10.1145/3366030.3366039.

[11] J. Villena-Román, J. García-Morera, M. Á. G. Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. U. López, Overview of TASS 2015, in: J. Villena-Román, J. García-Morera, M. Á. G. Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. U. López (Eds.), Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015), Alicante, Spain, September 15, 2015, volume 1397 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015, pp. 13–21.

[12] N. Das, S. Sagnika, A subjectivity detection-based approach to sentiment analysis, in: D. Swain, P. K. Pattnaik, P. K. Gupta (Eds.), Machine Learning and Information Processing, Springer Singapore, Singapore, 2020, pp. 149–160.

[13] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 2004, pp. 271–278. doi:10.3115/1218955.1218990.

[14] J. M. Wiebe, R. F. Bruce, T. P. O'Hara, Development and use of a gold-standard data set for subjectivity classifications, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, College Park, Maryland, USA, 1999, pp. 246–253. doi:10.3115/1034678.1034721.

[15] T. Wilson, J. Wiebe, Annotating opinions in the world press, in: Proceedings of the SIGDIAL 2003 Workshop, The 4th Annual Meeting of the Special Interest Group on Discourse and Dialogue, July 5-6, 2003, Sapporo, Japan, The Association for Computer Linguistics, 2003, pp. 13–22.

[16] M. Abdul-Mageed, M. Diab, Subjectivity and sentiment annotation of Modern Standard Arabic newswire, in: N. Ide, A. Meyers, S. Pradhan, K. Tomanek (Eds.), Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 110–118.

[17] I. Chaturvedi, E. Cambria, R. E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Inf. Fusion 44 (2018) 65–77. doi:10.1016/j.inffus.2017.12.006.

[18] M. Geva, Y. Goldberg, J. Berant, Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 1161–1166. doi:10.18653/v1/D19-1107.

[19] P. Röttger, B. Vidgen, D. Hovy, J. B. Pierrehumbert, Two contrasting data annotation paradigms for subjective NLP tasks, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 175–190. doi:10.18653/v1/2022.naacl-main.13.

[20] F. Antici, F. Ruggeri, A. Galassi, K. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barrón-Cedeño, A corpus for sentence-level subjectivity detection on English news articles, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 273–285.

[21] S. Rustamov, E. Mustafayev, M. Clements, Sentence-level subjectivity detection using neuro-fuzzy models, in: A. Balahur, E. van der Goot, A. Montoyo (Eds.), Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 108–114.

[22] F. Benamara, B. Chardon, Y. Mathieu, V. Popescu, Towards context-based subjectivity analysis, in: H. Wang, D. Yarowsky (Eds.), Proceedings of 5th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 2011, pp. 1180–1188.

[23] F. Antici, L. Bolognini, M. A. Inajetovic, B. Ivasiuk, A. Galassi, F. Ruggeri, SubjectivITA: An italian corpus for subjectivity detection in newspapers, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021, volume 12880 of *LNCS*, Springer, 2021, pp. 40–52. doi:10.1007/978-3-030-85251-1_4.

[24] C. Banea, R. Mihalcea, J. Wiebe, Multilingual subjectivity: Are more languages better?, in: C.-R. Huang, D. Jurafsky (Eds.), Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 28–36.

[25] M. Abdul-Mageed, M. Diab, M. Korayem, Subjectivity and sentiment analysis of Modern Standard Arabic, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 587–591.

[26] C. Banea, R. Mihalcea, J. Wiebe, Sense-level subjectivity in a multilingual setting, Comput. Speech Lang. 28 (2014) 7–19. doi:10.1016/j.csl.2013.03.002.

[27] J. M. Struß, F. Ruggeri, A. Barrón-Cedeño, F. Alam, D. Dimitrov, A. Galassi, G. Pachov, I. Koychev, P. Nakov, M. Siegel, M. Wiegand, M. Hasanain, R. Suwaileh, W. Zaghouani, Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles, in: [55], 2024.

[28] R. Suwaileh, M. Hasanain, F. Hubail, W. Zaghouani, F. Alam, ThatiAR: Subjectivity detection in arabic news sentences, arXiv: 2406.05559 (2024).

[29] Z. Sheikh Ali, W. Mansour, T. Elsayed, A. Al-Ali, AraFacts: The first large Arabic dataset of naturally occurring claims, in: N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, S. Touileb (Eds.), Proceedings of the Sixth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 2021, pp. 231–236. URL: https://aclanthology.org/2021.wanlp-1.26.

[30] F. Ruggeri, F. Antici, A. Galassi, K. Korre, A. Muti, A. Barrón-Cedeño, On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection, in: R. Campos, A. M. Jorge, A. Jatowt, S. Bhatia, M. Litvak (Eds.), Text2Story@ECIR, volume 3370 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 103–111.

[31] J. Köhler, G. K. Shahi, J. M. Struß, M. Wiegand, M. Siegel, T. Mandl, M. Schütz, Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[32] K. Krippendorff, Computing krippendorff's alpha-reliability, 2011. URL: https://repository.upenn.edu/handle/20.500.14332/2089.

[33] M. Fasulo, L. Babboni, L. Tedeschini, AI wizards at checkthat! 2025: Enhancing transformer-based embeddings with sentiment for subjectivity detection in news articles, in: [56], 2025.

[34] S. M. A. Hashmi, S. Aamir, M. Anas, T. Usmani, F. Alvi, A. Samad, Investigators at checkthat! 2025: Using llms to improve fact-checking, in: [56], 2025.

[35] M. Heil, D. Bang, DS@GT at checkthat! 2025: Detecting subjectivity via transfer-learning and corrective data augmentation, in: [56], 2025.

[36] M. Ahmad, A. N. Chy, Csecu-learners at checkthat! 2025: Multilingual transformer-based approach for subjectivity detection in news articles across multilingual and zero-shot settings, in: [56], 2025.

[37] A. Elbouanani, E. Dufraisse, A. Tuo, A. Popescu, Cea-list at checkthat! 2025: Evaluating llms as detectors of bias and opinion in text, in: [56], 2025.

[38] S. C. Jaiswal, R. Kumar, Iiit surat at checkthat! 2025: Identifying subjectivity from multilingual text sequence, in: [56], 2025.

[39] K. Gurumurthy, A. Shrivastava, P. K. Rajpoot, P. Devadiga, B. Hazarika, M. Jain, M. Sharma, A. Suneesh, A. B. Suresh, A. U. Baliga, Tifin at checkthat! 2025: Cross-lingual subjectivity classification in news through monolingual, multilingual, and zero-shot learning, in: [56], 2025.

[40] G. Burel, P. Lisena, E. Daga, R. Troncy, H. Alani, ClimateSense at CheckThat! 2025: Combining fine-tuned large language models and conventional machine learning models for subjectivity and scientific web discourse analysis, in: [56], 2025.

[41] M. T. A. Shawon, F. Haq, M. A. Mia, G. S. M. Mursalin, M. I. Khan, CUET_KCRL at checkthat!2025: Ensemblenet with roberta-large for subjectivity detection in news articles, in: [56], 2025.
[42] S. Meyer, M. Roth, nlu@utn at checkthat! 2025: Combining bias sensitivity, linguistic features, and persuasion cues in an ensemble for subjectivity detection, in: [56], 2025.
[43] A. Sahitaj, J. Li, P. W. Neves, P. S. Fedor Splitt, C. Jakob, V. Solopova, V. Schmitt, XplaiNLP at checkthat! 2025: Multilingual subjectivity detection with finetuned transformers and prompt-based inference with large language models, in: [56], 2025.
[44] S. Debnath, D. Das, JU_NLP at checkthat! 2025: A confidence-guided transformer-based approach for multilingual subjectivity classification, in: [56], 2025.
[45] K. Alexander, M. Z. Ullah, D. Gkatzia, NapierNLP at checkthat! 2025: Detecting subjectivity with llms and model fusion, in: [56], 2025.
[46] T. B. Beltrán, R. Pan, J. A. G. Díaz, R. V. García, UmuTeam at checkthat! 2025: Language-specific versus multilingual models for fact-checking, in: [56], 2025.
[47] M. del Carmen Toapanta-Bernabé, M. Ángel Garcia-Cumbreras, L. A. Ureña-López, D. D. M. Intriago, J. S. Holguín-Reyes, Sinai-UGPLN at checkthat! 2025: A hybrid sbert–logistic regression framework for segment-level subjectivity detection in english news, in: [56], 2025.
[48] M. A. Rahman, M. A. Amin, M. S. Dewan, M. J. Hasan, M. A. Rahman, SmolLab_SEU at checkthat! 2025: How well do multilingual transformers transfer across news domains for cross-lingual subjectivity detection?, in: [56], 2025.
[49] A. Aditya, R. Jambulkar, S. Pal, Arcturus at checkthat! 2025: Deberta-v3-base for multilingual subjectivity detection in news articles, in: [56], 2025.
[50] M. Al-Smadi, Qu-nlp at checkthat! 2025: Multilingual subjectivity in news articles detection using feature-augmented transformer models with sequential cross-lingual fine-tuning, in: [56], 2025.
[51] R. Padmashri, K. V., V. Srikumar, D. Thenmozhi, Checkmates at checkthat! 2025: Transformer-based models for subjectivity classification, in: [56], 2025.
[52] I. Diaz, J. Barco, J. Hernández, E. Lee-Romero, G. Bel-Enguix, cepanca_UNAM at checkthat! 2025: Using bert-based classifiers for detection of subjectivity, in: [56], 2025.
[53] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. doi:10.18653/V1/D19-1410.
[54] A. A. M. Sardar, K. Fatema, M. A. Islam, JUNLP at CheckThat! 2024: Enhancing check-worthiness and subjectivity detection through model optimization, in: [55], 2024.
[55] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, 2024.
[56] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.