

# Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity in News Articles

Notebook for the CheckThat! Lab at CLEF 2023

Andrea Galassi<sup>1,\*</sup>, Federico Ruggeri<sup>1,\*</sup>, Alberto Barrón-Cedeño<sup>1</sup>, Firoj Alam<sup>2</sup>, Tommaso Caselli<sup>3</sup>, Mucahid Kutlu<sup>4</sup>, Julia Maria Struß<sup>5</sup>, Francesco Antici<sup>1</sup>, Maram Hasanain<sup>2</sup>, Juliane Köhler<sup>5</sup>, Katerina Korre<sup>1</sup>, Folkert Leistra<sup>3</sup>, Arianna Muti<sup>1</sup>, Melanie Siegel<sup>6</sup>, Mehmet Deniz Türkmen<sup>4</sup>, Michael Wiegand<sup>7</sup> and Wajdi Zaghouani<sup>2</sup>

<sup>1</sup>University of Bologna, Italy

<sup>2</sup>Qatar Computing Research Institute, HBKU, Qatar

<sup>3</sup>University of Groningen, Netherlands

<sup>4</sup>TOBB University of Economics and Technology, Türkiye

<sup>5</sup>University of Applied Sciences Potsdam, Germany

<sup>6</sup>Darmstadt University of Applied Sciences, Germany

<sup>7</sup>University of Klagenfurt, Austria

## Abstract

We describe the outcome of the 2023 edition of the CheckThat! Lab at CLEF. We focus on subjectivity (Task 2), which has been proposed for the first time. It aims at fostering the technology for the identification of subjective text fragments in news articles. For that, we produced corpora consisting of 9,530 manually-annotated sentences, covering six languages —Arabic, Dutch, English, German, Italian, and Turkish. Task 2 attracted 12 teams, which submitted a total of 40 final runs covering all languages. The most successful approaches addressed the task using state-of-the-art multilingual transformer models, which were fine-tuned on language-specific data. Teams also experimented with a rich set of other neural architectures, including foundation models, zero-shot classifiers, and standard transformers, mainly coupled with data augmentation and multilingual training strategies to address class imbalance. We publicly release all the datasets and evaluation scripts, with the purpose of promoting further research on this topic.

---

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

\*Corresponding author.

✉ a.galassi@unibo.it (A. Galassi); federico.ruggeri6@unibo.it (F. Ruggeri); a.barron@unibo.it (A. Barrón-Cedeño); fialam@hbku.edu.qa (F. Alam); t.caselli@rug.nl (T. Caselli); m.kutlu@etu.edu.tr (M. Kutlu); julia.struss@fh-potsdam.de (J. M. Struß); francesco.antici@unibo.it (F. Antici); mhasanain@hbku.edu.qa (M. Hasanain); juliane.koehler@fh-potsdam.de (J. Köhler); aikaterini.korre2@unibo.it (K. Korre); f.a.leistra@student.rug.nl (F. Leistra); arianna.muti2@unibo.it (A. Muti); melanie.siegel@h-da.de (M. Siegel); m.turkmen@etu.edu.tr (M. D. Türkmen); michael.wiegand@aau.at (M. Wiegand); wzaghouani@hbku.edu.qa (W. Zaghouani)

ORCID 0000-0001-9711-7042 (A. Galassi); 0000-0002-1697-8586 (F. Ruggeri); 0000-0003-4719-3420 (A. Barrón-Cedeño); 0000-0001-7172-1997 (F. Alam); 0000-0003-2936-0256 (T. Caselli); 0000-0002-5660-4992 (M. Kutlu); 0000-0001-9133-4978 (J. M. Struß); 0000-0000-0000-0000 (F. Antici); 0000-0001-7172-1997 (M. Hasanain); 0000-0002-7175-5895 (J. Köhler); 0000-0002-9349-9554 (K. Korre); 0009-0004-4981-5961 (F. Leistra); 0000-0002-3387-6557 (A. Muti); 0000-0002-5064-5750 (M. Siegel); 0000-0002-5403-1078 (M. Wiegand); 0000-0003-1521-5568 (W. Zaghouani)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

# 1. Introduction

The CheckThat! Lab [1, 2] was organized for the 6<sup>th</sup> time in the framework of CLEF 2023. This paper presents an overview of Task 2 on the identification of subjectivity in news articles, which is organized for the first time.<sup>1</sup>

In objective sentences, the information is usually presented in a straightforward way. Instead, subjective sentences often include the use of specific vocabulary, figures of speech, or other elements that make it more difficult to analyze by machine learning models.

In the context of fact-checking, objective sentences can be directly fed to a fact-checking pipeline for verification, while subjective ones require an additional processing step, it aims at extracting a claim or simply discarding its information. Moreover, the presence of subjective content may be a useful feature that could facilitate downstream tasks in the fact-checking pipeline, such as political bias [4] and factuality reporting [5].

Given a set of sentences taken from a news article,<sup>2</sup> Task 2 requires classifying each of the sentences as subjective or objective. A sentence is considered **subjective** if its contents are based on or influenced by personal feelings, tastes, or opinions. Otherwise, the sentence is considered **objective**. The task is offered in Arabic, Dutch, English, Italian, German, and Turkish, with an additional multilingual setting.<sup>3</sup>

The task attracted 12 participants, for a total of 40 final submissions. Submitted approaches include large language models, generative models, such as ChatGPT and GPT-3, pre-training over multilingual data, data augmentation techniques, ensembles of classifiers, and feature selection. Transformer-based architectures showed to be the most successful, especially when pre-trained over multilingual data or when considering augmented data. Nonetheless, the task is not yet solved and there is still room for improvement.

The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the multilingual data we produced for the task. Section 4 overviews the models proposed by the different participants and the results they obtained in the task. Section 5 closes with final remarks and potential future work.

## 2. Related Work

Previous studies have explored the contribution of subjectivity detection (SD) technology to well-known downstream tasks, such as sentiment analysis [7, 8, 9] and bias detection [10, 11]. SD can also influence other tasks such as claim extraction [12, 13] and, crucially for our context, fact-checking [14, 15, 16, 17].

The inherent difficulty of providing a practical definition of subjectivity [8, 18] has led to several formulations for the task at hand, as it is often influenced by domain-specific assumptions and a lack of a schematic definition, in particular for data collection. In previous work, corpora for SD were developed in several different ways, such as relying on domain-specific assumptions [19, 12, 20, 21, 22] or statistical methods [23, 24]. We instead rely on a prescriptive approach [25], in

---

<sup>1</sup>Refer to Alam et al. [3], Da San Martino et al. [4], Nakov et al. [5], Haouari et al. [6] to read about Tasks 1, 3, 4, and 5, respectively.

<sup>2</sup>We note that Turkish dataset utilizes sentences taken from tweets.

<sup>3</sup>All the data and scripts are publicly available at <https://checkthat.gitlab.io/clef2023/task2/>.

**Table 1**  
Statistics of the datasets for all six languages.

Language	Training			Development			Test		
	Tot	OBJ	SUBJ	Tot	OBJ	SUBJ	Tot	OBJ	SUBJ
Arabic	1,185	905	280	297	227	70	445	363	82
Dutch	800	489	311	200	107	93	500	263	237
English	830	532	298	219	106	113	243	116	127
German	800	492	308	200	123	77	291	194	97
Italian	1,613	1,231	382	227	167	60	440	323	117
Turkish	800	422	378	200	100	100	240	129	111

which SD is conceived as a step that can contribute to tasks such as claim extraction and fact-checking, and data are collected framing the task to domain-specific objectives and proposing pragmatic annotation.

Following Chaturvedi et al. [8], we distinguish between syntactic and semantic solutions for SD. The first category of approaches relied on keyword spotting [19, 12] or lexicons [20, 21, 22, 26]. In contrast, the semantic category encompasses statistical methods [23, 24] or neural architectures, such as convolutional neural networks [27], deep belief networks [28], and transformer architectures like BERT [29]. To the best of our knowledge, a systematic approach for SD leveraging state-of-the-art language models is yet to be proposed.

For what concerns language coverage, most studies have focused on English alone. Some contributions have extended to other languages, such as Arabic [30], German [30], French [26, 30], Italian [29], Romanian [13, 30], and Spanish [30]. Most of these attempts have mainly relied on machine translation and monolingual ontologies, which inevitably introduce noise when jumping across languages. In order to produce the datasets for this task, we annotate corpora in six languages from scratch, relying on native (or near-native) speakers of all languages, and following a common set of language agnostic guidelines [31].

### 3. Datasets

All datasets considered in the task were created following the annotation guidelines presented in [31]. In total, about 10k sentences were considered for the task. Table 1 gives statistics of the corpora in all languages, while Table 2 shows examples of subjective and objective sentences.

#### 3.1. Arabic

Arabic news articles were annotated by three native speakers. The annotators chosen for the subjectivity annotation task have diverse Arabic-speaking backgrounds, including Egyptian, Yemeni, and Bahraini. Each annotator is proficient in Modern Standard Arabic (MSA) but brings their own dialect and unique forms of expression. One annotator has expertise in linguistics and computational skills, while the other two annotators specialize in the humanities, specifically digital and political domains.

**Table 2**

Examples of subjective and objective sentences in annotated datasets.

Language	Sentence	Class
Arabic	الدكتور سامي الخيمي واللواء بهجت سليمان سفيران للأسد في حرب لفظية طاحنة.	SUBJ
	وكما هو معلوم فوجود الأوزون يحمي الحياة على الأرض من الأشعة فوق البنفسجية المنبعثة من الشمس.	OBJ
Dutch	<i>De nieuwe status van Bonaire, Sint Eustatius en Saba is een stap naar verbetering.</i>	SUBJ
	<i>Dante slaagde erin om de hel te verruilen voor de relatief milde vlammen van het Vagevuur.</i>	OBJ
English	<i>While it's misguided to put all focus or hope onto one section of the working class, we can't ignore this immense latent power that logistics workers possess.</i>	SUBJ
	<i>Workers would have a 24 percent wage increase by 2024, including an immediate 14 percent raise.</i>	OBJ
German	<i>Für die Pandemie-Macher ist es zugleich von strategischer Bedeutung, die Kontrollgruppe der Ungeimpften zu eliminieren – und dies möglichst schnell.</i>	SUBJ
	<i>Der andere Angeklagte bekundete, er könne sich an den ganzen Vorgang nicht erinnern.</i>	OBJ
Italian	<i>Hanno festeggiato il matrimonio come se non ci fosse il coronavirus.</i>	SUBJ
	<i>Tutti sono stati identificati e multati per aver violato le norme anti contagio per il contenimento del fenomeno epidemico.</i>	OBJ
Turkish	<i>Kılıçdaroğlu laikliğe aykırı davranmaya devam ediyor: Bu sefer Kur'an öpüp başına koydu.</i>	SUBJ
	<i>Akşener basına seslendi: Emekçilerin günlük hayatlarını yaşanır hale getirin.</i>	OBJ

We ensured annotators' suitability for the task by selecting university-level annotators with strong Arabic language background, including research experience or relevant degrees. They underwent a screening test and received two to three weeks of training, which included group discussion of annotation tasks, guideline reading, and meetings with the annotation lead annotator.

Data Collection involved four phases. In Phase 1, Arabic sentences from news articles were selected, filtered, and parsed. In Phase 2, sentences from the 12 most frequent news domains were chosen for annotation. Due to labeling skew, Phase 3 focused on selecting sentences with a higher probability of subjectivity using an SVM classifier. Finally, in Phase 4, the annotated sentences were reviewed, filtering out uncertain labels, and acquiring the majority label per sentence for the released dataset.

### 3.2. Dutch

All the Dutch sentences were sourced from the DPG Media 2019 dataset [32]<sup>4</sup>. This dataset contains partisanship annotations for Dutch newspaper articles based on two methods: publisher-level and article-level. For the task, only the article-level annotations were retained as they were based on the actual content of the article, thus ensuring a higher annotation quality. All articles annotated as containing some form of partisanship have been gathered and then split into sentences. Next, a total of 1,500 sentences have been manually annotated by one native speaker. In order to evaluate the generalizability of the participating systems, articles from publishers that were not present in the training set have been intentionally kept for testing purposes.

### 3.3. English

For English, we use the corpus created by Antici et al. [31] for the training and development splits. The corpus was created by annotating sentences from articles on controversial topics published in popular outlets.<sup>5</sup> Six annotators took part in the annotation effort, with each instance being judged by two of them. Annotators gathered together later on to discuss and solve disagreements, relying on a seventh annotator to solve conflicts when necessary. We develop a novel test set following the same procedure, containing 243 sentences that come from the same news outlets as the other partitions. The Krippendorff’s alpha inter-annotator agreement (IAA) on the test set was 0.85 (nearly perfect agreement), similar to the 0.83 of the training and development splits.

### 3.4. German

The training, development and test set for German has been assembled by randomly selecting sentences from the CT!2022FAN-Corpus [33] consisting of news articles that have been annotated according to the factuality of their main claim [34]. Each sentence has been annotated by two annotators. In total five native speakers were involved in the annotation process. Conflicts were solved by asking a third annotator for their judgement.

### 3.5. Italian

The training and development data for Italian is mostly derived from the SubjectivITA corpus [29] and consists of about 1,841 sentences. Rather than using the original annotation, we re-annotated the corpus following our up-to-date guidelines. The re-annotation resulted in the class-switching of 157 sentences. As for the test set, we released a novel collection of 440 sentences, gathered from popular Italian news outlets.<sup>6</sup> The annotation follows the same methodology used for the English dataset, involving five annotators plus one to solve conflicts. The IAA score on this novel test set is 0.91, which corresponds to nearly perfect agreement.

<sup>4</sup><https://github.com/dpgmedia/partisan-news2019>

<sup>5</sup>The outlets are [tribunemag.co.uk](http://tribunemag.co.uk), [spectator.co.uk](http://spectator.co.uk), [shtfplan.com](http://shtfplan.com), [vdare.com](http://vdare.com), [theweek.com](http://theweek.com), [frontpagemag.com](http://frontpagemag.com), [economist.com](http://economist.com), and [theguardian.com](http://theguardian.com).

<sup>6</sup>We considered the following websites: [corriere.it](http://corriere.it), [avantionline.it](http://avantionline.it), [ilpost.it](http://ilpost.it), [avvenire.it](http://avvenire.it), [repubblica.it](http://repubblica.it), [ilfattoquotidiano.it](http://ilfattoquotidiano.it), [ilgiornale.it](http://ilgiornale.it), [ansa.it](http://ansa.it), [ilfoglio.it](http://ilfoglio.it), [liberoquotidiano.it](http://liberoquotidiano.it).

### 3.6. Turkish

In order to construct the Turkish dataset, instead of using news articles to extract sentences, we utilized sentences in tweets. In particular, we first crawled Turkish tweets tracking keywords about politics. Subsequently, we removed similar tweets and then selected candidate tweets to be annotated. As judging tweets with incomplete and/or multiple sentences would be problematic, we manually selected a single sentence from each tweet and discarded the unsuitable ones. Two annotators first judged each sentence independently. Subsequently, they discussed with each other to reach an agreement on the sentences they disagreed on. We discarded the sentences on which the two annotators disagreed even after their discussion.

### 3.7. Multilingual

The multilingual dataset is composed of sentences sampled from the other datasets. For training, we proposed to use data from other datasets. We proposed a development set resulting from the aggregation of 50 subjective and 50 objective sentences randomly sampled from the respective development set in all the languages. The same procedure was followed for the test set, using the test sets from the other languages. While some teams followed our partition for training and development, others preferred to create their own splits.

## 4. Overview of the Systems and Results

Task 2 is formally defined as follows. Given a sentence  $s$ , extracted either from a news article or from a tweet (as in the case of Turkish), determine whether  $s$  is influenced by the subjective view of its author (class SUBJ) or presents an objective view of the covered topic (class OBJ).

A total of 12 teams participated to this task, with most teams targeting more than one language, be it with the same or with different approaches. The participants experimented with multiple models from the BERT family, as well as with generative models.

Table 3 offers a snapshot of the approaches, whereas Table 4 reports the performance results for all submissions, ranked on the basis of macro-averaged  $F_1$ .<sup>7</sup>



For the baselines, we implemented a logistic regressor trained on a multilingual Sentence-BERT [45] representation of the data. For each language, we trained the baseline on the respective training data alone.






















Five of the teams experimented with the use of generative pre-trained transformers (GPT) [36, 38, 43] with different levels of success. Team DWReCo [36] obtained the top performance in English and first runner-up in Turkish by using GPT-3 to reduce class imbalance with propaganda style-based data augmentation. The styles are identified from the journalistic checklist to identify subjective news. Team **Fraunhofer SIT** [38] used GPT-3 as well, but their few-shot classification barely afforded to beat the baseline. Team **TUDublin** [43] also performed data augmentation, this time using ChatGPT. Still, their classification model, built on top of M-BERT did not improve over the baseline in any of the languages they participated in. Team

---

<sup>7</sup>We decided to use the macro-averaged  $F_1$  rather than the  $F_1$  on the positive classes, as usually done in binary tasks, to overcome its limitation in contexts where the distribution of the classes is heavily imbalanced [44].

**Table 3**

Overview of the approaches to **Task 2**. The numbers in the language box refer to the position of the team in the official ranking; =part of the official submission; =considered in internal experiments.

Team	Languages						Models											Misc								
	Multilingual	Arabic	Dutch	German	English	Italian	Turkish	BERT	RoBERTa	XLNet	RoBERTa	GigaBERT	M-BERT	M-DeBERTa	S-BERT	SetFit	ChatGPT	GPT-3	BART	LSTM	Gradient Boosting	Multi-lingual training	Data augmentation	Feature Selection	Ensemble	
Accenture [35]	3	5	7	8	3	4																				
Awakened				10																						
DWReCo [36]			4	1	2																					
ES-VRAI [37]	-																									
Fraunhofer SIT [38]			5	6																						
Gpachov [39]				2																						
KUCST				4																						
NN [40]	1	1	2	2	5	2	3																			
tarrekko	-	-	-	-	-	-	-																			
Thesis Titan [41]	2	2	1	1	3	1	1																			
TOBB ETU [42]	4	5	3	3	9	5	6																			
TUDublin [43]				11	6																					

- Run submitted after the deadline.

**TOBB ETU** employed ChatGPT to directly classify the texts, experimenting with both zero- and few-shot classification.

The second best performance in English was obtained by team **Gpachov** [39], who used an ensemble of three distinct models: XLM-RoBERTa, Sentence BERT (S-BERT), and SetFit.

As observed in previous years, paying attention to all the languages paid back again. Team **Thesis Titan** [41] developed multiple fine-tuned models using mDeBERTaV3-base [46], starting from a newly developed multilingual dataset. While keeping the training data fixed, they used language-specific validation sets to optimize the models for each language, as well as to identify optimal language-specific hyperparameters. This approach resulted in the top performance in Dutch, German, Italian and Turkish, being the second best in Arabic and the multilingual setting. Team **NN** [40] relied on the multilingual XLM-RoBERTa. This approach resulted in the top performance in the multilingual setting, as well as in Arabic, with a top-3 performance in Dutch, German, Italian, and Turkish.

The baseline logistic regressors are competitive in all settings, obtaining a score of at least 0.64 and often bitting participant approaches. At least half of the submissions surpassed the baseline, with a difference with the best approaches that range from 18 percentage points (German) to 6 percentage points (English).

Next we visit the landscape for the multilingual and for each language setting.

**Table 4**

Results for the official submissions for the multilingual and for all six languages.

Team	F1	Team	F1	Team	F1
<b>Multilingual</b>		<b>English</b>		<b>Italian</b>	
1 NN [40]	81.97	- tarrekko *	78.19	1 Thesis Titan [41]	75.75
- tarrekko *	81.16	1 DWReCo [36]	78.18	- tarrekko *	71.61
2 Thesis Titan [41]	81.00	2 Gpachov [39]	77.34	2 NN [40]	71.01
- ES-VRAI [37]	77.96	3 Thesis Titan [41]	76.78	3 Accenture [35]	65.52
3 <i>baseline</i>	73.56	4 KUCST *	73.07	4 <i>baseline</i>	63.70
4 TOBB ETU [42]	66.62	5 NN [40]	72.84	5 TOBB ETU [42]	63.35
<b>Arabic</b>		6 Fraunhofer SIT [38]	72.72	6 TUDublin [43]	45.92
1 NN [40]	78.75	7 <i>baseline</i>	71.98	<b>Turkish</b>	
- tarrekko *	78.66	8 Accenture [35]	68.90	1 Thesis Titan [41]	89.94
2 Thesis Titan [41]	77.53	9 TOBB ETU [42]	63.46	- tarrekko *	87.01
3 Accenture [35]	72.53	10 Awakened *	60.41	2 DWReCo [36]	84.11
4 <i>baseline</i>	65.75	11 TUDublin [43]	40.32	3 NN [40]	81.21
5 TOBB ETU [42]	64.51	<b>German</b>		4 Accenture [35]	78.11
<b>Dutch</b>		1 Thesis Titan [41]	81.52	5 <i>baseline</i>	77.40
† 1 Thesis Titan [41]	81.43	2 NN [40]	74.13	6 TOBB ETU [42]	70.16
- tarrekko *	77.74	- tarrekko *	73.08		
2 NN [40]	75.57	3 TOBB ETU [42]	71.19		
3 TOBB ETU [42]	73.01	4 DWReCo [36]	69.82		
4 <i>baseline</i>	66.68	5 Fraunhofer_SIT [38]	68.39		
5 Accenture [35]	62.32	6 <i>baseline</i>	63.65		
		7 Accenture [35]	25.58		

- Run submitted after the deadline.

† Team involved in the preparation of the data.

\*No working note submitted.

**Multilingual.** Five teams submitted runs to the multilingual setting. **NN [40]** obtained the first position with their approach based on XLM-RoBERTa. It is interesting to notice that, in the monolingual settings, team **Thesis Titan** obtained a better score than team **NN** in 5 out of 6 languages, and than team **tarrekko** in 4 out of 6 languages. Nevertheless, their approach is not cross-language, since each a model is built independently for each setting.

**Arabic.** Five teams submitted their results, with team **NN** obtaining the best result of 0.79. The three best approaches obtained a similar score (from 0.78 to 0.79), largely surpassing the baseline score of 0.66.

**Dutch.** Five teams participated, with **Thesis Titan** obtaining the best result of 0.81, surpassing the baseline by about 15 percentage points.



**English.** Out of 11 submissions, only seven surpassed the approach, with team **DWReCo** [36] obtaining the first place. The four best approaches achieved similar scores, falling within the range of [0.77,0.78]. Similarly, the three approaches ranked between fourth and sixth position obtained a comparable result, with a score of approximately 0.73, which is slightly higher than the baseline.

**German.** We received seven submissions, and **Thesis Titan** achieved the highest result of 0.82, surpassing the baseline by 18 percentage points.

**Italian.** Six teams participated, and **Thesis Titan** achieved the highest result of 0.76, surpassing the second-best result from team **NN** by 0.05 points and the baseline by 0.12 points.

**Turkish.** Team **Thesis Titan** obtained the highest results among the six submissions, outperforming the baseline by approximately 13 percentage points.

## 5. Conclusion

We have presented a detailed overview of Task 2 of the CheckThat! Lab of CLEF 2023. The lab was focused on the detection of subjectivity in sentences extracted from news articles. Following the objectives of CLEF, we offered the task in six different languages and in a multilingual setting, thus fostering multilinguality.

Most of the submissions focused on the use of pre-trained models, some exploited more recent dialogue-based technologies such as ChatGPT. The most successful approaches incorporated additional knowledge in their model through multilingual pre-training of the models or data augmentation. The best macro F1 scores ranged between 0.75 to 0.82, showing that the task can be successfully addressed but is not yet completely solved.

Future work will be centered on extending high-quality multilingual datasets and broadening the scope by including document-level classification settings.

## Acknowledgments

We are thankful to the volunteers that helped with the annotation of the data such as A. Bardi, A. Fedotova, and K. Ebermanns. The work of A. Galassi is supported by the European Commission NextGeneration EU programme, PNRR-M4C2-Investimento 1.3, PE00000013-“FAIR” - Spoke 8. A. Muti is supported by the program *Progetti di formazione per la ricerca: Big Data per una regione europea più ecologica, digitale e resiliente*—Alma Mater Studiorum—Università di Bologna, Ref. 2021-15854. K. Korre is supported by the PON programme FSE REACT-EU, Ref. DOT1303118. The work related to the Arabic language was partially made possible by NPRP grant NPRP13S-0206-200281 and NPRP 14C-0916-210015 from the Qatar National Research Fund (a member of Qatar Foundation). The work related to the Turkish language was funded by the Scientific and Technological Research Council of Turkey (TUBITAK) ARDEB 3501 Grant No 120E514. The work related to the German data has partially been funded by the BMBF (German Federal Ministry of Education and Research) under the grant no. 01FP20031J. The responsibility for the

contents of this publication lies with the authors. The findings achieved herein are solely the responsibility of the authors.

## References

- [1] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The CLEF-2023 CheckThat! Lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), *Advances in Information Retrieval*, Springer Nature Switzerland, Cham, 2023, pp. 506–517.
- [2] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouni, Overview of the CLEF-2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, 2023.
- [3] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouni, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content, in: [47], 2023.
- [4] G. Da San Martino, F. Alam, M. Hasanain, R. N. Nandi, D. Azizov, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 3 on political bias of news articles and news media, in: [47], 2023.
- [5] P. Nakov, F. Alam, G. Da San Martino, M. Hasanain, R. N. Nandi, D. Azizov, P. Panayotov, Overview of the CLEF-2023 CheckThat! lab task 4 on factuality of reporting of news media, in: [47], 2023.
- [6] F. Haouari, Z. Sheikh Ali, T. Elsayed, Overview of the CLEF-2023 CheckThat! lab task 5 on authority finding in twitter, in: [47], 2023.
- [7] A. Stepinski, V. O. Mittal, A fact/opinion classifier for news articles, in: W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, July 23-27, 2007, ACM, 2007, pp. 807–808. URL: <https://doi.org/10.1145/1277741.1277919>. doi:10.1145/1277741.1277919.
- [8] 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. URL: <https://doi.org/10.1016/j.inffus.2017.12.006>. doi:10.1016/j.inffus.2017.12.006.
- [9] S. Clematide, S. Gindl, M. Klenner, S. Petrakis, R. Remus, J. Ruppenhofer, U. Waltinger, M. Wiegand, MLSA — a multi-layered reference corpus for German sentiment analysis, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 3551–3556. URL: [http://www.lrec-conf.org/proceedings/lrec2012/pdf/125\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/125_Paper.pdf).

- [10] D. Aleksandrova, F. Lareau, P. A. Ménard, Multilingual sentence-level bias detection in Wikipedia, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 42–51. URL: <https://aclanthology.org/R19-1006>. doi:10.26615/978-954-452-056-4\_006.
- [11] C. Hube, B. Fetahu, Neural based statement classification for biased language, in: J. S. Culpepper, A. Moffat, P. N. Bennett, K. Lerman (Eds.), Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019, ACM, 2019, pp. 195–203. URL: <https://doi.org/10.1145/3289600.3291018>. doi:10.1145/3289600.3291018.
- [12] 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. URL: <https://aclanthology.org/W03-1014>.
- [13] C. Banea, R. Mihalcea, J. Wiebe, Sense-level subjectivity in a multilingual setting, *Comput. Speech Lang.* 28 (2014) 7–19. URL: <https://doi.org/10.1016/j.csl.2013.03.002>. doi:10.1016/j.csl.2013.03.002.
- [14] L. L. Vieira, C. L. M. Jerônimo, C. E. C. Campelo, L. B. Marinho, Analysis of the subjectivity level in fake news fragments, in: C. de Salles Soares Neto (Ed.), *WebMedia '20: Brazillian Symposium on Multimedia and the Web*, São Luís, Brazil, November 30 - December 4, 2020, ACM, 2020, pp. 233–240. URL: <https://doi.org/10.1145/3428658.3430978>. doi:10.1145/3428658.3430978.
- [15] 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: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, iiWAS 2019, Munich, Germany, December 2-4, 2019, ACM, 2019, pp. 15–24. URL: <https://doi.org/10.1145/3366030.3366039>. doi:10.1145/3366030.3366039.
- [16] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. D. S. Martino, A. Abdelali, N. Durrani, K. Darwish, A. Al-Homaid, W. Zaghrouani, T. Caselli, G. Danoe, F. Stolk, B. Bruntink, P. Nakov, Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society, in: *Findings of EMNLP 2021*, 2021, pp. 611–649.
- [17] P. Nakov, F. Alam, S. Shaar, G. Da San Martino, Y. Zhang, COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 997–1009. URL: <https://aclanthology.org/2021.ranlp-1.113>.
- [18] 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. URL: <https://aclanthology.org/W03-2102/>.
- [19] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 486–497.
- [20] 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.

- [21] H. Yu, V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP '03, Association for Computational Linguistics, USA, 2003, p. 129–136. URL: <https://doi.org/10.3115/1119355.1119372>. doi:10.3115/1119355.1119372.
- [22] J. Villena-Román, J. García-Morera, M. Á. G. Cumbreiras, 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. Cumbreiras, 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. URL: <http://ceur-ws.org/Vol-1397/overview.pdf>.
- [23] 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. URL: <https://aclanthology.org/P04-1035>. doi:10.3115/1218955.1218990.
- [24] F. Sha, F. C. N. Pereira, Shallow parsing with conditional random fields, in: M. A. Hearst, M. Ostendorf (Eds.), Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003, The Association for Computational Linguistics, 2003, pp. 213–220. URL: <https://aclanthology.org/N03-1028/>.
- [25] 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. URL: <https://ceur-ws.org/Vol-3370/paper10.pdf>.
- [26] F. Benamara, B. Chardon, Y. Mathieu, V. Popescu, Towards context-based subjectivity analysis, in: Proceedings of 5th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 2011, pp. 1180–1188. URL: <https://aclanthology.org/I11-1132>.
- [27] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers, The Association for Computer Linguistics, 2014, pp. 655–665. URL: <https://doi.org/10.3115/v1/p14-1062>. doi:10.3115/v1/p14-1062.
- [28] I. Chaturvedi, Y. Ong, I. Tsang, R. Welsch, E. Cambria, Learning word dependencies in text by means of a deep recurrent belief network, *Knowledge-Based Systems* 108 (2016). doi:10.1016/j.knosys.2016.07.019.
- [29] F. Antici, L. Bolognini, M. A. Inajetovic, B. Ivasiuk, A. Galassi, F. Ruggeri, Subjectivita: An italian corpus for subjectivity detection in newspapers, in: CLEF, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 40–52.
- [30] C. Banea, R. Mihalcea, J. Wiebe, Multilingual subjectivity: Are more languages better?, in: Proceedings of the 23rd International Conference on Computational Linguistics (Coling

- 2010), Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 28–36. URL: <https://aclanthology.org/C10-1004>.
- [31] F. Antici, A. Galassi, F. Ruggeri, K. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barrón-Cedeño, A corpus for sentence-level subjectivity detection on english news articles, 2023. [arXiv:2305.18034](https://arxiv.org/abs/2305.18034).
- [32] C.-L. Yeh, B. Loni, M. Hendriks, H. Reinhardt, A. Schuth, Dpgmedia2019: A dutch news dataset for partisanship detection, 2019. [arXiv:arXiv:1908.02322](https://arxiv.org/abs/1908.02322).
- [33] G. K. Shahi, J. M. Struß, T. Mandl, J. Köhler, M. Wiegand, M. Siegel, CT-FAN: A Multilingual dataset for Fake News Detection, 2022. URL: <https://doi.org/10.5281/zenodo.6555293>. doi:10.5281/zenodo.6555293.
- [34] 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: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.
- [35] S. Tran, P. Rodrigues, B. Strauss, E. Williams, Accenture at CheckThat! 2023: Impacts of back-translation on subjectivity detection, in: [47], 2023.
- [36] I. B. Schlicht, L. Khellaf, D. Altiok, Dwreco at CheckThat! 2023: Enhancing subjectivity detection through style-based data sampling, in: [47], 2023.
- [37] H. T. Sadouk, F. Sebbak, H. E. Zekiri, Es-vrai at CheckThat! 2023: Enhancing model performance for subjectivity detection through multilingual data aggregation, in: [47], 2023.
- [38] R. A. Frick, Fraunhofer sit at CheckThat! 2023: Can llms be used for data augmentation & few-shot classification? detecting subjectivity in text using chatgpt, in: [47], 2023.
- [39] G. Pachov, D. Dimitrov, I. Koychev, P. Nakov, Gpachov at CheckThat! 2023: A diverse multi-approach ensemble for subjectivity detection in news articles, in: [47], 2023.
- [40] K. Dey, P. Tarannum, M. A. Hasan, S. R. H. Noori, Nn at CheckThat! 2023: Subjectivity in news articles classification with transformer based models, in: [47], 2023.
- [41] F. Leistra, T. Caselli, Thesis titan at CheckThat! 2023: Language-specific fine-tuning of mdebertav3 for subjectivity detection, in: [47], 2023.
- [42] M. Deniz Türkmen, G. Coşgun, M. Kutlu, TOBB ETU at CheckThat! 2023: Utilizing chatgpt to detect subjective statements and political bias, in: [47], 2023.
- [43] E. Shushkevich, J. Cardiff, Tudublin at CheckThat! 2023: Chatgpt for data augmentation, in: [47], 2023.
- [44] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, *BMC genomics* 21 (2020) 1–13.
- [45] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: <https://aclanthology.org/D19-1410>. doi:10.18653/v1/D19-1410.
- [46] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. [arXiv:2111.09543](https://arxiv.org/abs/2111.09543).
- [47] 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.

## A. Approaches Summary per Participant

**Accenture** [35] employed several language-specific pre-trained language models which were fine-tuned to the downstream task. They also applied data augmentation through back-translation to dim the class imbalance in the training datasets.

**Awakened** implemented an ensemble of classifiers including a BiLSTM and BART. The embedding produced by each models is concatenated and fed to a final classification layer.

**DWReCo** [36] experimented with propaganda style-based data augmentation via GPT-3 to address class imbalance. The styles are identified from the journalistic checklist to identify subjective news.

**ES-VRAI** [37] used M-BERT and made use of oversampling techniques to address class imbalance.

**Fraunhofer SIT** [38] employed GPT-3 for few-shot classification.

**Gpachov** [39] applied an ensemble of three distinct models: XLM-RoBERTa, Sentence BERT (S-BERT), and SetFit.

**KUCST** used a Gradient Boosting classifier with BERT-based encoding and a subset of carefully selected features as inputs.

**NN** [40] used the multilingual XLM-RoBERTa, fine-tuned on 100 languages from 2.5TB of filtered CommonCrawl data.

**tarrekko** did not provide additional information.

**Thesis Titan** [41] developed multiple fine-tuned models using mDeBERTaV3-base [46] starting from a newly developed multilingual dataset. While keeping the training data fixed, they have used language-specific validation sets to optimize the models for each language, as well as to identify optimal language-specific hyperparameters.

**TOBB ETU** [42] employed ChatGPT to classify the texts. They explored zero-shot and few-shot classification. In the former, they show a few mistakes of zero-shot classification method on the training set.

**TUDublin** [43] experimented with M-BERT and made use of ChatGPT to perform data augmentation of the available training datasets.