Overview of the CLEF-2024 CheckThat! Lab Task 2 on Subjectivity in News Articles

Notebook for the CheckThat! Lab at CLEF 2024

Julia Maria Struß[1,*], Federico Ruggeri[2,*], Alberto Barrón-Cedeño[2], Firoj Alam[3], Dimitar Dimitrov[4], Andrea Galassi[2], Georgi Pachov[4], Ivan Koychev[4], Preslav Nakov[5], Melanie Siegel[6], Michael Wiegand[7], Maram Hasanain[3], Reem Suwaileh[8] and Wajdi Zaghouani[9]

[1]*University of Applied Sciences Potsdam, Germany*
[2]*University of Bologna, Italy*
[3]*Qatar Computing Research Institute, HBKU, Qatar*
[4]*Sofia University, Bulgaria*
[5]*Mohamed bin Zayed University of Artificial Intelligence, UAE*
[6]*Darmstadt University of Applied Sciences, Germany*
[7]*University of Vienna, Austria*
[8]*Hamad Bin Khalifa University, Qatar*
[9]*Northwestern University in Qatar, Education City, Doha, Qatar*

Abstract

We present an overview of Task 2 of the seventh edition of the CheckThat! lab at the 2024 iteration of the Conference and Labs of the Evaluation Forum (CLEF). The task focuses on subjectivity detection in news articles and was offered in five languages: Arabic, Bulgarian, English, German, and Italian, as well as in a multilingual setting. The datasets for each language were carefully curated and annotated, comprising over 10,000 sentences from news articles. The task challenged participants to develop systems capable of distinguishing between subjective statements (reflecting personal opinions or biases) and objective ones (presenting factual information) at the sentence level. A total of 15 teams participated in the task, submitting 36 valid runs across all language tracks. The participants used a variety of approaches, with transformer-based models being the most popular choice. Strategies included fine-tuning monolingual and multilingual models, and leveraging English models with automatic translation for the non-English datasets. Some teams also explored ensembles, feature engineering, and innovative techniques such as few-shot learning and in-context learning with large language models. The evaluation was based on macro-averaged F1 score. The results varied across languages, with the best performance achieved for Italian and German, followed by English. The Arabic track proved particularly challenging, with no team surpassing an F1 score of 0.50. This task contributes to the broader goal of enhancing the reliability of automated content analysis in the context of misinformation detection and fact-checking. The paper provides detailed insights into the datasets, participant approaches, and results, offering a benchmark for the current state of subjectivity detection across multiple languages.

Keywords

subjectivity classification, fact-checking, misinformation detection

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
julia.struss@fh-potsdam.de (J. M. Struß); federico.ruggeri6@unibo.it (F. Ruggeri); a.barron@unibo.it (A. Barrón-Cedeño); fialam@hbku.edu.qa (F. Alam); ilijanovd@fmi.uni-sofia.bg (D. Dimitrov); a.galassi@unibo.it (A. Galassi); gpachov@uni-sofia.bg (G. Pachov); koychev@fmi.uni-sofia.bg (I. Koychev); preslav.nakov@mbzuai.ac.ae (P. Nakov); melanie.siegel@h-da.de (M. Siegel); michael.wiegand@univie.ac.at (M. Wiegand); mhasanain@hbku.edu.qa (M. Hasanain); rsuwaileh@hbku.edu.qa (R. Suwaileh); wajdi.zaghouani@northwestern.edu (W. Zaghouani)
0000-0001-9133-4978 (J. M. Struß); 0000-0002-1697-8586 (F. Ruggeri); 0000-0003-4719-3420 (A. Barrón-Cedeño); 0000-0001-7172-1997 (F. Alam); 0000-0003-1308-180X (D. Dimitrov); 0000-0001-9711-7042 (A. Galassi); 0009-0008-1546-3805 (G. Pachov); 0000-0002-5064-5750 (M. Siegel); 0000-0002-5403-1078 (M. Wiegand); 0000-0002-7466-178X (M. Hasanain); 0000-0002-7466-178X (R. Suwaileh); 0000-0003-1521-5568 (W. Zaghouani)

1. Introduction

The CheckThat! lab is organized for the 7th time within CLEF 2024. This paper presents an overview of Task 2 on detecting subjectivity in news articles; the task is organized for a second time. The ability to discern between subjective and objective content has become increasingly important in the domain of natural language processing (NLP), especially in the context of news articles during the last decades. The distinction between subjective statements, which reflect personal opinions, beliefs, or biases, and objective statements, which present factual information, is fundamental for various applications such as sentiment analysis, opinion mining, and fact-checking. Task 2 aims to advance research in this area by providing a platform for the development and the evaluation of systems capable of subjectivity detection across multiple languages. Specifically, Task 2 challenges participants to distinguish whether a given sentence from a news article expresses the subjective view of the author or presents an objective account of the topic. This binary classification task is crucial for improving the reliability and the transparency of information processing systems, particularly in an era where misinformation and biased reporting are prevalent.

The task is designed to be language-agnostic, with datasets provided in five languages: Arabic, Bulgarian, English, German, and Italian. Additionally, a multilingual scenario is introduced in order to evaluate the robustness of the participating systems across diverse language contexts. The datasets have been carefully annotated to ensure a high quality and consistency, following prescriptive guidelines for language-agnostic subjectivity detection. We evaluated the participating systems based on macro-averaged F1 score, which balances precision and recall across the subjective and the objective classes, ensuring a fair assessment of system performance. The task's comprehensive evaluation framework, coupled with the diverse linguistic datasets, provides a rigorous benchmark for advancing the state-of-the-art in subjectivity detection.

In the rest of the paper, we offer an overview of Task 2, detailing the datasets, the evaluation measures, and the submission guidelines. We also present the results and the methodologies of the participating systems, highlighting the progress and the challenges when developing robust subjectivity detection models. By fostering collaboration and innovation in this critical area, Task 2 contributes to the broader goal of enhancing the reliability of automated content analysis in the digital age.

2. Related Work

Subjectivity detection has been approached in different contexts and domains. Early work on the topic was mainly conducted in the context of sentiment analysis [1, 2] focusing on English, but later also exploring multilingual approaches [3, 4]. Subjectivity detection has also been approached as part of bias detection [5, 6], claim extraction [7], or in the context of fact-checking [8, 9]. Our task is motivated by fact-checking.

Naturally, the definition of subjectivity varies with the task and the direction it is approached from. Previous work has developed corpora by relying on different assumptions to detect subjectivity, such as domain-specific heuristics based on keyword spotting [1, 10, 11], statistical methods [12], and based on annotation guidelines [13, 14, 15]. According to Chaturvedi et al. [16], the first solutions are known as *syntactic* approaches, while the latter two sets are *semantic* approaches. Syntactic methods suffer from limited applicability as they are tied to domain- and language-specific knowledge. For this reason, semantic approaches, especially based on annotation guidelines, have recently been preferred. Nevertheless, interpretation biases, edge cases, and annotation ambiguity are notable challenges when adopting annotation guidelines, which hinder the development of high-quality datasets [14, 17]. These risks are limited by carrying out a data annotation methodology based on the prescriptive paradigm [18] and framing subjectivity detection for the task of fact verification [19]. This has also been formulated and adopted in the previous iteration of Task 2 in the sixth edition of the CheckThat! lab [20], and we follow it in this iteration as well. Compared to the previous iteration of the task, we dropped Dutch and Turkish due to the lack of support, and we added Bulgarian as a new language.

Table 1
Statistics about the data for all five languages.

	Training			Development			Development-Test			Test		
Language	**Total**	**OBJ**	**SUBJ**	**Total**	**OBJ**	**SUBJ**	**Total**	**OBJ**	**SUBJ**	**Total**	**OBJ**	**SUBJ**
Arabic	1,185	905	280	297	227	70	445	363	82	748	425	323
Bulgarian	729	406	323	106	59	47	208	116	92	250	143	107
English	830	532	298	219	106	113	243	116	127	484	362	122
German	800	492	308	200	123	77	291	194	97	337	226	111
Italian	1,613	1,231	382	227	167	60	440	323	117	513	377	136

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

Language	**Sentence**	**Class**
Arabic	وجدت بوحيد 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

Our data annotations cover different granularity levels, including sentence [7, 21], segment [22], and document [23]. While the majority of the available corpora are in English, there are several attempts to extend subjectivity detection to other languages, such as Arabic [24, 25], German [24], French [22, 24], Italian [23], Romanian [26, 24], and Spanish [24]. However, they mostly rely on machine translation and ontologies to scale to multiple languages, which introduces noise in the annotation.

The task of subjectivity detection was also part of the 2023 edition of the CLEF CheckThat! lab [27]. The task was offered in 6 languages [20]: Arabic, Dutch, English, German, Italian, and Turkish.

3. Datasets

The task offered datasets in five different languages with a total of more than 10k sentences manually annotated following the guidelines in [19]. Table 1 presents details on the dataset statistics. Some sample instances for each language are given in Table 2.

3.1. Arabic

The dataset consists of sentences from news articles, including sources such as AraFact [28]. The complete data collection and annotation process involved several phases. In the *article selection* phase, we selected 1,159 news articles from AraFact [28]. Additionally, we manually searched for opinionated articles from various Arabic news outlets, eventually selecting 221 articles. These articles were parsed and segmented, resulting in 15,947 sentences. During the *sentence selection* phase, we selected 4,524 sentences after various filtering steps, where different existing classifiers were used to classify the sentences into subjective and objective, and the sentences were selected based on their decisions.

Finally, in the *sentence annotation* phase, we annotated the sentences using in-house and Amazon Mechanical Turk (MTurk) annotators. We adhered to standard qualification tests and ensured majority agreement among 3 to 5 annotators. A label was selected for each sentence where at least two annotators agreed. The inter-annotator agreement (pair-wise Cohen's kappa) was $\kappa = 0.538$. More details about the data collection and annotation process can be found in [29].

For the lab, we used all data from CheckThat! 2023 [20], specifically using the training and the development datasets as the training and development sets, respectively, and the test data as the development-test set. Based on the data collection and annotation procedure discussed above and in [29], we developed a new dataset for testing.

3.2. Bulgarian

The Bulgarian corpus is based on a fake news dataset [30], provided by the Bulgarian Association of PR Agencies[1]. It was annotated by Bulgarian students of journalism and further cleaned and deduplicated by the authors. The cleaning process consisted of removing Unicode symbols and URL encoding schemes. Afterwards, we segmented the articles into sentences, and we filtered out all sentences shorter than 5 tokens. We then selected a total of 1,293 random sentences for annotation. The annotation was conducted by three native Bulgarian speakers: two annotators and a consolidator, according to the guidelines of [19]. The process consisted of multiple annotation batches, each of size 200. After each batch was annotated, we accepted the examples with matching labels, and we sent all conflicting annotations for analysis and discussion between the annotators. In cases of disagreement after these initial discussions, the consolidator was involved to help reach a decision. We measured *Krippendorff's Alpha* for the inter-annotator agreement for every batch, and it ranged from 0.6 (worse batches, lots of discussions until labels are agreed upon for every sentence) to 0.85 (good batches, very few disputes).

3.3. English

The English dataset is partially based on NewsSD-ENG [19], a corpus of 1,049 sentences labeled by seven annotators following guidelines for subjectivity detection tailored to an information retrieval setting [31]. We used 830 examples for training and 219 for development. The Krippendorff's alpha [32] measuring the inter-annotator agreement on the whole corpus is 0.83. The test partition of the CheckThat! lab 2023 Task 2 edition [20] is used as a dev-test set. We further developed a new test set following the same data collection methodology for NewsSD-ENG, which comprises several stages to ensure high quality. First, we sampled 15 news articles on controversial topics, with a total of 490 sentences. Then, eight annotators labeled the sentences as subjective or objective. Each sentence was annotated by at least two annotators. Then, the annotators engaged in a discussion phase to resolve any disagreements. The unresolved conflicts were addressed by a third annotator who decided on the final label. The inter-annotator agreement on the new test set measured with Krippendorff's alpha was 0.86.

3.4. German

The German dataset was assembled by randomly selecting sentences from the CT 2022 FAN-Corpus [33] consisting of news articles originally annotated according to the factuality of their main claim. The 800 manually annotated sentences for training and the 200 instances for development are from the 2023 edition of the task [20]. As an additional development set (dev-test), the 2023 test data is used. For the test set, 360 new sentences were randomly sampled from the CT!2022FAN-Corpus (instances not included in the other partitions). They were annotated following the guidelines outlined in [19]. 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 three native speakers, all co-authors of this paper. The annotators achieved a substantial

[1] http://www.bapra.bg

agreement [34], with Fleiss' $kappa = 0.696$ ($p < 0.0001$, $z = 22.1$). The final label was obtained by majority voting.

3.5. Italian

The dataset is based on the SubjectivITA corpus [23], re-annotated using the guidelines introduced in Antici et al. [19]. The corpus contains 1,841 sentences split into 1,613 for training and 227 for development. The inter-annotator agreement measured with Fleiss' $kappa$ [34] is 0.65. A novel test set is defined following the same methodology used for the English dataset. In particular, we collected 513 sentences from 14 news articles targeting controversial topics. Eight annotators labeled each sentence as subjective or objective. The inter-annotator agreement on the new test set measured with Krippendorff's alpha [32] is 0.79.

4. Overview of the Systems and Results

Fifteen teams participated in this task, submitting 36 valid runs. Seven teams submitted valid runs for more than one language, with three teams participating in all six language settings, including the multilingual one. All teams participated in the English subtask.

Table 3 shows the results achieved by the individual teams for each language. At least two teams improved over the baseline for all languages, except for Bulgarian. Notable performance was achieved for Italian and German, followed by English, with scores above 0.70. For Arabic, none of the teams achieved a macro-F1 score above 0.50. The team with the most stable results across all languages was Nullpointer [35]: with the exception of the English subtask, they always ranked among the top-3 teams. All teams used neural networks, with transformer-based models being the most frequent choice. Some teams used language-specific monolingual transformer models, others chose multilingual models and

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			**Bulgarian**			**English**		
1	IAI Group	0.495	1	(baseline)	0.753	1	HYBRINFOX	0.744
2	Nullpointer †	0.491	2	Nullpointer	0.717	2	Tonirodriguez	0.737
3	(baseline)	0.485	3	HYBRINFOX	0.715	3	SSN-NLP	0.712
4	SemanticCuetSync	0.480	4	IAI Group	0.582	4	Checker Hacker	0.708
5	Tonirodriguez	0.465	5	JUNLP	0.364	5	JK_PCIC_UNAM	0.708
6	HYBRINFOX	0.455	**Italian**			6	SINAI	0.703
7	JUNLP	0.362	1	JK_PCIC_UNAM	0.792	7	FactFinders	0.695
German			2	HYBRINFOX	0.784	8	Vigilantes	0.695
1	Nullpointer	0.791	3	Nullpointer	0.743	8	Eevvgg	0.695
2	IAI Group	0.730	4	(baseline)	0.650	9	Nullpointer	0.689
3	(baseline)	0.699	5	IAI Group	0.586	10	Indigo	0.639
4	HYBRINFOX	0.697				11	(baseline)	0.635
Multilingual						12	SemanticCuetSync	0.627
–	Nullpointer *	0.712				13	JUNLP	0.560
1	HYBRINFOX	0.685				14	CLaC	0.450
2	(baseline)	0.670				15	IAI Group	0.449
3	IAI Group	0.629						

† Team involved in the preparation of the data.
* Submitted after the official deadline.

some teams used English models in combination with automatic translation.

Table 4
Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking.

Team	Language						Model												Misc			
	Multilingual	Arabic	Bulgarian	English	German	Italian	BERT	RoBERTa	DistilBERT	Gemini	mBERT	mDeBERTa	Sentence-BERT	SetFit	Mistral-7B-Instruct	XLM RoBERTa	DeBERTa	BART	Llama	Sentiment-Analysis-BERT	Data Augmentation	Translating data	Multi-lingual Training	Feature Selection
Checker Hacker [36]				4				☑													☑			
CLaC [37]				14						☑											☑			
Eevvgg [38]				8			☑																	☑
FactFinders				7											☑									
HYBRINFOX [39]	1	6	3	1	4	2		☑	☑		☑											☑		☑
IAI Group [40]	3	1	4	15	2	5		☑								☑								
Indigo [41]				10									☑	☑										
JK_PCIC_UNAM [42]				5		1	☑																	☑
JUNLP		7	5	13			☑				☑													
Nullpointer [35]	-	2	2	1	9	3														☑		☑		
SemanticCuetSync [43]		4		12								☑							☑					
SINAI				6				☑																
SSN-NLP [44]				3				☑																☑
Tonirodriguez [45]		5		2								☑				☑	☑	☑					☑	
Vigilantes				8			☑																	

- The run was submitted after the official deadline, therefore not part of the official ranking.

4.1. Baselines

For all languages, our baseline was a multilingual SentenceBERT [46] model with a logistic regression classifier on top of it. In particular, we consider 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.

4.2. Results per Language

Multilingual Three teams submitted runs to the multilingual subtask with two beating the baseline. Team **HYBRINFOX** [39] ranked first with their hybrid approach combining an ensemble of RoBERTa [47] and SentenceBERT [46] with the rule-based-expert-system VAGO [48]. The late submission by team **Nullpointer** [35] achieved a better score by almost 3%-points using a system tailored to English and translating all non-English instances using Google Translate.

Arabic The Arabic subtask attracted six teams with only two surpassing the baseline of 0.485. The **IAI Group** [40] achieved a macro F1 score of 0.495, with their XLM-RoBERTa approach, which was the top-ranked system for Arabic. The second ranked team **Nullpointer** [35] fine-tuned a sentiment-based transformer model using data augmentation and used a custom classifier that assigned a higher weight to the objective class.

Bulgarian Four teams submitted runs to the Bulgarian subtask with none surpassing the baseline scoring 0.753. The two systems at the second and third rank – team **Nullpointer** [35] and team **HYBRINFOX** [39] – achieved similar scores of 0.717 and 0.715, respectively.

English The English subtask attracted the highest number of participants with 15 teams submitting valid runs, 10 of which surpassed the baseline. Team **HYBRINFOX** was ranked first (0.744) followed by **Tonirodriguez** [45] and **SSN-NLP** [44]. The teams at positions four to nine achieved similar results in the range [0.69, 0.71]. Team **Indigo** [41] was ranked 10th, passing the baseline by a small margin.

German Three teams submitted runs for the German subtask with only two surpassing the baseline. Team **Nullpointer** achieved the first place with a score of 0.791 surpassing the **IAI Group** ranking second and the baseline by 6 and 10%-points, respectively.

Italian Four teams submitted valid runs to the Italian subtask with team **JK_PCIC_UNAM** ranking first obtaining a score of 0.792, closely followed by team **HYBRINFOX**. The third-positioned team **Nullpointer** follows with a difference of 4%-points still surpassing the baseline by a margin.

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 **CheckerHacker** [49] used an ensemble of two transformer-based models: BERT-base-uncased and XLM-RoBERTa-base. For the ensemble, they used average probabilities. In addition to the ensemble, they implemented data augmentation.

Team **CLaC** [37] approached the task by leveraging an LLM (Google's Gemini[2]) for classification. They modeled the task as a multi-voter scenario where the LLM was used to create two semantically similar sentences for each sentence in the test set. Then, the same LLM predicted the subjectivity of each sentence using a single prompt. Finally, majority voting over the three labels was used to decide the final subjectivity label. Additionally, the prompt was contextualized by providing 600 randomly selected samples from the training set.

Team **Eevvgg** [38] compared a number of feature-based supervised models, namely Naïve Bayes, SVMs, logistic regression, decision trees and random forest with transformer-based models like BERT for the English subtask, incorporating additional syntactic features deemed as stance markers.

Team **FactFinders** fine-tuned the Mistral-7B-Instruct-v0.2 model on the data provided for English. (No further details are available, since this submission was not accompanied by a paper.)

Team **HYBRINFOX** [39] evaluated an ensemble combining a RoBERTa-based encoder, a Sentence-BERT encoder, and lexical features. The RoBERTa and SentenceBERT embeddings were concatenated with subjectivity scores extracted from a rule-based expert system based on the VAGO lexical database [48]. These scores covered aspects such as vagueness, subjectivity, detail, and objectivity. The enriched embeddings were then fed into the downstream classifier. Regarding training, only RoBERTa was fine-tuned, while the SentenceBERT model weights were frozen. The authors used machine translation with DeepL[3] for all non-English sub-tasks.

Team **IAI Group** [40] experimented with the multilingual XLM-RoBERTa model for all subtasks. They fine-tuned the model for each specific language.

Team **Indigo** [41] proposed a classifier based on SetFit [50], in two steps. In the first stage, the training data was used to fine-tune a transformer using contrastive learning. The resulting transformer generated sentence embeddings, which are fed into a Differentiable Linear Neural Layer. This model is known to require only a few training data samples, making it suitable for few-shot learning settings. In their evaluation, the authors show the effectiveness of their model compared to other conventional fine-tuning approaches (e.g., BERT or RoBERTa), especially when trained for only a few epochs.

[2] https://gemini.google.com
[3] https://www.deepl.com/en/translator

Team **JK_PCIC_UNAM** [42] used a BERT-based classifier for English and Italian. They fine-tuned two distinct BERT classifiers, each tailored to a specific language. In each setting, they enriched BERT-based embeddings with linguistic features, including the number of quotations, the percentage of nouns, verbs, adjectives, adverbs, and sentiment probabilities from input texts provided by pysentimiento [51].

Team **JUNLP** fine-tuned pre-trained language models (PLMs) for Arabic, Bulgarian, and English. They used BERT for English and multilingual BERT (mBERT) for Arabic and Bulgarian. The original training data was used without tuning the values of the hyperparameters.

Team **Nullpointer** [35] fine-tuned a BERT-based classifier for Arabic, Bulgarian, English, German, and Italian. The BERT model, initially pre-trained for sentiment analysis, was fine-tuned for each specific language, where the sentiment labels output by the model were mapped to subjectivity labels. They handled class imbalance, and translated all non-English data to English.

Team **SemanticCuetSync** [43] implemented and evaluated several models based on transformers and deep learning for English and Arabic. They experimented with linear regression, SVM, multinomial naïve Bayes, k-nearest neighbors, random forest, CNN, CNN+LSTM, CNN+BiLSTM, and two LLMs that were fine-tuned. They selected a LLaMA-3-8b model for their submission, as it yielded the best results for both languages.

Team **SINAI** fine-tuned RoBERTa-base for English on the data provided for the task.

Team **SSN-NLP** [44] explored keyword- to embedding-based representations, using kNN and random forests as well as neural networks, including transformers, all for English. They used a custom pre-processing pipeline, which included tokenization and part-of-speech (POS) tagging to produce additional features. Their best-performing model fine-tuned a RoBERTa-based classifier enriched with POS features concerning subjectivity and objectivity.

Team **ToniRodriguez** [45] fine-tuned two multilingual transformer-based classifiers and XLM-RoBERTa for English, German, and Italian. Eventually, the mDeBERTa-v3 model was chosen as the best-performing one. Finally, they applied zero-shot cross-lingual transfer to Arabic and Bulgarian.

Team **Vigilantes** fine-tuned BERT for English on the data provided for the task.

5. Conclusion and Future Work

We presented an overview of Task 2 from the CheckThat! lab at CLEF 2024. The task concerned the detection of subjective sentences in controversial news articles. The task was offered in five different languages and also in a multilingual setting.

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 with instruction tuning to detect subjectivity or for data augmentation. The most successful solutions coupled transformer-based classifiers with domain knowledge in the form of feature extraction, machine translation, and data augmentation, outperforming baselines by a large margin in most tasks, with the exception of Arabic and Bulgarian. The best macro F_1 scores ranged between 0.49 and 0.79, which shows that there is a lot of room for improvement, especially for certain languages.

In future work, we plan to increase the number of languages covered and focus on multi- and cross-lingual settings.

Acknowledgments

The work related to the German data has partially been funded by the BMBF (German Federal Ministry of Education and Research) under grant no. 01FP20031J. The responsibility for the contents of this publication lies with the authors.

The work of M. Hasanain, R. Suwaileh, F. Alam and W. Zaghouani is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund, which is a part of Qatar Research Development and Innovation Council (QRDI).

The work of D. Dimitrov, G. Pachov, and I. Koychev is partially financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

The work of A. Galassi is funded by European Commission's NextGenerationEU programme, PNRR-M4C2-Investimento 1.3, PE00000013 "FAIR", Spoke 8.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[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] 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. Zaghouani, 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.
[28] 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.
[29] R. Suwaileh, M. Hasanain, F. Hubail, W. Zaghouani, F. Alam, ThatiAR: Subjectivity detection in arabic news sentences, arXiv: 2406.05559 (2024).
[30] G. Karadzhov, P. Gencheva, P. Nakov, I. Koychev, We built a fake news / click bait filter: What happened next will blow your mind!, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 334–343. doi:10.26615/978-954-452-049-6_045.
[31] 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.
[32] K. Krippendorff, Computing krippendorff's alpha-reliability, 2011. URL: https://repository.upenn.edu/handle/20.500.14332/2089.
[33] 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.
[34] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[35] M. R. Biswas, A. Tasneem Abir, W. Zaghouani, Nullpointer at CheckThat! 2024: Identifying subjectivity from multilingual text sequence, in: [52], 2024.
[36] S. D. Zehra, K. Chandani, M. Khubaib, A. A. Aun Muhammed, F. Alvi, A. Samad, Checker Hacker at CheckThat! 2024: Detecting check-worthy claims and analyzing subjectivity with transformers, in: [52], 2024.
[37] S. Gruman, L. Kosseim, CLaC at CheckThat! 2024: A zero-shot model for check-worthiness and subjectivity classification, in: [52], 2024.
[38] E. Gajewska, Eevvgg at CheckThat! 2024: Evaluative terms, pronouns and modal verbs as markers of subjectivity in text, in: [52], 2024.
[39] M. Casanova, J. Chanson, B. Icard, G. Faye, G. Gadek, G. Gravier, P. Égré, HYBRINFOX at CheckThat! 2024 - task 2: Enriching BERT models with the expert system VAGO for subjectivity detection, in: [52], 2024.
[40] P. R. Aarnes, V. Setty, P. Galuščáková, IAI group at CheckThat! 2024: Transformer models and

data augmentation for checkworthy claim detection, in: [52], 2024.
[41] S. Sar, D. Roy, Indigo at CheckThat! 2024: Using Setfit: A resource efficient technique for subjectivity detection in news article., in: [52], 2024.
[42] K. Salas-Jimenez, I. Díaz, H. Gómez-Adorno, JK_PCIC_UNAM at CheckThat! 2024: Analysis of subjectivity in news sentences using transformers based models, in: [52], 2024.
[43] A. I. Paran, M. S. Hossain, S. H. Shohan, J. Hossain, S. Ahsan, M. M. Hoque, SemanticCuetSync at CheckThat! 2024: Finding subjectivity in news article using Llama, in: [52], 2024.
[44] P. Premnath, P. Vaithiya Subramani, B. B, N. R. Salim, SSN-NLP at CheckThat! 2024: From classic algorithms to transformers: A study on detecting subjectivity, in: [52], 2024.
[45] A. Rodríguez de la Torre, E. Golobardes Ribé, J. Suau Martínez, Tonirodriguez at CheckThat!2024: Is it possible to use zero-shot cross-lingual for subjectivity detection in low-resources languages?, in: [52], 2024.
[46] 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.
[47] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.
[48] B. Icard, V. Claveau, G. Atemezing, P. Égré, Measuring vagueness and subjectivity in texts: From symbolic to neural VAGO, in: Proceedings of the IEEE International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, 2023, pp. 395–401. doi:10.1109/WI-IAT59888.2023.00065.
[49] K. Chandani, D. E. Z. Syeda, Checker Hacker at CheckThat! 2024: Ensemble models for checkworthy tweet identification, in: [52], 2024.
[50] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, 2022. URL: https://arxiv.org/abs/2209.11055. arXiv:2209.11055.
[51] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, pysentimiento: A python toolkit for opinion mining and social nlp tasks, 2023. arXiv:2106.09462.
[52] 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.