XplaiNLP at CheckThat! 2025: Multilingual Subjectivity Detection with Finetuned Transformers and Prompt-Based Inference with Large Language Models

Notebook for the CheckThat! Lab at CLEF 2025

Ariana Sahitaj[1,2,*], Jiaao Li[1,2], Pia Wenzel Neves[1,2], Fedor Splitt[1,2], Premtim Sahitaj[1,2], Charlott Jakob[1,2], Veronika Solopova[1,2] and Vera Schmitt[1,2]

[1]*Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany*
[2]*German Research Center for Artificial Intelligence (DFKI), Berlin, Germany*

Abstract

This notebook reports the XplaiNLP submission to the CheckThat! 2025 shared task [1] on multilingual subjectivity detection. We evaluate two approaches: (1) supervised fine-tuning of transformer encoders, EuroBERT, XLM-RoBERTa, and German-BERT, on monolingual and machine-translated training data; and (2) zero-shot prompting using two LLMs: `o3-mini` for Annotation (rule-based labelling) and `gpt-4.1-mini` for DoubleDown (contrastive rewriting) and Perspective (comparative reasoning). The Annotation Approach achieves 1st place in the Italian monolingual subtask with an F_1 score of 0.8104, outperforming the baseline of 0.6941. In the Romanian zero-shot setting, the fine-tuned XLM-RoBERTa model obtains an F_1 score of 0.7917, ranking 3rd and exceeding the baseline of 0.6461. The same model also performs reliably in the multilingual task and improves over the baseline in Greek. For German, a German-BERT model fine-tuned on translated training data from typologically related languages yields competitive performance over the baseline. In contrast, performance in the Ukrainian and Polish zero-shot settings falls slightly below the respective baselines, reflecting the challenge of generalization in low-resource cross-lingual scenarios.

Keywords

Subjectivity Detection, Multilingual NLP, Zero-Shot Learning, Prompt-Based Inference

1. Introduction

Understanding whether a sentence expresses a personal opinion or presents information in a neutral and therefore objective way is important in many natural language processing tasks [2]. This distinction is particularly relevant in the context of news reporting, where objectivity is traditionally considered a core principle. Yet, subjective or evaluative language is often embedded in news texts through stylistic choices and subtle dialogic elements that influence how readers interpret information. [3] This effect is especially strong when opinionated language is presented in the style of factual reporting, causing evaluative statements to appear as objective observations [4]. A precise distinction between subjective and objective language is important for tasks such as sentiment analysis [5], stance detection [6], automated fact-checking [7, 8], propaganda detection [9], argument mining [10], and bias identification [11]. These applications rely on the ability to detect whether a statement reflects personal opinion, emotional language, or evaluative framing, or whether it is intended to convey factual content. Subjective sentences commonly include emotional terms, value judgments, or rhetorical elements such as irony or exaggeration [12]. However, even for human readers, it is not always simple to decide whether a sentence is subjective or not. Interpretations often depend on context and background knowledge,

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
ariana.sahitaj@campus.tu-berlin.de (A. Sahitaj); jiaao.li@campus.tu-berlin.de (J. Li); p.wenzel.2@campus.tu-berlin.de (P. Wenzel Neves); splitt@campus.tu-berlin.de (F. Splitt); sahitaj@tu-berlin.de (P. Sahitaj); c.jakob@tu-berlin.de (C. Jakob); veronika.solopova@tu-berlin.de (V. Solopova); vera.schmitt@tu-berlin.de (V. Schmitt)
0009-0002-0096-9383 (A. Sahitaj); 0009-0001-8340-0215 (P. Wenzel Neves); 0000-0003-3908-5681 (P. Sahitaj); 0009-0002-6262-9018 (C. Jakob); 0000-0003-0183-9433 (V. Solopova); 0000-0002-9735-6956 (V. Schmitt)

making critical thinking essential for distinguishing between evaluative language and factual reporting [13, 14]. This issue becomes more apparent in multilingual settings, as different languages signal subjectivity in diverse ways, through verb forms, word order, lexical choices, or stylistic conventions [15]. At the same time, many languages lack annotated resources for subjectivity detection, which poses an additional challenge for training reliable models [16].

This notebook describes our submission to the CheckThat! Lab at CLEF 2025 [17], which focused on sentence-level subjectivity detection across multiple languages. Our approach investigates two complementary approaches: (1) supervised fine-tuning of multilingual and monolingual transformer-based encoders on annotated datasets, and (2) zero-shot prompting with LLMs using natural language inference guided by explicit instructions. While the first approach relies on parameter-efficient adaptation of pretrained models, the second uses the contextual reasoning capabilities of LLMs to classify subjectivity without additional training.

The paper is structured as follows: In Section 2 we review related work on subjectivity detection. Section 3 introduces the dataset used in the shared task. Section 4 details our modeling approaches, including fine-tuned transformer models and zero-shot prompting strategies. Section 5 presents our evaluation results before we finally summarize our findings in Section 6 and outline directions for future work.

2. Related Work

The aim of subjectivity detection is to distinguish language that conveys private states, such as opinions, from language that presents information in a way that seem factual or in a neutral manner, regardless of whether the information is actually true [18]. Yu and Hatzivassiloglou proposed one of the earliest computational models for sentence-level subjectivity detection, using Bayesian classification. Riloff and Wiebe and Riloff et al. explored how subjectivity detection can improve information extraction by reducing false positives, especially in metaphorical or emotional contexts. These studies marked a shift toward integrating subjectivity classification into practical NLP pipelines. In 2006 Esuli et al. introduced SentiWordNet, a lexical resource assigning polarity and objectivity scores to WordNet synsets. While primarily intended for polarity detection, it also provides objectivity measures, implicitly supporting subjectivity detection tasks. Banea et al. addressed the scarcity of resources in non-English languages by developing a bootstrapping approach to build subjectivity lexicons using only seed lists and a basic corpus. This method made subjectivity analysis feasible for low-resource settings. Later, Chaturvedi et al. provided a comprehensive survey of both rule-based and automatic models. They emphasized that subjectivity detection is a important preprocessing step for sentiment analysis, as polarity classifiers may otherwise incorrectly label neutral statements as emotional content. Antici et al. introduced NewsSD-ENG, a sentence-level dataset with human-annotation subjectivity labels. Their experiments showed that multilingual transformer models (M-BERT and M-SBERT) clearly outperformed traditional classifiers (SVM and LR), with multilingual training improving performance and enabling robust cross-lingual subjectivity detection. Savinova and Del Prado reframed subjectivity detection as a regression task and fine-tuned a RoBERTa model to predict degrees of subjectivity in news texts. Their model aligned closely with human judgments and significantly outperformed a widely used rule-based regressor based on lexical patterns. Recent work by Shokri et al. evaluated large language models like GPT-3.5, GPT-4, and Gemini for sentence-level subjectivity detection in English news. They showed that while fine-tuned models perform well in-distribution, zero-shot and chain-of-thought prompting yield more robust generalization across diverse datasets. Also, Suwaileh et al. introduced ThatiAR, the first large-scale dataset for sentence-level subjectivity detection in Arabic news. They demonstrated that GPT-4, especially in few-shot settings, outperformed traditional and fine-tuned Arabic language models, highlighting the potential of LLMs in morphologically-rich and culturally diverse contexts.

3. Dataset

The dataset employed in this study originates from the shared task on subjectivity detection from the CheckThat! Lab 2025.[1] It is designed to evaluate the ability of computational systems to classify whether a sentence or short paragraph extracted from a news article expresses a subjective (**SUBJ**) or objective opinion (**OBJ**). The dataset comprises textual instances drawn from news sources in five languages: English, Italian, German, Bulgarian, and Arabic. For each language, the data is divided into four subsets: training, development, development-test, and test (with test labels withheld for evaluation purposes).

Table 1
Sentence and label distribution per language and dataset split. Each split shows total number of sentences and class distribution.

Language	Train			Dev			Dev-Test		
	Total	SUBJ	OBJ	Total	SUBJ	OBJ	Total	SUBJ	OBJ
English	830	298	532	462	240	222	484	122	362
Italian	1613	382	1231	667	177	490	513	136	377
German	800	308	492	491	174	317	337	111	226
Bulgarian	729	323	406	467	292	175	250	107	143
Arabic	2446	1055	1391	742	476	266	748	323	425

Table 1 provides a detailed overview of the sentence and label distribution across languages and dataset splits. Among the five languages, Arabic stands out with the largest training set (2,446 instances), while the remaining languages, Italian (1,613), English (830), German (800), and Bulgarian (729), are comparatively balanced in size. These differences likely reflect variation in source availability, annotation resources, and data curation priorities. The distribution of labels also varies by language and split: In the English training set, 64.10% of sentences are labeled as objective, while the remainder are subjective. The Italian data exhibits a strong bias towards objectivity across all splits, with over 73% of sentences labeled as OBJ. By contrast, the Bulgarian development set features a higher proportion of subjective content, highlighting potential cultural or editorial differences in reporting styles.

Table 2
Token statistics for each language on the test set.

Test Set					
Language	**Total**	**Avg. Length**	**Min**	**Max**	**Median**
English	300	28.79	2	114	28.0
Italian	299	28.34	2	113	23.0
German	347	31.46	8	117	28.0
Arabic	1036	42.39	5	175	38.0
Zero-Shot Romanian	206	34.56	2	151	30.0
Zero-Shot Ukrainian	297	28.17	2	114	26.0
Zero-Shot Greek	284	40.88	1	141	36.0
Zero-Shot Polish	351	29.43	4	97	26.0

Token-level statistics for the test set are summarized in Table 2, using the `xlm-roberta-base` tokenizer [26]. Arabic and Greek test sets have the longest sequences on average, while English, Italian, and Polish are more concise. These differences may impact model robustness across languages. During preprocessing, we identified several anomalous cases, particularly in the Bulgarian, German, and Italian splits, where open-ended quotation marks led to excessively long token sequences (often over 500 tokens). These likely stemmed from the tokenizer's handling of unmatched punctuation and were manually corrected to avoid distortions in length statistics.

[1] https://checkthat.gitlab.io/clef2025/task1/

3.1. Ambiguities and Hard-to-Translate Cases

To better understand common sources of labeling disagreement, we manually examined 20 sentences with annotation conflicts in the English development set. Five recurring themes emerged, each highlighting linguistic or contextual features that challenge binary subjectivity classification:

- **Immigration Discourse:** Statements such as *"Mr. Buchanan's criticism of immigration"* may appear factual but are often ideologically charged, subtly framing the topic in ways that evoke subjective interpretation.
- **Race and Social Commentary:** Sentences referencing phrases such as *"CRT anti-white curricula"* or *"diversity, equity, and inclusion"* are lexically neutral but semantically charged. The underlying ideological associations can trigger differing interpretations, often reflecting the annotator's sociopolitical context.
- **Media and Political Rhetoric:** Labels such as *"Lügenpresse"* or *"Treason Lobby"* embed explicit bias or contempt within declarative syntax, complicating detection by surface-level classifiers.
- **Sarcasm and Pragmatic Devices:** Utterances like *"What could possibly go wrong?"* rely on irony or context-based inference. Lacking overt opinion markers, they remain difficult to detect using standard lexical cues.
- **Framing in Economic and Environmental Topics:** Sentences such as *"bribed by a globalist billionaire"* combine factual assertions with emotionally charged language, blurring the line between reporting and commentary.

4. Approach

We explore two complementary approaches to multilingual subjectivity detection: supervised fine-tuning of transformer-based classifiers and zero-shot prompting with LLMs. The former trains task-specific classifiers on available annotated data, while the latter uses instruction-following capabilities of LLMs to perform inference without parameter updates. This section details the setup, training procedures, and reasoning strategies employed in both directions.

4.1. Fine-Tuned Transformers

Fine-tuning German-BERT with Translated Training Data The German training dataset was expanded by translating the other provided datasets into German, followed by fine-tuning a BERT model [27]. We ordered the languages from most to least similar to German, assuming that using the translation of training data of more similar languages will yield better results. First English (West Germanic, Indo-European), then Italian (Romance, Indo-European), Bulgarian (South Slavic, Indo-European) and Arabic (Semitic, Non-Indo-European) [28]. Even though Italian and Bulgarian are both Indo-European languages, Bulgarian is more distant from German than Italian in terms of Levenshtein distance [29]. By gradually adding more translated training data we monitored which additions improved the performance of the model as shown in Table 3.
In all training settings for the fine-tuning process, training data was shuffled, weight decay was set to 0.01, batch size was 32 and improved truncation and padding was applied. The number of epochs and the learning rate were adjusted to the size of the dataset, namely the amount of sentences. All experiments were conducted on a remote server equipped with an NVIDIA Tesla T4 GPU with 66 GB of RAM. Tokenizer and dataloader used for testing were the same ones that were used during training. The different fine-tuned models were compared regarding their macro F1-scores. The addition of the English, Italian, and Bulgarian translated training data increased the F1-score, but when including the Arabic data, it dropped. The F1-scores of the de-en-it-model and de-en-it-bg-model were very close, so we ran those fine-tuned models also on the dev-test dataset, showing a clear preference to include Bulgarian.

Table 3
Training Configuration and Performance Across Languages

Languages	Sentences	Epochs	Learning Rate	DEV macro-F1	DEV-TEST macro-F1
de	800	4	0.001	0.7253	-
de, en	1630	4	0.001	0.7405	-
de, en, it	3243	5	0.001	**0.7651**	0.7712
de, en, it, bg	3972	5	0.001	**0.7692**	**0.8172**
de, en, it, bg, ar	6418	5	0.004	0.7275	-

Monolingual Fine-tuning EuroBERT and XLM-RoBERTa-base To identify an effective architecture for monolingual subjectivity detection, we fine-tuned two transformer-based models: EuroBERT [30] and XLM-RoBERTa-base [31]. Both models are well-suited for sentence-level classification tasks, capable of capturing nuanced semantic and syntactic patterns. EuroBERT, a recently released multilingual model, is pre-trained primarily on European languages, aligning well with the linguistic coverage of our datasets [30]. XLM-RoBERTa-base was also selected for comparison due to its consistent cross-lingual performance and demonstrated effectiveness in prior sentence-level classification tasks [31, 32]. Experiments in monolingual setting were conducted on a remote server equipped with an NVIDIA H100 GPU with 80 GB of memory. On the provided datasets (excluding Arabic), we fine-tuned both models using a batch size of 16, a learning rate of 2e-5, and 15 training epochs. In the mean time, we employed early stopping with a patience value of 3 based on the macro-F1 score on the development set and applied temperature scaling post-training to calibrate prediction confidence [33]. We used the AdamW optimizer with a weight decay of 0.01 and employed Focal Loss [34] with class weighting to address label imbalance. To improve training efficiency and stability, we adopted mixed-precision training using PyTorch's AMP framework and applied gradient clipping with a maximum norm of 1.0 to prevent exploding gradients. After training, we evaluated both models on the dev-test datasets using a structured inference pipeline. Input sentences were tokenized using the same configuration as during training and passed through the model to obtain raw logits. These logits were optionally calibrated using temperature scaling and then converted to class probabilities via the softmax function. Due to the observed class imbalance in the training data across all languages, where the SUBJ class occupies only around 37% of instances on average and is consistently under-represented compared to OBJ, we thus applied a reduced classification threshold of 0.45 (instead of the standard 0.5) for predicting the SUBJ label. Final predictions were mapped to their corresponding labels (OBJ or SUBJ) and compared to gold-standard labels to compute accuracy and macro-F1, as specified by the shared task organizers.

4.2. LLM-Based Inference Strategies

In addition to our fine-tuned classifiers, we implement three zero-shot prompting strategies using LLMs. Each strategy frames subjectivity detection as an inference-only task and generates natural-language responses from one or more prompts per sentence. The three methods differ in reasoning style but operate directly on the input text without task-specific training, as seen in the overview in Figure 1.

4.2.1. Annotation Approach

In this inference strategy, each sentence is classified with a single API call to the OpenAI `o3-mini` model. [35] We supply the model with the full set of 14 decision rules [22], which define linguistic and rhetorical cues, to differentiate subjective from objective language. Upon receiving the model's natural-language response, we extract both, the verdict and the supporting explanation, in two stages: first, by parsing the response as JSON and reading the `verdict` and `explanation` fields; and, if JSON parsing fails, by performing a simple keyword match for "*subjective*" or "*objective*" within the raw text. In all cases, we archive both the generated explanation and the final label. The following example illustrates how each of our prompting strategies processes the same input sentence.

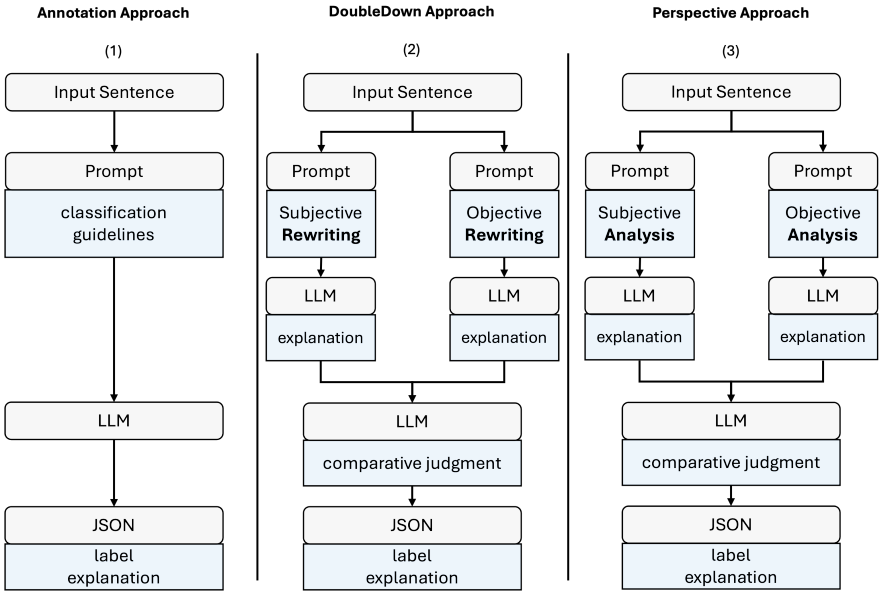

Figure 1: Overview of the three prompting strategies for subjectivity detection. In the **(1) Annotation approach**, the LLM classifies the input sentence based on explicit guidelines provided in the prompt. **(2) DoubleDown** rewrites the sentence in both subjective and objective styles and infers the label by comparing which version better preserves the original meaning. **(3) Perspective** generates explanations for both possible interpretations and selects the label based on the more convincing reasoning.

> **Input Sentence:** "Blanco established himself earlier in his career working for Dr. Luke's Kasz Money Productions."

> **Explanation:** The sentence provides factual information about Blanco's career and his affiliation with a production company. It does not include any indications of personal opinion, sarcastic remarks, or evaluative language by the author. Instead, it merely states a historical fact, which aligns with the criteria for an objective sentence. **Label:** OBJ

4.2.2. DoubleDown Approach

The DoubleDown approach reframes classification as a two-way comparison instead of a direct label. Rather than asking the model to judge the original sentence outright, it first generates two rewrites, one that leans into a subjective style (opinion, personal evaluation or emotional language) and a second that adopts a more objective tone (factual content, neutral phrasing). Both rewrites come from separate `gpt-4.1-mini` prompts. [36] In the final step, the original sentence plus both rewrites are submitted to a third prompt, which determines which version best preserves the original meaning, and on that basis classifies the input as subjective or objective. Below, you can see how the sentence is reformulated in both its subjective and objective versions.

> **Subjective rewrite:** In my view, Blanco really made a name for himself early on thanks to his work with Dr. Luke's Kasz Money Productions collaboration that, to me, marked a crucial turning point in his career.
> **Objective rewrite:** Blanco worked earlier in his career at Dr. Luke's Kasz Money Productions.

Since the objective rewrite more closely reflects the original sentence and presents it as a factual career statement with only minor evaluative elements, the model classifies it as **OBJ**.

4.2.3. Perspective Approach

The *Perspective* approach frames the classification task as a stepwise evaluation of contrasting interpretations. Rather than directly predicting a label, the model is prompted to analyze the input from two distinct angles, subjective and objective, before making a decision. All prompts in this approach are processed by `gpt-4.1-mini`. [36] The process is divided into three steps. First, the model receives a prompt asking it to explain why the sentence might be considered subjective, focusing on aspects such as evaluative language or personal judgment. In the second step, the same sentence is analyzed from an objective perspective, highlighting features like factuality, neutrality, or verifiability. These two prompts are issued separately to ensure that the analyses are independently reasoned. In the third step, the outputs from the previous prompts are provided as input to a final classification prompt. The model is asked to compare both analyses and determine which interpretation is more convincing. It then outputs a classification along with a justification. This approach prioritizes analytical comparison over generative rewriting like in the *DoubleDown* approach, as seen below.

> **Explanation:** The statement contains elements that can be viewed both subjectively and objectively. The subjective analysis points out that the phrase "established himself" is open to interpretation, as what qualifies as "established" can vary by individual perception, making it a somewhat evaluative judgment. The objective analysis highlights that the statement refers to a verifiable fact: Blanco worked for Dr. Luke's Kasz Money Productions earlier in his career. This part can be independently confirmed. However, the key phrase "established himself" goes beyond merely stating a fact about employment; it implies a level of success, recognition, or impact, which is inherently subjective because these concepts differ across perspectives. Therefore, while the statement contains a factual component, the primary assertion involves a subjective judgment. Given this, the subjective analysis is more convincing because the core claim revolves around the idea of "establishing oneself", which is not a strictly objective measure. **Label:** SUBJ

5. Experiments and Results

Table 4 summarizes the official evaluation results of our submitted systems on the final test set, as provided by the shared task organizers. We submitted one system per subtask, using a combination of zero-shot prompting and fine-tuned transformer models. The approach and model were chosen individually for each language, based on preliminary development set results and practical observations. In the Italian monolingual subtask, our system reached the highest macro-F_1 score on the leaderboard (0.8104) using the *Annotation Approach* with the `o3-mini` model. This suggests that prompt-based classification can work well when the input data closely follows the structure of the rules described in the prompt. For Romanian, we used a fine-tuned XLM-RoBERTa model and achieved an F1 score of 0.7917. This result placed us third overall and indicates that the model was able to generalize reasonably well, even though labeled subjectivity data in Romanian was not part of our fine-tuning. In the multilingual setting, we again used a fine-tuned *XLM-RoBERTa*, which reached 0.7186 macro-F_1, clearly outperforming the baseline. The model showed stable results across several languages, including Greek and Polish. For German, we fine-tuned a *German-BERT* model using translated training data from related languages. This approach led to solid performance (F_1 = 0.7269) and confirms that adding training data from similar languages can be helpful when working with limited resources, as also observed by Solopova et al. for german language. In English, we used the *Annotation Approach*, which achieved 0.7228 and outperformed the baseline. This result supports the idea that rule-based prompting can be effective in high-resource settings where the classification cues are well captured by the guidelines. Performance in the zero-shot subtasks varied. In Ukrainian, our system scored 0.6124, slightly below the baseline. In Greek, it reached 0.4750, showing a moderate improvement over the baseline. For Polish, the model scored 0.5665, which was slightly below the baseline (0.5719). These results suggest that zero-shot performance depends not only on the model itself, but also on the similarity between the training

and test languages, and the phrasing patterns in the input data. Due to the shared task submission protocol, only one system could be submitted per language. This restriction limited our ability to systematically compare multiple approaches across all languages. Consequently, the selected system for each subtask reflects a pragmatic decision based on development performance and informal testing, rather than a globally optimal configuration. This was especially relevant for zero-shot settings, where generalization is influenced by a combination of linguistic similarity, domain coverage, and how well task framing aligns with the model's training data. Among the prompt-based methods, the Annotation strategy proved more robust than the more complex comparative prompting variants. XLM-RoBERTa consistently outperformed EuroBERT across all settings tested during development, particularly in multilingual and cross-lingual tasks.

Table 4
Evaluation results for our approach across subtasks, compared to the baselines and top-ranked scores. Rows are sorted by how close our F_1 score is to the top result. Bold values in *Our* F_1 indicate scores above the baseline.

Subtask	Baseline F_1	Our F_1	Top Score F_1	Approach
Monolingual Italian	0.6941	**0.8104**	**0.8104**	Annotation
Zero-Shot Romanian	0.6461	**0.7917**	0.8126	XLM-RoBERTa
Multilingual Subjectivity	0.6390	**0.7186**	0.7550	XLM-RoBERTa
Zero-Shot Ukrainian	0.6296	0.6124	0.6424	XLM-RoBERTa
Zero-Shot Greek	0.4159	**0.4750**	0.5067	XLM-RoBERTa
Monolingual English	0.5370	**0.7228**	0.8052	Annotation
Monolingual German	0.6960	**0.7269**	0.8520	German-BERT
Zero-Shot Polish	0.5719	0.5665	0.6922	XLM-RoBERTa

6. Conclusion and Future Work

We presented a multilingual system for subjectivity detection using two main approaches: fine-tuned transformer models and zero-shot prompting with LLMs. Our results in the CheckThat! 2025 shared task show that both directions can be effective, depending on language and resource availability. For fine-tuned models, XLM-RoBERTa delivered the most consistent performance and was used in several subtasks, including multilingual and zero-shot settings. For German, we observed that fine-tuning a German-BERT model with translated training data led to competitive results. On the prompting side, the Annotation Approach with the o3-mini model performed well in high-resource languages, such as Italian and English, where classification rules were clearly reflected in the data. Due to the submission constraint of only one system per language, we could not test all combinations of models and approaches systematically. Our choices were based on limited development set results and informal comparisons. This affected our ability to fully explore the strengths and weaknesses of each approach across languages, especially for zero-shot cases. Another important limitation was the lack of broader context for each sentence. Since the task involved classifying isolated sentences, it was often difficult to judge subjectivity accurately without context. This made the task especially challenging when subjective language relied on surrounding sentences. Another issue is the imbalance of label distributions across languages, that is most notably in the Bulgarian development set, where subjective sentences dominate. These imbalances can lead models to internalize and amplify misleading associations, potentially reinforcing biases and over-predicting subjectivity in certain languages or cultural contexts. For future work, it would be valuable to systematically compare prompting and fine-tuned approaches across languages and subtasks under controlled conditions. In particular, we aim to better understand which types of tasks or linguistic features favor instruction-based inference over supervised training. Additionally, exploring more flexible combinations of prompting and fine-tuning, e.g., via model ensembling or fallback strategies such as few-shot prompting or confidence-based model switching, could help improve performance, especially in low-resource or zero-shot settings.

Acknowledgments

This research is funded by the Federal Ministry of Research, Technology and Space (BMFTR, reference: 03RU2U151C) in the scope of the research project news-polygraph.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.

[2] 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, 2003, pp. 129–136.

[3] J. Ruotsalainen, J. Hujanen, M. Villi, A future of journalism beyond the objectivity–dialogue divide? hybridity in the news of entrepreneurial journalists, Journalism 22 (2021) 2240–2258.

[4] F.-J. Rodrigo-Ginés, J. Carrillo-de Albornoz, L. Plaza, A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it, Expert Systems with Applications 237 (2024) 121641.

[5] I. Chaturvedi, E. Cambria, R. E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Information Fusion 44 (2018) 65–77.

[6] P. Kasnesis, L. Toumanidis, C. Z. Patrikakis, Combating fake news with transformers: a comparative analysis of stance detection and subjectivity analysis, Information 12 (2021) 409.

[7] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[8] P. Sahitaj, I. Maab, J. Yamagishi, J. Kolanowski, S. Möller, V. Schmitt, Towards automated fact-checking of real-world claims: Exploring task formulation and assessment with llms, arXiv preprint arXiv:2502.08909 (2025).

[9] T. Scheffler, V. Solopova, M. Popa-Wyatt, The telegram chronicles of online harm, Journal of Open Humanities Data 7 (2021).

[10] P. Sahitaj, R. Ruiz-Dolz, A. Sahitaj, A. Nizamoglu, V. Schmitt, S. Mohtaj, S. Möller, From construction to application: Advancing argument mining with the large-scale kialoprime dataset, in: Computational Models of Argument, IOS Press, 2024, pp. 229–240.

[11] 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.

[12] G. Palshikar, M. Apte, D. Pandita, V. Singh, Learning to identify subjective sentences, in: Proceedings of the 13th International Conference on Natural Language Processing, 2016, pp. 239–248.

[13] B. Liu, et al., Sentiment analysis and subjectivity., Handbook of natural language processing 2 (2010) 627–666.

[14] A. Sahitaj, P. Sahitaj, S. Mohtaj, S. Möller, V. Schmitt, Towards a computational framework for distinguishing critical and conspiratorial texts by elaborating on the context and argumentation with llms, Working Notes of CLEF (2024).

[15] F. M. S. Eid, M. S. A. Mutahar, Bridging rhetorical differences: Arabic textual metaphors in academic writing and translation, European Journal of Arts, Humanities and Social Sciences 2 (2025) 183–201.

[16] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, P. Kazienko, Learning personal human biases and representations for subjective tasks in natural language processing, in: 2021 IEEE international conference on data mining (ICDM), IEEE, 2021, pp. 1168–1173.
[17] F. Ruggeri, A. Muti, K. Korre, J. M. Struß, M. Siegel, M. Wiegand, F. Alam, R. Biswas, W. Zaghouani, M. Nawrocka, B. Ivasiuk, G. Razvan, A. Mihail, Overview of the CLEF-2025 CheckThat! lab task 1 on subjectivity in news article, ????
[18] J. Wiebe, T. Wilson, R. Bruce, M. Bell, M. Martin, Learning subjective language, Computational linguistics 30 (2004) 277–308.
[19] E. Riloff, J. Wiebe, W. Phillips, Exploiting subjectivity classification to improve information extraction, in: AAAI, 2005, pp. 1106–1111.
[20] A. Esuli, F. Sebastiani, et al., Sentiwordnet: A publicly available lexical resource for opinion mining., in: LREC, volume 6, 2006, pp. 417–422.
[21] C. Banea, R. Mihalcea, J. Wiebe, A bootstrapping method for building subjectivity lexicons for languages with scarce resources., in: LREC, volume 8, 2008, pp. 2–764.
[22] 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. URL: https://aclanthology.org/2024.lrec-main.25/.
[23] E. Savinova, F. M. Del Prado, Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements, in: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, 2023, pp. 305–314.
[24] M. Shokri, V. Sharma, E. Filatova, S. Jain, S. Levitan, Subjectivity detection in english news using large language models, in: Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, 2024, pp. 215–226.
[25] R. Suwaileh, M. Hasanain, F. Hubail, W. Zaghouani, F. Alam, Thatiar: subjectivity detection in arabic news sentences, arXiv preprint arXiv:2406.05559 (2024).
[26] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[27] B. Chan, T. Möller, M. Pietsch, T. Soni, Bert base german cased, 2020. URL: https://huggingface.co/google-bert/bert-base-german-cased.
[28] L. Campbell, Historical linguistics, Edinburgh University Press, 2013.
[29] M. Serva, F. Petroni, Indo-european languages tree by levenshtein distance, EPL (Europhysics Letters) 81 (2008) 68005. URL: http://dx.doi.org/10.1209/0295-5075/81/68005. doi:10.1209/0295-5075/81/68005.
[30] N. Boizard, H. Gisserot-Boukhlef, D. M. Alves, A. Martins, A. Hammal, C. Corro, C. Hudelot, E. Malherbe, E. Malaboeuf, F. Jourdan, G. Hautreux, J. Alves, K. El-Haddad, M. Faysse, M. Peyrard, N. M. Guerreiro, P. Fernandes, R. Rei, P. Colombo, Eurobert: Scaling multilingual encoders for european languages, 2025. URL: https://arxiv.org/abs/2503.05500. arXiv:2503.05500.
[31] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[32] J. M. Struß, F. Ruggeri, A. Barron-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: G. Faggioli, N. Ferro, P. Galuščakova, A. G. S. de Herrera (Eds.), CLEF 2024 Working Notes : Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Notebook for the CheckThat! Lab at CLEF 2024, 2024, pp. 287 – 298. URL: https://ceur-ws.org/Vol-3740/paper-25.pdf.
[33] A. S. Mozafari, H. S. Gomes, W. Leão, S. Janny, C. Gagné, Attended temperature scaling: A practical approach for calibrating deep neural networks, 2019. URL: https://arxiv.org/abs/1810.11586. arXiv:1810.11586.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. URL: https://arxiv.org/abs/1708.02002. arXiv:1708.02002.

[35] OpenAI, Openai o3-mini, 2025. URL: https://openai.com/index/openai-o3-mini/, accessed: 2025-05-28.

[36] OpenAI, Gpt-4.1, 2024. URL: https://openai.com/index/gpt-4-1/, accessed: 2025-05-28.

[37] V. Solopova, V. Herman, C. Benzmüller, T. Landgraf, Check news in one click: NLP-empowered pro-kremlin propaganda detection, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 44–51. URL: https://aclanthology.org/2024.eacl-demo.6/.